

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/3333312>

Modern Methods of Speech Processing

Article in *IEEE Transactions on Speech and Audio Processing* · February 1994

DOI: 10.1109/89.260357 · Source: IEEE Xplore

CITATIONS

158

READS

96

1 author:



Oded Ghitza

Boston University

57 PUBLICATIONS 1,245 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Look at my HomePage: [View project](#)

All content following this page was uploaded by [Oded Ghitza](#) on 23 September 2014.

The user has requested enhancement of the downloaded file. All in-text references [underlined in blue](#) are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition

Oded Ghitza, *Senior Member, IEEE*

Abstract—Auditory models that are capable of achieving human performance in tasks related to speech perception would provide a basis for realizing effective speech processing systems. Saving bits in speech coders, for example, relies on a perceptual tolerance to acoustic deviations from the original speech. Perceptual invariance to adverse signal conditions (noise, microphone and channel distortions, room reverberations) and to phonemic variability (due to nonuniqueness of articulatory gestures) may provide a basis for robust speech recognition. A state-of-the-art auditory model that simulates, in considerable detail, the outer parts of the auditory periphery up through the auditory nerve level is described. Speech information is extracted from the simulated auditory nerve firings, and used in place of the conventional input to several speech coding and recognition systems. The performance of these systems improves as a result of this replacement, but is still short of achieving human performance. The shortcomings occur, in particular, in tasks related to low bit-rate coding and to speech recognition. Since schemes for low bit-rate coding rely on signal manipulations that spread over durations of several tens of ms, and since schemes for speech recognition rely on phonemic/articulatory information that extend over similar time intervals, it is concluded that the shortcomings are due mainly to a perceptually related rules over durations of 50–100 ms. These observations suggest a need for a study aimed at understanding how auditory nerve activity is integrated over time intervals of that duration. We discuss preliminary experimental results that confirm human usage of such integration, with different integration rules for different time-frequency regions depending on the phoneme-discrimination task.

I. INTRODUCTION

IN building speech-processing systems, we make the axiomatic assumption that the speech waveform contains the entire information produced by the human speech production mechanism. In a particular speech-processing system, however, only part of this information is used, depending on the task for which the system is designed. For speech recognition tasks, for example, only phonemic information is needed. For speech coding tasks, on the other hand, information associated with the quality of speech is also required.

Traditionally, the task-related information is extracted from the speech waveform (or its Fourier representation) using statistical inference methods. This paper discusses an alternative way to extract this information that relies on processing principles derived from properties of the auditory system. The premise of this approach is that such processing principles can provide a basis for realizing effective speech processing

systems. Saving bits in speech coders, for example, relies on perceptual tolerance to acoustic deviations from the original speech. Perceptual invariance to adverse signal conditions (noise, microphone and channel distortions, room reverberations) and to phonemic variability (due to nonuniqueness of articulatory gestures) may provide a basis for robust speech recognition.

The potential advantages that can be gained by utilizing auditory models for speech processing depend on how accurate the models are in mimicking human performance. And building such accurate models depend on the amount of knowledge we have about the auditory system. This knowledge is acquired by combining data that has been collected in both psychophysical and physiological studies of the auditory system. Studies of cochlear mechanics and studies of mechanical to neural transduction in the cochlea provide insight into the processing of sounds in the pre-auditory-nerve stages of the auditory periphery. Studies of the population response of single auditory nerve fibers in the cat to speech-like signals provide a rich source of information concerning the principles by which such sounds are encoded in the auditory nerve. In contrast, little is known, at present, about the functional operation of auditory nuclei beyond the auditory nerve.

Current research activity in auditory modeling is mostly devoted to the study of the auditory periphery. Several papers have been published that examine how the response of the cochlea may be processed to provide a relevant representation of the speech signal. Each study utilizes a computational model to simulate either the direct firing activity or another related representation of the cochlear output. However, the manner in which this information is processed differs among the studies, reflecting differences in the structural properties of the central processor hypothesized by each study. These structural properties can be described by using a two-component characterization: the place/nonplace component, that indicates if the central processor utilizes explicit knowledge about the fibers' tonotopic place of origin in the cochlear partition, and the rate/temporal component, that indicates whether the central processor uses instantaneous firing rate measurements alone or higher-order firing statistics (e.g., the interspike interval statistics). Using this two-component criterion, the following categories are traditionally used: (1) place/rate category, where the central processor possesses explicit knowledge of place and uses only instantaneous rate information [7], [24], [30], (2) place/temporal category, where place information is used together with detailed temporal information of local neural responses [8], [14], [31], [32], and (3) nonplace/temporal

Manuscript received March 2, 1993; revised September 15, 1993.

The author is with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9214434.

category, where place information is omitted altogether and the only sources of information are the temporal properties of the global neural response [1], [9].¹

This paper focuses on a model that belongs to the last category. Detailed modeling of the auditory periphery (up through the auditory nerve level) is used to simulate the firing activity of the auditory nerve. The model consists of 190 cochlear channels, distributed from 200 Hz to 7000 Hz according to the frequency-position relation suggested by [15]. Each channel comprises Goldstein's nonlinear model of the human cochlea [13], followed by an array of five level-crossing detectors that simulate the auditory nerve fibers innervating one inner hair cell. At this junction of the model the speech sound is represented as a 950-dimensional point process (190 channels times five levels per channel) that simulates the firing activity of the auditory nerve fibers.

One way to achieve human performance is to process the simulated auditory nerve firing activity according to principles derived from properties of post auditory-nerve nuclei. Unfortunately, such information is not available at present. Our approach, therefore, is to process the simulated auditory nerve firing patterns according to principles that are motivated by observed properties of actual auditory nerve response. In Section II, a representation of this kind is described as an ensemble interval histogram (EIH). Conceptually, the EIH is a measure of the spatial (tonotopic) extent of coherent activity across the simulated auditory nerve. This measure signifies different physical properties of the acoustic stimulus depending on frequency: Information at low frequencies is represented in terms of the extent of coherent firing activity that is phase-locked to the underlying resolved components, while information at high frequencies is represented as the extent of coherent instantaneous rate of the firing activity driven by a wideband, unresolved (in frequency) temporal event. This representation is in accordance with observed properties of auditory nerve response, which show a diminishing degree of phase locking of the neural firings to the driving stimulus, as the frequency of the stimulus increases (e.g., [18]).

As pointed out before, the advantages that can be gained by utilizing auditory models for speech processing depend on how accurate the models are in mimicking human performance. In Sections III and IV we evaluate the capability of the EIH to achieve human performance. Using psychophysical results as a reference we examine the extent to which speech processing systems that use the auditory model mimic human performance, in tasks related to speech recognition and to speech coding. In Section III we evaluate the EIH in the context of a speech recognition task (after [10]). We first identify a psychophysical experiment that addresses such a task, the Diagnostic Rhyme Test, or DRT [33] and present data we have collected on how accurate human listeners are in performing the task. Next, we simulate the DRT procedure. The auditory periphery is replaced by the EIH and the higher auditory elements by an automatic speech recognition system. Errors are displayed in terms of a distribution among six

phonemically distinctive features. The error patterns of the simulated procedure are then compared with those of the human subjects, to provide a quantitative measure on how close the model performance is to the actual human performance. The evaluation process shows that performance greatly improves by replacing a conventional speech representation by the EIH, but is still short of achieving human performance.

In Section IV we evaluate the EIH in the context of speech coding. Following the methodology used in Section III, we first identified a psychophysical experiment that addresses a speech coding task (the Mean-Opinion-Score, or MOS, a test which is widely used to assess quality of speech coders) and collected data on how human listeners score the quality of synthetic speech produced by different speech coding systems. Then, we use the EIH as a basis for a system aiming at predicting the MOS. Since the EIH is computed from simulated auditory nerve responses modeled in considerable detail, we hypothesize that it contains only perceptually-relevant speech information. If this hypothesis holds, perceptual differences between the original and the coded speech should be realistically reflected in the EIH domain. To examine this hypothesis, we mimicked the MOS test by measuring the L_2 -norm of the difference between the EIHs of the original and the coded speech. For comparison purposes we also examined the L_2 -norm in the cepstral domain. The comparison shows that EIH provides better MOS predictions for coders with a bit-rate of 16 kbit/s and above, but underrates the quality assessment of CELP-type of coders (i.e., coders with a bit-rate of 8 kbit/s and below).

The evaluation process described in Sections III and IV demonstrates that generally, performance is improved by replacing conventional speech representation methods by the auditory model, but is still short of achieving human performance. The evaluation process indicates that shortcomings occur, in particular, in tasks related to low bit-rate coding and to speech recognition. In Section V-A we discuss possible reasons for this inadequacy. Since schemes for low bit-rate coding rely on signal manipulations that spread over durations of several tens of ms, and since schemes for speech recognition rely on phonemic/articulatory information that extend over similar time intervals, it is concluded that the shortcomings are due mainly to the inappropriate manner by which we currently measure a distance between two speech segments. A typical way to represent a speech segment is in terms of an ordered sequence of observations, sampled at a uniform rate. (The observations can be expressed, for example, in terms of EIH or in terms of Fourier power spectrum). Hence, to measure the distance between two speech segments we must define a distance metric between two **sequences** of observations. This, in turn, requires a definition of a distance metric between two **single** observations. In the experiments of Sections III and IV, the distance between two single EIH observations is measured in the L_2 -norm sense. The distance between two sequences of EIH observations is measured as the accumulated sum of distances between corresponding single observations along a path that best matches the sequences in the sense of minimizing the accumulated sum. The experimental results of Sections III and IV suggest that both, the L_2 -norm (of

¹ For an excellent overview of auditory models, the reader is referred to the theme issue "Representation of Speech in the Auditory Periphery", edited by S. Greenberg, *Journal of Phonetics*, Volume 16, No. 1, January 1988.

the difference between two single EIH observations) and the accumulated sum (for two sequences of EIH observations) are inadequate measures for the purpose of predicting human performance. They also imply a need for a study aimed at understanding how auditory nerve activity is integrated over durations of 50–100 ms. In Section V-B we discuss preliminary experimental results that confirm human usage of such integration, with different integration rules for different time-frequency regions depending on the phoneme-discrimination task.

II. THE ENSEMBLE INTERVAL HISTOGRAM (EIH) REPRESENTATION

This section describes the EIH model. The model comprises two stages, one that models the pre-auditory-nerve section and one that stands for the post auditory-nerve section of the auditory periphery. The pre-auditory-nerve section has been modeled in considerable detail, guided by the physiological structure of the auditory periphery. In contrast, the post auditory-nerve section is represented in an heuristic manner since little is known, so far, about the operation of auditory functions associated with that part of the auditory periphery. In Section II-B, we review the pre-auditory-nerve portion of the auditory periphery. The considerations in modeling this part of the auditory periphery are discussed next: the model for the mechanical displacement of the basilar membrane is described in Section II-C, and the model for the mechanical-to-neural transduction is described in Section II-D. The second stage of the EIH model is discussed in Section II-E.

Before describing the EIH model, let us briefly describe the structure of the mammalian auditory system.² This will put into perspective the state of our knowledge and what part of it is being utilized in our current models.

A. A Brief Description of the Auditory Pathway

The auditory system is symmetrically organized around a midline between the left and the right sides (see Fig. 1). The most peripheral part at each side consists of the **external and middle ears**, the **cochlea**, and the **auditory nerve**. This portion of the auditory system will be described in detail, in Section II-B. The auditory nervous system, which receives its inputs from both left and right auditory nerves, consists of several neural nuclei that can be grouped into three major parts—the auditory brainstem, the auditory midbrain, and the auditory cortex. Each nucleus can be divided into different zones, characterized by their morphological structure, neurophysiological response, and their input/output mappings. The main information flow is along an ascending pathway that begins at the external ears and ends in the auditory cortex. In parallel, there is an information flow in a descending pathway that begins at the auditory cortex and can be traced all the way down to the middle ears.

Fig. 1 shows the major nuclei of the mammalian auditory system. The auditory nerve at each side projects to its corresponding **cochlear nucleus** which is the first nucleus in

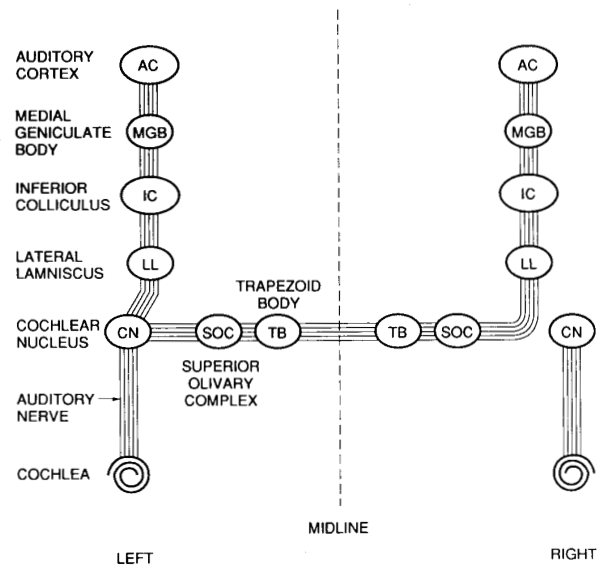


Fig. 1. Major nuclei of the mammalian auditory system.

the auditory brainstem. The projections of various divisions in each cochlear nucleus form two parallel branches, one on the same side of the cochlear nucleus and the other across the midline. These projections, on both sides, proceed either directly or indirectly via other brainstem nuclei (the **trapezoid body** and the **superior olivary complex**), to the auditory midbrain nuclei (the **lateral lemniscus** and the **inferior colliculus**). The inferior colliculus provides the major ascending path to the **medial geniculate body** which, in turn, provides inputs to the **primary auditory cortex**.

B. Physiological Basis for the EIH Model

A diagram of the pre-auditory-nerve portion of the auditory periphery is shown in Fig. 2(a). This part of the auditory periphery comprises three distinct parts: the outer ear, the middle ear, and the inner ear. The outer ear consists of the **pinna** (the ear surface surrounding the canal in which sound is funneled) and the **external canal**. Sound waves are guided through the outer ear to the middle ear, which consists of the **eardrum** (which moves due to the sound pressure) and a mechanical transducer comprises the **hammer**, the **incus** and the **stapes** (which conveys the motion of the eardrum into mechanical vibrations along the inner ear). The inner ear consists of the **cochlea**, which is a fluid-filled chamber partitioned by the basilar membrane, and the **auditory nerve**. The mechanical vibrations at the entrance of the cochlea (a 2 1/2 turn, snail-like tube, shown as a straight tube in Fig. 2(b)) excites the fluid inside the cochlea and cause the basilar membrane to vibrate at places associated with the frequencies of the input acoustic wave. Distributed along the basilar membrane (in a dense but discrete manner) are sensors called **inner hair cells (IHC)** that are innervated by the auditory nerve fibers and transform the mechanical displacement of the basilar membrane into firing activity at the nerve fibers.

² For an introductory survey of the auditory system the following references are recommended: [27] and [19].

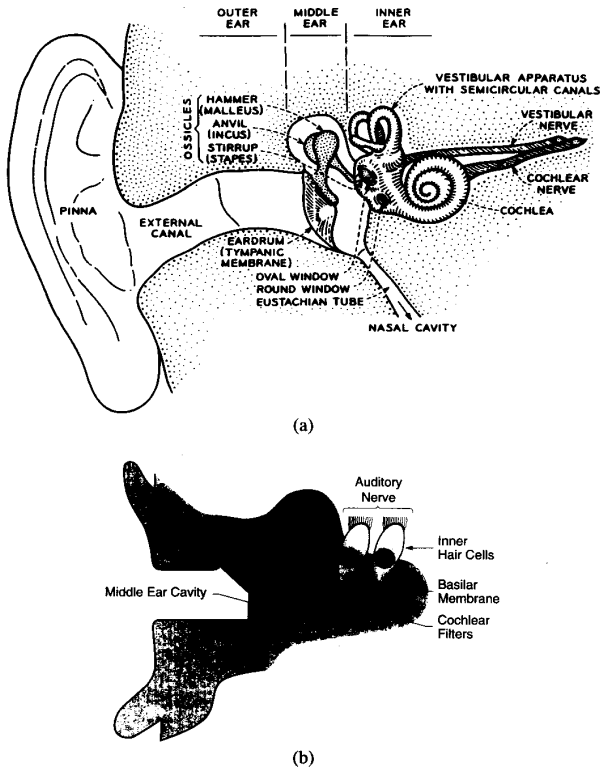


Fig. 2. (a) A physiological model of the pre-auditory-nerve parts of the auditory periphery. (b) A simplified view of (a). The 2 1/2 turn, snail-like shape of the cochlea is shown as a straight tube.

The mechanical displacement of the basilar membrane, at any given place, can be viewed as the output signal of a band-pass filter whose frequency response has a resonance peak at a frequency which is characteristic of the place. This resonance frequency is called characteristic frequency (CF). The log of CF is approximately proportional to distance along the membrane, and the distribution of the inner hair cells (IHC's) along the cochlear partition is essentially uniform. (There are some 4000 IHC's along the basilar membrane). The displacement of the basilar membrane is reflected in the AC component of the IHC receptor potential. The transformation from mechanical motion to receptor potential involves several nonlinearities, the most relevant to the present discussion being the half-wave rectification which is a consequence of the unidirectional depolarization of the IHC. Each IHC is innervated by approximately ten auditory-nerve fibers, whose spontaneous activity ranges between 0 and 100 discharges per second. The spontaneous rate is highly correlated with the fiber diameter and the size of the synaptic region between the fiber and inner hair cell. The spontaneous discharge rate is also correlated with the threshold of response. For any given CF region, fibers with high spontaneous rate tend to have between 5 and 20 dB lower threshold than units with low rates of background activity. Occasionally, low-spontaneous units may have as much as 40–60 dB higher thresholds than high-spontaneous units of comparable CF [22], [23].

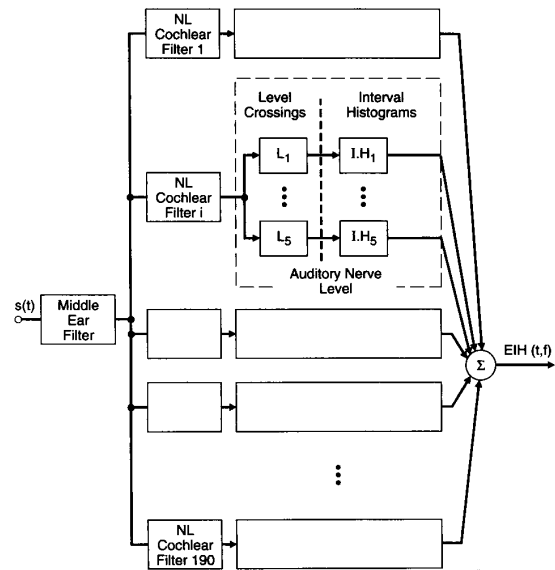


Fig. 3. The ensemble interval histogram (EIH) model.

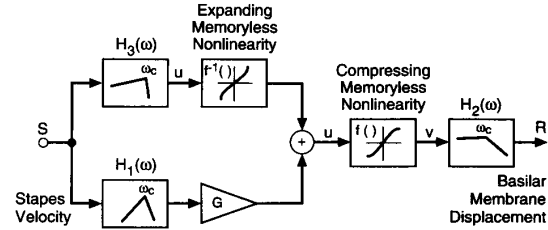


Fig. 4. The MBPNL cochlear model (after [13]).

C. Modeling the Mechanical Displacement of the Basilar Membrane

The EIH model is schematically illustrated in Fig. 3. Its first stage represents the auditory periphery up through the level of the auditory nerve. The middle ear is modeled as a high-pass filter, with a cutoff frequency at 1000 Hz and a slope of 20 dB/decade. The mechanical displacement of the basilar membrane is sampled by 190 IHC channels distributed from 200 Hz to 7000 Hz according to the frequency-position relation suggested by [15]

$$F = A(10^{ax} - 1) \quad (1)$$

where F is frequency in Hz, x is the normalized distance along the membrane (i.e., $0 < x < 1$, $x = 0$ is the apex) and the appropriate constants for the human cochlea are $A = 165.4$ and $a = 2.1$. The corresponding mechanical motion in each channel is simulated using a model of the human cochlea suggested by Goldstein (1990). The model is termed multibandpass nonlinear filter (MBPNL). It operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed psychophysical behavior. The MBPNL model is shown in Fig. 4. The lower signal processing path (H_1 - H_2) is a compressive nonlinear filter that represents the sensitive, narrowband compressive

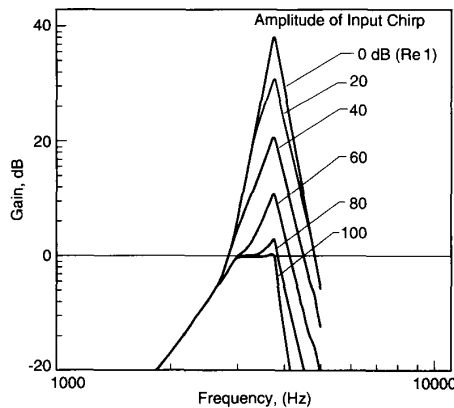


Fig. 5. Iso-input frequency response of the MBPNL model at $CF = 3400$ Hz.

nonlinearity at the tip of the basilar-membrane response. The upper signaling path (H3-H2) is a linear filter that represents the insensitive, broadband linear tail response of basilar-membrane response (after [13]).

The "iso-input" frequency response of an MBPNL filter at CF of 3400 Hz is shown in Fig. 5. For an input signal $s(t) = A \sin(\omega_o t)$, with A and ω_o fixed, the MBPNL behaves as a linear system, with a fixed "operating point" on the nonlinear curves f and f^{-1} of Fig. 4, determined by A . Fig. 5 shows the iso-input frequency response of the system for different values of A : For a given A , a discrete "chirp" signal was presented to the MBPNL, with a slowly changing frequency. Changes in ω_o occurred only after the system reached steady-state, for a proper gain measurement. For a 0 dB input level ($A = 1$), the gain at CF is approximately 40 dB. As the input level increases the gain drops, the bandwidth increases and CF shifts. These observations are in accordance with psychophysical behavior.

Fig. 6 illustrates the high degree of overlap among the simulated cochlear channels. It shows the iso-input response at input level of 60 dB. Only one of every ten channels is drawn.

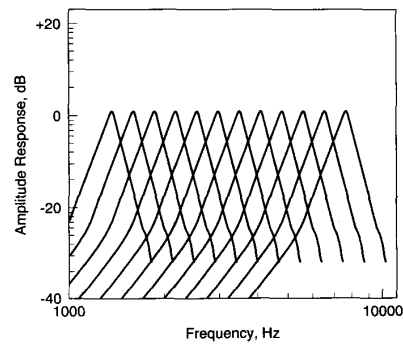


Fig. 6. An illustration of the degree of overlap among the cochlear channels. Shown are the iso-input response at input level of 60 dB. Only one of every ten channels is drawn.

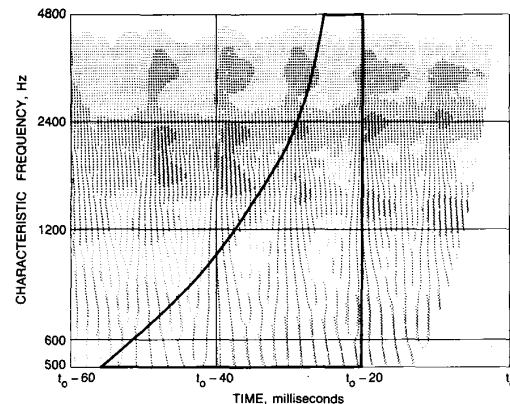


Fig. 7. Simulated auditory-nerve activity for the first 60 ms inside the vowel [o] in the word "gob." The abscissa represents time and the ordinate represents the characteristic frequency of the IHC channels. Note the logarithmic scale of the characteristic frequency, which represents the place-to-frequency mapping at the basilar membrane. In the figure, a level-crossing occurrence is marked as a single dot, and the output activity of each level-crossing detector is plotted as a separate trace. Each IHC channel contributes five parallel traces, with the lower trace representing the lower-threshold level-crossing detector. If the magnitude of the filter's output-signal is low, only one level will be crossed, as is the case for the very top channels of the figure. However, for large signal magnitudes, several levels will be activated, creating a "darker" area of activity. The figure also illustrates how the length of the analysis window in each channel is related to its CF.

D. Modeling the Mechanical to Neural Transduction

The ensemble of nerve fibers innervating a single IHC is simulated with an array of level-crossing detectors at the output of each cochlear filter (i.e., each level-crossing detector is equivalent to a fiber of specific threshold). A neural firing is simulated as the positive-going level crossing. The thresholds are distributed across a range of positive levels, to account for the half-wave rectification nature of the IHC receptor potential. The values assigned to the level j of every filter is a random Gaussian variable, with a mean, L_j , and a standard deviation, $\sigma_j = 0.2L_j$. The mean values, $\{L_j\}_{j=1}^5$, are uniformly distributed on a log scale over the amplitude range of the MBPNL output. The randomness in the values of the j th level across the cochlear channels simulates the fact that diameters and synapse-connection sizes of fibers innervating the same side of different IHC's along the cochlear partition have a certain amount of intrinsic variability (which is characteristic of most physiological systems). The model omits

certain properties of the auditory-nerve function. By modeling the nerve-fiber firing activity as a point process produced by a level-crossing detector, the probabilistic nature of the neural firing mechanism is essentially neglected. Thus, the level crossings should be interpreted as the **combined** firing activity of a collection of fibers originating in different IHC's located in a range along the basilar membrane small enough to ensure similar cochlear tuning characteristics.

The outputs of the level-crossing detectors represent the discharge activity of the auditory-nerve fibers in terms of a 950-dimensional point process (190 channels times five levels per channel). Fig. 7 shows simulated auditory-nerve activity for the first 60 ms inside the vowel [ah] in the word "gob." The abscissa represents time and the ordinate represents the characteristic frequency of the IHC channels. Note the logarithmic scale of the characteristic frequency, which represents

the place-to-frequency mapping on the basilar membrane. In the figure, a level-crossing occurrence is marked as a single dot, and the output activity of each level-crossing detector is plotted as a separate trace. Each IHC channel contributes five parallel traces, with the lower trace representing the lower-threshold level-crossing detector. If the magnitude of the filter's output signal is low, only one level will be crossed, as is the case for the very top channels of Fig. 7. However, for large signal magnitudes, several levels will be activated, creating a "darker" area of activity.

E. Measure of Synchrony and Instantaneous Rate Across the Simulated Fibers

The 950-dimensional point process at the output of the level-crossing detectors serves as the input to the second stage of the EIH. With the purpose of achieving human performance, this point process, which simulates the auditory nerve firing activity, would have to be processed following principles derived from properties of post auditory-nerve nuclei. Unfortunately, such information is not available at present. Our approach, therefore, is to apply processing principles that are motivated by observed properties of actual auditory nerve response.

Measurements of firing responses of cats' auditory-nerve fibers (e.g., [3]–[6], [29], [37]) show a significant difference between the properties of the firing patterns of low CF (say, below 1000 Hz) and high CF fibers. This difference is determined mainly by the mechanical properties of the basilar membrane. At low CF's, harmonics are resolved with high precision and neural discharges of auditory nerve fibers are phase locked to the underlying driving component. That is, synchrony (between neural discharges and the underlying driving component) is maintained. At high CF's, frequency resolution is poor and the phase-locking of the discharges is greatly reduced. The instantaneous rate of firing, however, conveys temporal information with fine time resolution.

It is also evident from these measurements that as the sound pressure level increases, more fibers fire in **coherence** with certain temporal properties of the stimulus waveform. For this reason, we suggest the spatial (tonotopic) extent of such coherent activity as a measure of the perceptual importance of the underlying temporal events. The extent of coherent activity signifies different properties of the underlying stimulus, depending on frequency. At low frequencies, it is the extent of coherent firing activity **phase-locked** to an underlying resolved component. At high frequencies, it is the extent of coherent **instantaneous rate** of the firing activity driven by a wideband, unresolved (in frequency) temporal event. Obviously, there is no distinct boundary between these auditory nerve regions. Rather, the change in properties is gradual (e.g., [18]).

In the model, the amount of coherent neural activity across the simulated fiber array is measured by determining the similarity in the short-term interval probability density functions of individual level-crossing detectors. An estimate of the interval probability density function of a given level can be obtained by computing a histogram of the intervals from the point process data produced by the level-crossing

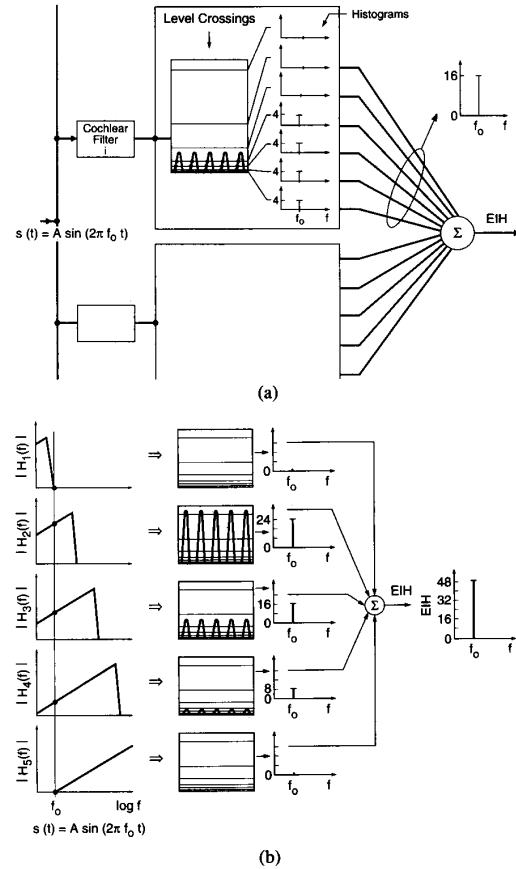


Fig. 8. (a) The contribution of the i th channel to the Ensemble Interval Histogram for an input signal $s(t) = A \sin(2\pi f_0 t)$. Only the lowest four level-crossing detectors are contributing nonzero histograms to the ensemble. (b) The contribution of five successive channels $|H_i(f)|$, $i=1,2,\dots,5$, to the EIH, for an input signal $s(t) = A \sin(2\pi f_0 t)$. Channel i contributes to the f_0 bin of the EIH provided that $A|H_i(f_0)|$ exceeds any of the level-crossing thresholds.

detector. Only intervals between successive upward-going level crossings are considered. Since we prefer an auditory representation in the frequency domain, the histogram of the **inverse intervals** is computed. This is accomplished by distributing the reciprocal of the intervals (i.e., converting the intervals into units of frequency) in a histogram consisting of successive bins, ranging from 0 to, say, 4000 Hz. (We shall discuss the considerations involved in choosing the bin allocation, the window length and the number of bins momentarily). The similarity between all individual interval probability density functions is measured by collecting the individual histograms into one ensemble interval histogram (EIH) (via summing the corresponding histogram bins across all levels and all channels). The resulting representation is the EIH. The tonotopic extent of a coherent neural activity generated by an underlying temporal event is encoded as the magnitude of the corresponding bin in the EIH.

To illustrate this point, consider the case where the input signal is $s(t) = A \sin(2\pi f_0 t)$. First, consider the channel with CF equal to f_0 (see Fig. 8(a)). For a given amplitude A ,

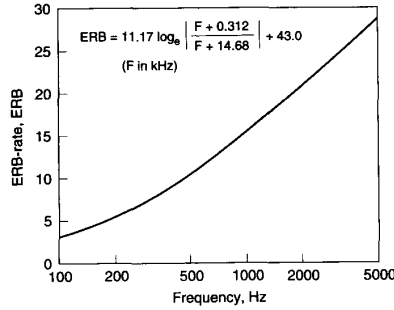


Fig. 9. ERB-rate of the human auditory filters, as a function of CF. Suggested by [26], based on psychophysical experiments. See text for a definition of ERB.

the cochlear filter output will activate only part of the level-crossing detectors, depending on the value of A . For a given detector, the time interval between two successive positive-going level crossings is $1/f_o$. Since the histogram is scaled in units of frequency, this interval contributes a count to the f_o bin. For the input signal in the illustrated example, all of the intervals are the same, resulting in a histogram where the magnitude of each bin, save one (f_o), is zero. As the signal amplitude A increases, more levels are activated. As a result, this cochlear filter contributes additional counts to the f_o bin of the EIH. Since the crossing levels are equally distributed on a log-amplitude scale, the magnitude of any EIH bin is related, in some fashion, to decibel units. However, this relation is not a straightforward one since there are several sources contributing counts to the f_o bin in a nonlinear manner. Fig. 8(b) shows an input signal $s(t) = A \sin(2\pi f_o t)$ driving five adjacent cochlear channels. For the sake of simplicity, we assume that the filters are linear, with an amplitude response $|H_i(f)|$ and a phase response $\phi_i(f)$, $i=1,2,\dots,5$. Due to the shape of the filters, more than one cochlear channel will contribute to the f_o bin. In fact, all the cochlear filters which produce $s_i(t) = A|H_i(f_o)| \sin(2\pi f_o t + \phi_i(f_o))$ will contribute to the f_o bin of the EIH, provided that $A|H_i(f_o)|$ exceeds any of the level crossing thresholds. In Fig. 8(b) only cochlear filters 2, 3 and 4 are contributing nonzero histograms to the EIH. The number of counts is different for each filter, depending on the magnitude of $A|H_i(f_o)|$.

Two factors affect the properties of the interval histogram, the bin allocation over the frequency range and the choice of the window length. Motivated by the tonotopic organization along the auditory pathway we assign the bins according to the ERB-rate scale (shown in Fig. 9, after [26]), which is related to the psychophysical critical band assignment, or the Bark scale. In the following paragraph we will define the ERB-rate scale.

Let $|H(f)|$ be a unimodal frequency response of a filter, and let $|H(f_o)|$ be the maximum gain of the filter, at frequency f_o . The equivalent rectangular bandwidth (ERB, in Hz) of $|H(f)|$ is defined as follows

$$\text{ERB} = \frac{\int |H(f)|^2 df}{|H(f_o)|^2}. \quad (2)$$

In words, ERB is the **bandwidth** of an hypothetical rectangular filter, with a gain of $|H(f_o)|$, such that the integral over its

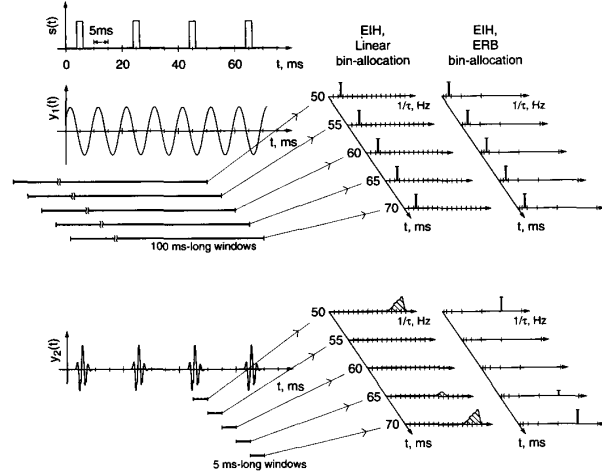


Fig. 10. An illustration of the relationship between the bandwidth characteristics of the filters, the window length and the bin allocation. Depending on frequency, EIH signifies different physical properties of the acoustic stimulus, from synchrony at low frequencies to instantaneous rate at high frequencies.

frequency response is equal to the integral over $|H(f)|$. Using psychophysical measurements of the ERB of human auditory filters, [26] derived the following quadratic fit, as a function of the center frequency of the auditory filter

$$\text{ERB} = 6.23F^2 + 93.39F + 28.52. \quad (3)$$

Where F is frequency in kHz. (Note that very similar ERB's can be derived from (2) of Greenwood, where an ERB at a given frequency corresponds to a constant distance of 0.85 mm on the basilar membrane [25]). Using the ERB of the auditory filter as a unit of measurement, Moore and Glasberg [26] suggested the **ERB-rate scale** which relate number of ERB's to frequency. This scale was obtained by integrating the reciprocal of (3), yielding

$$\text{ERB-rate} = 11.17 \log_e \frac{F + 0.312}{F + 14.675} + 43.0. \quad (4)$$

Where F is frequency in kHz. (The constant 43.0 was chosen to make the number of ERB's = 0 when $F = 0$.) To summarize, (3) specifies the ERB of a human auditory filter at a given frequency F , and (4) determines the number of successive ERB's which covers the frequency range $[0, F]$. Using the ERB-rate scale, we quantized the frequency range $[0, 4000]$ Hz into 32 bins, which is roughly the number of ERB's in this frequency range. Momentarily, we will illustrate how the choice of bin allocation determine the properties of the EIH.

Another parameter that affects the properties of the interval histogram is the size of the observation window. Motivated, again, by the tonotopic organization along the auditory pathway, we set the window length to be inversely proportional to center frequency. That is, at time t_0 , intervals produced by a level-crossing detector located at center frequency CF_0 are collected over a window of length $\frac{10}{CF_0}$ that ends at time t_0 (see Fig. 7).

Fig. 10 illustrates the relationship between the bandwidth characteristics of the filters, the window length and the bin

allocation. The figure is organized from left to right. It shows the response of two hypothetical cochlear filters, H_1 and H_2 , to a pulse-train input, $s(t)$, with a pulse every 20 ms. The center frequency of H_1 is 100 Hz, and that of H_2 is 2000 Hz, with bandwidths of 30 Hz and 300 Hz, respectively. The bandwidths of the filters dictate the properties of their outputs. Thus, H_1 , which resolves the frequency component at 100 Hz, results in a sinusoidal output $y_1(t)$. In contrast, the output of H_2 is wider in bandwidth and follows sharp temporal changes of $s(t)$. (In the limit, with the pulse-width approaching zero, $y_2(t)$ is the impulse response of the filter). In the example of Fig. 10, EIH is produced at a uniform rate, once every 5 ms, and only interval histograms of the **zero crossings** are considered. Because of our choice of window length, zero-crossings of $y_1(t)$ are collected over a 100 ms window, and zero-crossings of $y_2(t)$ are collected over a 5 ms window. Fig. 10 shows the location of five successive windows, for $y_1(t)$ and $y_2(t)$. The interval histograms for these windows are shown in the right-hand-side of the figure. The figure shows typical interval histograms for two choices of bin allocation, linear (with, say, 128 bins over [0, 4000] Hz and a fine frequency resolution) and ERB (with 32 bins over the same frequency range). In the case of linear bin allocation, the narrow-band signal $y_1(t)$ contributes identical intervals only to the 100-Hz bin. And since the window length of H_1 is much longer than the frame rate, the interval histograms hardly change from frame to frame. In contrast, the wide-band signal $y_2(t)$ contributes intervals of different values, resulting in histograms that extend over several bins. And since the window length of H_2 is similar to the frame rate, the histograms change rapidly from frame to frame, demonstrating high temporal resolution. In the case of the ERB bin allocation, bins at low frequencies are narrow, resulting in fine frequency resolution, similar to the frequency resolution of the histograms with linear bin allocation. However, bins at high frequencies are wide—covering a filter bandwidth (e.g., 300 Hz at CF=2000 Hz)—and a frequency range of one ERB bin contains several linearly allocated bins. Therefore, the interval count at this ERB bin equals to the **sum** of intervals over all the linearly allocated bins at that frequency range. In other words, at time t_0 , ERB bins at high frequencies contain the overall number of intervals collected over the window, irrespective of the shape of the interval pdf. Therefore, we view the changes in time at the high frequency bins as a measure of instantaneous rate.

Using the example of Fig. 10, let us consider now the properties of the **ensemble** of interval histogram, collected over several successive filters. If we consider the filters surrounding H_1 , they all resolve the 100 Hz frequency component and contribute intervals to the same bin (located at 100 Hz). Therefore, the magnitude of the EIH at this bin can be viewed as a measure of the number of level-crossing detectors and the number of successive filters that are synchronized (or phase-locked) to the 100 Hz underlying component. If we consider the filters surrounding H_2 , they all result in output signals similar to $y_2(t)$ of Fig. 10. The change with time of the corresponding EIH bin can, therefore, be viewed as a measure of the extent of coherent instantaneous-rate activity across the array.

In summary, choosing a bin allocation and a window length that are matched to the bandwidth characteristics of the cochlear filters provide a unified representation that exhibits fine frequency resolution at low CF's (based on a measure of synchrony), and fine temporal resolution at high CF's (based on a measure of instantaneous rate). There is no distinct boundary that signals the switch from one type of measurement to the other. Rather, the change in properties is gradual.

III. USING THE EIH IN A RECOGNITION TASK

In this section we examine the extent to which the EIH is capable of mimicking human performance in a task related to speech recognition. We confine ourselves to a recognition task with a minimal cognitive load since we are aiming at measuring the performance of the EIH, in isolation from the other parts of the evaluation system. We first identify a psychophysical experiment that addresses such a task, the diagnostic rhyme test (DRT) [33] and collect data on how accurate a human listener is in performing the task. Next, we simulate the DRT procedure. The auditory periphery is replaced by the EIH and the higher auditory elements by an automatic speech recognition system especially designed to keep the errors due to the decision process to a minimum. Errors are displayed in terms of a distribution among six phonemically distinctive features. The error patterns of the simulated procedure are compared with those of the human, to provide quantitative measure on how close the model performance is to the actual human performance. The DRT is described in Section III-A, the simulated DRT in Section III-B and the experimental results in Section III-C.

A. Brief Description of Voiers' DRT

The psychophysical experiment that we selected is the one used in the standard DRT, suggested by Voiers [33]. In general, the DRT test attempts to evaluate how well phonemic information is perceived by a human listener. The test is divided into two parts, measuring the human performance (via psychophysical experiments) and deriving an intelligibility score. For our purposes only the first part, i.e., the data collection, is relevant. As we shall see momentarily, the test is appropriate for our needs for two reasons. First, from the acoustic point of view the DRT database spans the speech subspace associated with initial diphones in a uniform manner. Hence, the performance of the EIH is examined in terms of its capability to represent all possible cells in that speech subspace. And second, it allows us to separate the effects of the auditory periphery from those due to cognition.

Voiers' DRT database covers initial diphones of spoken words of the consonant-vowel-consonant (CVC) type. Table I shows the list of words used in the DRT. The list consists of 96 pairs of confusable words spoken in isolation. Words in a pair differ only in their initial consonants. The consonants are equally distributed among six phonemic distinctive features (16 word-pairs per feature) and among eight vowels. The feature classification follows the binary system suggested by

TABLE I
STIMULUS WORDS USED IN THE DRT

Voicing		Nasality	Sustention
Voiced—Unvoiced	Nasal—Oral	Sustained—Interrupted	
veal—feel	meat—beat	vee—bee	
bean—peen	need—deed	sheet—cheat	
gin—chin	mitt—bit	vill—bill	
dint—tint	nip—dip	thick—tick	
zoo—Sue	moot—boot	foo—pooh	
dune—tune	news—dues	shoes—choose	
voal—foal	moan—bone	those—doze	
goat—coat	note—dote	though—dough	
zed—said	mend—bend	then—den	
dense—tense	neck—deck	fence—pence	
vast—fast	mad—bad	than—Dan	
gaff—calf	nab—dab	shad—chad	
vault—fault	moss—boss	thong—tong	
daunt—taunt	gnaw—daw	shaw—chaw	
jock—chock	mom—bomb	von—bon	
bond—pond	knock—dock	vox—box	
Compactness		Sibilant	Graveness
Compact—Diffuse	Sibilated—Unsibilated	Grave—Acute	
yield—wield	zee—thee	weed—reed	
key—tea	cheep—keep	peak—teak	
hit—fit	jilt—gilt	bid—did	
gill—dill	sing—thing	fin—thin	
coop—poop	juice—goose	moon—noon	
you—rue	chew—coo	pool—tool	
ghost—boast	Joe—go	bowl—dole	
show—so	sole—thole	fore—thor	
keg—peg	jest—guest	met—net	
yen—wren	chair—care	pent—tent	
gat—bat	jab—dab	bank—dank	
shag—sag	sank—thank	fad—thad	
yawl—wall	jaws—gauze	fought—thought	
caught—taught	saw—thaw	bond—dong	
hop—fop	jot—got	wad—rod	
got—dot	chop—cop	pot—tot	

TABLE II
STIMULUS WORDS USED IN THE DALT

Voicing		Nasality	Sustention
Voiced—Unvoiced	Nasal—Oral	Sustained—Interrupted	
seethe—seed	teethe—teeth	screen—screed	
dish—ditch	jib—gyp	ring—rig	
goof—goop	mood—moot	noon—nude	
both—boat	brogue—broke	moan—mode	
chief—cheep	liege—leash	gleam—glebe	
give—gib	ridge—rich	rim—rib	
soothe—sued	prove—proof	tomb—tube	
jove—job (e)	loathe—loath	gloam—globe	
rev—reb	led—let	hen—head	
calve—cab	have—half	lam—lab	
froth—fraught	jaws—joss	brawn—broad	
slav—slob	hodge—hotch	nom—nob	
flesh—fletch	peg—peck	gem—jeb	
path—pat	lathe—lath	fan—fad	
frothe—fraud	flaws—floss	long—log	
bosh—botch	fob—fop	bomb—bob	
Graveness		Compactness	Sibilant
Grave—Acute	Compact—Diffuse	Sibilated—Unsibilated	
peach—peak	sheave—sheathe	league—lead	
kiss—kith	skim—skin	hick—hit	
sues—soothe	rufe—ruth	fuge—feud	
poach—poke	oaf—oath	bloke—bloat	
breeze—breathe	neap—neat	creek—creep	
bridge—brig	miff—myth	sling—slim	
truce—truth	rube—rude	dog—loop	
clothes—clothe	lobe—load	rogue—robe	
bess—beth	deaf—death	mesh—mess	
badge—bag	dab—dad	lag—lad	
ross—wroth	shawm—shawn	dog—daub	
notch—knock	top—tot	cock—cot	
ledge—leg	web—wed	egg—ebb	
mass—math	raff—rath	knack—nap	
maws—mothe	trough—troth	gong—gone	
bodge—bog	sauve—swathe	chock—chop	

Jakobson, Fant, and Halle [16].³ The vowels are [ee] (like in peen) and [it] (like in bit) for high-front vowels, [eh] (like in zed) and [at] (like in fast) for high-back vowels, [oo] (like in zoo) and [oh] (like in note) for low-front vowels, and [aw] (like in boss) and [ah] (like in bond) for low-back vowels.

The psychophysical procedure is very carefully controlled. The listeners are well trained and are very familiar with the database, including the voice quality of the individual speakers. A one-interval two-alternative forced choice (1I2AFC) paradigm is used. First, the subject is presented visually with a pair of rhymed words. Then, one word of the pair (selected at random) is presented aurally and the subject is required to indicate which of the two words was played. This procedure is repeated until all the words in the database have been presented. The errors can be displayed either in terms of a confusion matrix (between consonants), or as a distribution among the six phonemic distinctive features.

³ The six features are *Voicing*, *Nasality*, *Sustention*, *Sibilant*, *Graveness*, and *Compactness*. The *Voicing* feature characterizes the nature of the source, being periodic or nonperiodic. The *Nasality* feature indicates the existence of a supplementary resonator. The terms *Sustention* and *Sibilant* are due to Voiers. They correspond respectively to the continuant-interrupted and strident-mellow contrasts of Jakobson, Fant and Halle. Finally, *Graveness* and *Compactness* represent broad resonance features of the speech sound, related to place of articulation.

Two points are noteworthy. First, since it is a discrimination test, we can categorize the DRT as a speech recognition task. And second, due to the controlled nature of the test procedure, we assume that all cognitive information needed for the discrimination task is available to the listener prior to the aural presentation. (Of course, we also assume that the subject is indeed utilizing all this information.) If this assumption is correct, an error in identifying the word is due mainly to inaccuracy in the internal auditory representation of the stimulus. Hence, the error list provided by the test reflects errors in the internal human auditory representation during the discrimination task.

B. Simulating the DRT (After [10])

In the simulation, the peripheral part of the auditory pathway is modeled by the EIH, and the cognitive process is replaced by an array of recognizers, one for each pair of words in the DRT database. The errors due to the recognition procedure should be kept to a minimum, so that the overall detected errors are due mainly to inaccuracies in the front-end representation.

1) *The Recognition System:* A recognizer in the array represents one DRT word-pair. For a test word (from a given word-pair), the recognizer makes a maximum likelihood deci-

sion between two hidden Markov (HMM) word models, one for each word in the pair. An HMM word-model is defined as a left-to-right phonemic sequence. Each state of the HMM word-model represents one phonemic unit. The recognizer used is an HMM recognizer with time-varying states, suggested by Ghitza and Sondhi [12]. In this recognizer, a state of the HMM represents one phonemic unit in terms of a time-varying mean sequence of ordered frames—a template—and a block covariance matrix that characterizes the intra-frame statistical dependence within the phonemic unit. The particular phonemic unit that was selected is a diphone. In this way the dynamic nature of coarticulation between two successive phonemes is represented accurately, and the ability to discriminate between the initial consonants of the DRT words is improved.

For every word in the DRT database, the HMM word-model has a trivial transition matrix that allows only one deterministic transition (from the first diphone to the second). For such a simple transition matrix, the HMM model reduces to a modified version of the dynamic time warping (DTW) method of speech recognition (e.g., [36]). In this framework, a word-model for a C_iVC_f word is a concatenation of two state-templates C_iV and VC_f , and the decision between the two word models is a **minimum distance** decision. The input-word and the word-models are represented in terms of ordered sequences of frames (or observation vectors). The distances between the word-model sequences and the input-word sequence are measured as defined in the DTW method: we first define $d(\mathbf{a}, \mathbf{b})$ to be the distance between any two single observation vectors, \mathbf{a} and \mathbf{b} as

$$d(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})' \sum^{-1} (\mathbf{a} - \mathbf{b}). \quad (5)$$

Here, $'$ denotes vector transpose, and \sum is a covariance matrix whose (i, j) th entry is the covariance between the i th and the j th components of the observation vector. For $\sum = I$, $d(\mathbf{a}, \mathbf{b})$ is simply the L_2 -norm of the difference between \mathbf{a} and \mathbf{b} . In terms of d , we can define the distance between the sequences \mathbf{O}^i and \mathbf{O}^j by the usual DTW procedure. Let $\mathbf{o}_m^i, m = 1, 2, \dots, M$ be the observations in sequence \mathbf{O}^i , and $\mathbf{o}_n^j, n = 1, 2, \dots, N$ the observations in sequence \mathbf{O}^j . We define $D(\mathbf{O}^i, \mathbf{O}^j)$ as

$$D(\mathbf{O}^i, \mathbf{O}^j) = \frac{1}{N} \min_{m(n)} \sum_{n=1}^N d(\mathbf{o}_n^j, \mathbf{o}_{m(n)}^i). \quad (6)$$

The mapping $m(n)$ is constrained such that $m(1) = 1$ and $m(N) = M$. Thus $D(\mathbf{O}^i, \mathbf{O}^j)$ is the average distance between corresponding observation vectors in the two sequences, after the sequence \mathbf{O}^j has been optimally warped onto the sequence \mathbf{O}^i .

2) *Simulating the One-Interval-Two-Alternative-Forced-Choice Paradigm*: To eliminate effects due to variability among speakers, the simulation is done on a speaker-dependent basis. Every speaker provides two repetitions of the DRT word-list, one for training and one for testing. As a part of the training phase, a vocabulary of diphones is obtained for each speaker by segmenting the training part of the database manually. The inventory covers all the diphones that appear

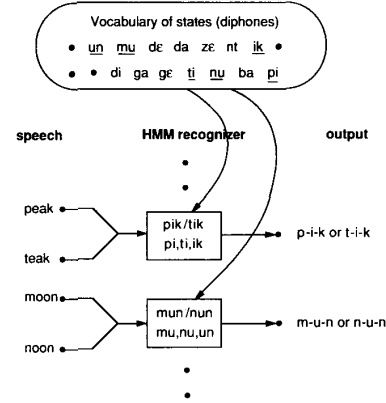


Fig. 11. An illustration of the DRT simulation procedure. To test the word-pair *peak/teak*, for example, the state models for the diphones *pi*, *ti* and *ik* are drawn from the states vocabulary, along with the appropriate transition matrix that allows only the necessary transitions. The recognizer is then presented with the word *peak*, and produces a phonemic transcription which is either *p-i-k* or *t-i-k*. If the first transcription occurs, the result of the simulated discrimination task is considered to be correct. Otherwise, an error is registered. Next, the word *teak* is tested. Identical state models and transition rules are used, and the same sequence of steps is repeated. This concludes the test for this word-pair. To test the next word-pair (e.g., *moon/noon*) the recognizer is loaded with new state models (*mu*, *nu* and *un*) and a new transition matrix, and the above procedure is repeated.

in the DRT word list. If several tokens of a particular diphone appear in the DRT word list, the diphone is represented by only one of these tokens.

The testing phase is a simulation of the 1I2AFC paradigm. For testing a particular word-pair, the recognizer is first loaded with the appropriate diphone-state models (drawn from the inventory) and transition matrices. This step simulates the effect of the visual presentation of the word-pair to the listener. Then, the two words are presented one at a time to the recognizer, analogously to the aural presentation to the listener. Based on the recognizer's phonemic transcription, it is decided whether or not the word was correctly recognized. This procedure is repeated until all the word pairs in the database have been scanned. The overall error list is displayed in terms of a distribution among the 6 phonemic features.

Fig. 11 illustrates the simulation procedure. To test the word-pair *peak/teak*, for example, the state models for the diphones *pi*, *ti* and *ik* are drawn from the state vocabulary, along with the appropriate transition matrix that allows only the necessary transitions. The recognizer is then presented with the word *peak*, and produces a phonemic transcription which is either *p-i-k* or *t-i-k*. If the first transcription occurs, the result of the simulated discrimination task is considered to be correct. Otherwise, an error is registered. Next, the word *teak* is tested. Identical state models and transition rules are used, and the same sequence of steps is repeated. This concludes the test for this word-pair. To test the next word-pair (*moon/noon*, in Fig. 11), the recognizer is loaded with new state models (*mu*, *nu* and *un*) and a new transition matrix, and the above procedure is repeated. Note that in testing the word-pair *peen/bean*, the state model for the diphone *pi* (which is required to model the word *peen*) is the same state model used previously for the word *peak*.

C. Experimental Results

1) *Signal conditions:* Three male speakers were used, two from Voiers' database (speakers RH and CH). Each speaker provided two repetitions of the DRT word-list, one for training and one for testing. The signals were lowpass filtered to 3600 Hz and sampled at an 8-kHz rate. A "noisy" version of the testing repetitions was created by adding white noise to the original ("clean") signals. The signal-to-noise ratio (SNR) was 10 dB.⁴ The noisy version was sent to Dynastat Inc. (a company established by Voiers) for the psychophysical evaluation. To comply with Dynastat's procedure, the processed words were recorded at a rate of a word every 1.3 s. For the recordings to sound continuous over time, we first set the variance of the white noise generator to a level that remained unchanged until all the words in the DRT word list had been recorded in sequence. To record a particular word, the speech signal was amplified (or attenuated) by a gain factor that was calculated in advance, to maintain the desired global SNR.

As pointed out before, the DRT simulation is performed on a speaker-dependent basis—to eliminate errors due to between speaker variability. A single-speaker simulation procedure (training and testing) is repeated for every speaker in the database.

For training, the vocabulary of diphones was created from the clean repetition of the DRT word-list assigned for training. For testing, we used the same noisy versions that were sent to Dynastat Inc.

2) *Analysis Methods:* We tested two speech representation methods, the EIH and the Fourier power spectrum. The auditory model for the EIH representation is described in Section II. The model uses 190 cochlear channels and five thresholds per channel. An EIH observation was computed once every 10 ms. The interval statistics at time t_0 are collected from all 950 (190 times 5) threshold detectors, using all simulated firing records which exist in the windows that end at time t_0 (see Fig. 7). The length of each window is 10/(CF), where CF is the characteristic frequency of the cochlear channel. The interval statistics are collected into a 32-bin vector, where the bins divide the frequency range [0, 4000] Hz according to the ERB-rate scale. The sum over the 32 bins was normalized to 1, to eliminate the effect of "loudness."

For the Fourier power spectrum, an eleventh-order cepstral representation was computed every 10 ms. That is, the feature-vector under consideration is the Fourier **spectral envelope**. At time t_0 , the cepstral coefficients are derived from the tenth-order LPC coefficients, computed from a 30 ms-long Hamming window centered at t_0 . The first cepstral coefficient (c_0) was set to 0 and only the next 10 coefficients were used. In this way, the envelope is normalized in the sense that the average value of the LPC log spectrum is 0.

⁴The SNR was defined using global measurements. First, the total energy, E_{Tot} , of the original (noise-free) word was computed. Then, the average energy per digital sample, E_{Samp} , was determined, by dividing E_{Tot} by the number of sample points in the signal. E_{Samp} was used to set the variance of a white noise generator to a level dependent on the desired global-SNR. This definition of global-SNR overestimates the actual signal to noise ratio during the consonantal segments since the magnitude of the noise is largely dependent on the amplitude of the vocalic portion of each word.

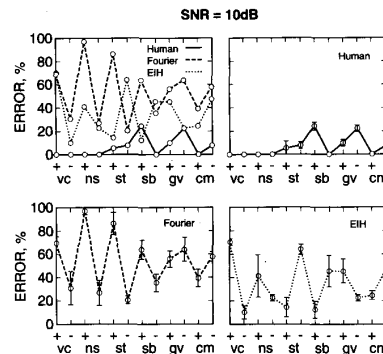


Fig. 12. Distribution of errors made by the human listener, an EIH with an ERB bin allocation and a Fourier power spectrum, in a 10 dB signal-to-noise ratio. The left-upper plot is a superposition of the other three plots, excluding the confidence intervals. The abscissa of every plot indicates the six phonetic features: "vc" is for Voicing, "ns" for Nasality, "st" for Sustention, "sb" for Sibilation, "gv" for Graveness, and "cm" for Compactness. The "+" sign stands for attribute present and the "-" sign for attribute absent. The line connecting the measurements is only for display purposes, to enable the reader to distinguish between error patterns that belong to a particular parameter value. The noise is additive and white, and the signal-to-noise ratio is defined using global measurements (see text). Note that an error value of 100% in a given phonemic category means that all 16 words in the category were mistakenly identified as their DRT counterparts.

As for the recognition system, the distance measures d and D of (5) and (6) were used, respectively. For d , we used $\sum = I$ (i.e., d is the L_2 -norm of the difference vector). The same distance measures were used for both analysis methods. We shall discuss the implications of this choice for d and D in Section V.

3) *Results:* The raw data that summarizes the outcome of one experimental run (i.e., one speaker) are organized in the form of a matrix with 12 rows and 16 columns. Each row represents a phonemic category (attribute present or attribute absent for each of the six phonemic features—see Table I), and each column represents a word in the DRT word-list, associated with the corresponding row. For the simulated procedure, an entry in the matrix is a binary number, a 0 (for a correct answer) or 1 (for an error). For the psychophysical procedure, the value of a matrix element is an integer number between 0 and 8 and it indicates the number of listeners who made a mistake in identifying the corresponding word (8 is the number of listeners participating in the test). Calculating statistics across the columns, we computed the average error and the 95% confidence interval for every row (i.e., for every phonemic category we averaged across all vowels and all listeners). Then, we averaged across all speakers.

Fig. 12 shows the resulting error distribution. The abscissa of every plot indicates the 12 phonemic categories: "vc" is for Voicing, "ns" for Nasality, "st" for Sustention, "sb" for Sibilation, "gv" for Graveness, and "cm" for Compactness. The "+" sign stands for attribute present and the "-" sign for attribute absent. The figure contains four plots, where the left-upper plot is a superposition of the other three plots, excluding the confidence intervals. Note that the line connecting the measurements is only for display purposes, to enable the reader to distinguish between error patterns that belong to a particular parameter value.

Three points are noteworthy. First, the human observer performs much better than the EIH and Fourier power spectrum. Second, on the average EIH is more robust to noise than the Fourier power spectrum, in agreement with previous reports (e.g., [9]). And third, errors made by the Fourier power spectrum analyzer are mainly in the presence of Voicing, Nasality, Sustention and Sibilant, while errors made by the EIH are mainly in the presence of Voicing and the absence of Sustention. (We will show later (in Section V-A) that other measures of interval statistics lead to improved results, and that the EIH performance can be brought closer to that of a human observer (see Fig. 16).)

IV. USING THE EIH TO PREDICT MEAN-OPINION-SCORE (MOS) OF SPEECH CODERS

In this section we examine the extent to which the EIH is capable of mimicking human judgement of speech quality. We follow the methodology used in Section III: first, we identify a psychophysical experiment that addresses this task. Then, we collect data on how the human listener scores the quality of synthetic speech produced by different speech coding systems. Finally, we use the EIH as a basis for a system aiming at predicting the human scores.

The psychophysical experiment that was selected is the MOS test, which is widely used to assess the quality of speech coders. It is a subjective test that can be categorized as a rating procedure. Subjects are presented, once, with a speech sentence and are requested to score its quality using a scale of five grades. The grades (and their numerical aliases) are Excellent (5), Good (4), Fair (3), Poor (2) and Bad (1). The MOS is the mean score, averaged over the database and the subjects.

It is obviously interesting to find out how adequate EIH is in predicting MOS. Since EIH is computed from simulated auditory nerve responses modeled in considerable detail, we may hypothesize that it contains only perceptually-relevant speech information. If this hypothesis holds, perceptual differences between the original and the coded speech should be realistically reflected in the EIH domain. To examine this hypothesis, we mimicked the MOS test by measuring the L_2 -norm of the difference between the EIHs of the original and the synthetic speech. For reference purposes we also compared the original and the coded speech in the cepstral domain, using the L_2 -norm metric. (We are aware of numerous methods for an objective quality assessment of coded speech (e.g., [20], [28], [35]), but a detailed comparison is beyond the scope of this paper).

A. Database

We collected MOS scores for 18 different speech coding conditions. The assessed coders were a 64 kbit/s μ -law PCM, two versions of a 32 kbit/s ADPCM, a 16 kbit/s LD-CELP, two versions of a 8 kbit/s CELP and three versions of a 6.6 kbit/s CELP.⁵ The ADPCM and LD-CELP coders were tested

in four tandem configurations.⁶ The speech material contained sentences spoken by four male and four female speakers, each contributing two sentences.

B. Analysis Methods

For each coding condition, the coded waveform was first aligned (in time) with the original waveform. The EIH representation was computed as described in Section III-C-2. The original and the coded speech signals were sampled synchronously. At time t_0 , an L_2 -norm of the difference between the normalized EIHs of the original and the coded speech was calculated first. This "frame" distance was then weighted by the "loudness" of the original speech at this time instant. (Our definition of loudness is the sum over the 32 bins, prior to normalization). The "EIH-distance" between a coded speech sentence and the original is defined as the mean value of the weighted frame distances over the entire sentence.

The cepstral representation was computed as described in Section III-C-2. The original and the coded speech signals were sampled synchronously. At time t_0 , an L_2 -norm of the difference between the normalized cepstra of the original and the coded speech was calculated first. This "frame" distance was then weighted by the c_0 of the original. The "Cepstral-distance" between a coded speech sentence and the original is defined as the mean value of the weighted frame distances over the entire sentence.

C. Experimental Results

Figs. 13 and 14 summarize the results. The abscissa indicates the MOS and the ordinate the EIH-distance (see Fig. 13) or the Cepstral-distance (see Fig. 14). We grouped the results according to coder identity (left) or according to tandem configuration (right). A particular coding condition can be located by a cross-alignment of the left and the right plots. MOS values, by definition, are between 1 (bad) and 5 (excellent). For the coders we used, EIH-distance values are between 0 (perfect match to original) and 8, and Cepstral-distance values are between 0 (perfect match) and 6. In our database, for example, μ -PCM has an MOS of 4.2, an EIH-distance of 1.2 and a Cepstral-distance of 1.2.

Fig. 13 shows that EIH successfully predicts MOS of coders with a bit-rate of 16 kbit/s and above. For this group of coders, EIH-distances and MOS are linearly related. However, MOS of the CELP coders are poorly predicted. These coders stand as a separate group in the MOS-EIH plane, with EIH-distances that are above the linear regression line that fits the other coders. Below, in Section V-A, we shall discuss possible reasons for this behavior.

Fig. 14 shows the extent to which the Cepstral-distance predicts the MOS. The scatter among the coders with a bit-rate of 16 kbit/s and above is higher than the scatter observed in Fig. 13. However, the group of CELP coders is properly merged with the rest of the coders.

⁵ μ -law PCM and ADPCM are described by [17]. LD-CELP is described by [2]. A description of CELP can be found in [21].

⁶ A tandem connection is a configuration used in communication networks, where several speech coders are connected in series. In a 2-tandem connection of ADPCM coders, for example, the synthetic speech produced by one ADPCM coder is passed through a second ADPCM coder. The synthetic speech produced by the second coder is the signal used for testing. See [17].

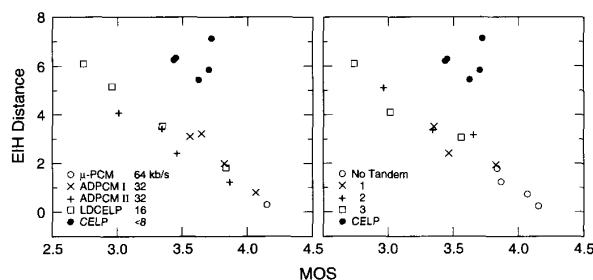


Fig. 13. EIH-distance versus MOS. The results are grouped according to coder identity (left) or according to tandem configuration (right). A particular coding condition can be located by a cross-alignment of the left and the right plots.

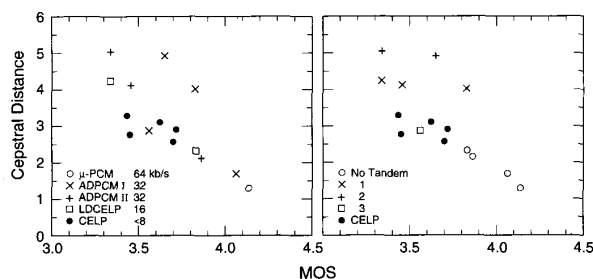


Fig. 14. Cepstral-distance versus MOS. Figure legend is as in Fig. 13.

V. DISCUSSION

A. Possible Reasons for Deficiency in Performance

The evaluation process of Sections III and IV demonstrates that generally, performance is improved by replacing a conventional speech representation by the auditory model, but is still short of achieving human performance. The shortcomings occur, in particular, in tasks related to low bit-rate coding and to speech recognition. In this section we discuss possible reasons for this deficiency. From the outset, we assume that EIH is an appropriate representation for an acoustic signal. This is so since the EIH is constructed from a detailed simulation of the human auditory-nerve firing patterns, using rules derived from general properties of observed auditory nerve activity in the cat (e.g., synchrony in low CF's, instantaneous rate in high CF's). The deficiency, therefore, may be the result of inappropriate use of the EIH in the context of speech processing.

The experiments in Section IV demonstrate that using the L_2 -norm in the EIH domain is appropriate only if the coded speech is of high quality (bit-rate of over 16 kbit/s). However, this same measure fails to predict MOS of CELP coded speech. What is the reason for this behavior? Fig. 15 shows the representation of the vowel [ir] (like in bit) produced by a male speaker. The upper plot contains the high-resolution Fourier magnitude spectrum computed by a 128-point discrete Fourier transform, using a 20-ms Hamming window. It also contains the envelope fit achieved by an tenth-order all-pole polynomial-fit. The plot uses a log-frequency/decibel scale. The lower plot is the EIH representation of the same signal. The EIH plot is in a linear-ERB/linear-EIH scale. (Note that the abscissa spans the frequency range [250, 4000] Hz. In this

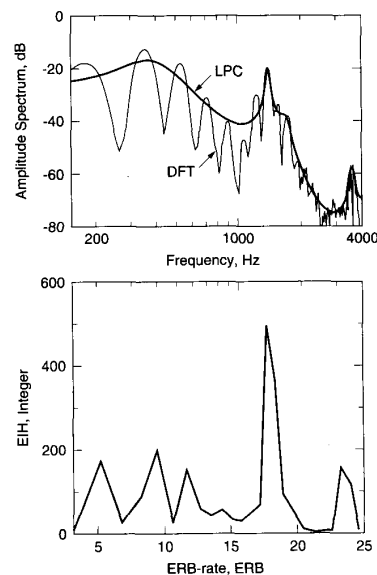


Fig. 15. Amplitude spectrum and EIH representations of the vowel [ir] (like in bit) produced by a male speaker. The upper plot contains the high-resolution Fourier magnitude spectrum computed by a 128 point discrete Fourier transform, using a 20-ms Hamming window. It also contains the envelope fit achieved by an tenth-order all-pole polynomial-fit. The plot uses a log-frequency/decibel scale. The lower plot is the EIH representation of the same signal. The EIH plot is in a linear-ERB/linear-EIH scale. Note that the abscissa spans the frequency range [250, 4000] Hz. In this frequency range, the ERB-rate scale overlaps with the log-frequency scale.

frequency range, the ERB-rate scale overlaps with the log-frequency scale.) Since it reflects the tonotopic organization along the auditory pathway, the ERB-rate is a natural scale for the EIH, and was used during the experiments of Sections III and IV. The EIH spans the first 1000 Hz by approximately half the number of bins, and with fine frequency resolution. This results in a detailed representation of the first few harmonics, as shown in the figure. In turn, an L_2 -norm of the difference between two such EIHs is very sensitive to the value of the first few individual frequency components. Consequently, EIH is a good predictor of MOS only if the two speech observations under comparison are similar enough (such that differences are localized in time and frequency). Indeed, this is the case for those coders in our database with a bit-rate of 16 kbit/s and above. For the CELP coders, however, differences are not localized. Saving bits in low bit-rate coders relies on perceptual tolerance to changes that extend over durations of several frames. In order to achieve synthetic speech with an acceptable quality, these coding schemes are designed to spread the differences between the original speech segment and the synthesized segment smoothly over the entire segment. Consequently, a comparison between two single observation taken at the same time instant, one from the original waveform and one from the CELP coded waveform, will result in an irrelevant large errors due to the misalignment of the harmonics. Note that the Cepstral-distance is insensitive to the location of individual harmonics since the L_2 -norm is measured between the spectral envelopes (via the 11th-order cepstra).

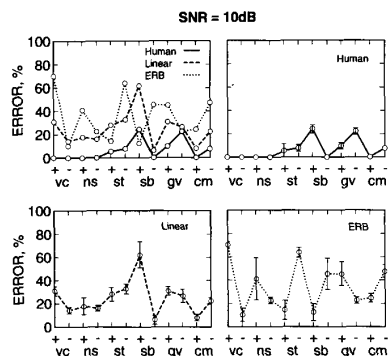


Fig. 16. Distribution of errors made by the human listener, an EIH with an ERB bin allocation (from Fig. 12) and an EIH with a linear bin allocation (from [10]) in a 10 dB signal-to-noise ratio. Figure legend is as in Fig. 12.

A similar observation can be made when examining the recognition experiment of Section III. Speech recognizers rely on measuring a distance between phonemic/articulatory information that extend over durations of 50–100 ms, very much like low bit-rate coders. In the simulation of the 112AFC procedure we compared two speech segments of such duration, the state-template C_iV and the input segment, each represented as an ordered sequence of single observations. Even when these speech segments carry the same phonemic information, they usually differ in their length and in the way this information is manifested over the segment duration. Two sources of errors are noteworthy. First, an L_2 -norm type of measure between two EIH observations is inappropriate. For recognition tasks, we seek a measure that will reflect the **phonemic** distance between the speech observations. The exact location of individual harmonics is not relevant to the identity of a phoneme, and large errors due to misalignment of the harmonics will mask the useful error information. (This observation holds even for speaker-dependent speech recognition tasks, such as the one examined in Section III). Fig. 16 demonstrates this point. The right-upper plot (performance of a human listener) and the left-lower plot (EIH with an ERB bin allocation) are from Fig. 12. The right-lower plot (EIH with a linear bin allocation) is from [10]. In that study, the error distribution was obtained using the same method described in Section III. The parameters of the MBPNL cochlear filters and the allocation of the level-crossing detectors were as in Section III-C-2, but the interval-histogram bins were assigned differently, using 128 linearly spaced bins ranging from 0 to 4000 Hz. Mathematically, EIH was treated as a log spectrum (recall that the levels are equally distributed on a log scale) and the feature-vector was a 25th-order cepstral fit, obtained by computing the inverse DFT of the EIH and by truncating the resulting cepstral series.

Although the overall number of errors of the two EIH configurations in Fig. 16 is almost the same, the error distribution of the EIH with the linear bin allocation better predicts the error distribution of the human observer. Since the same recognizer have been used in both experiments, the improvement in performance is due to the difference between the feature vectors which is mainly dictated by the

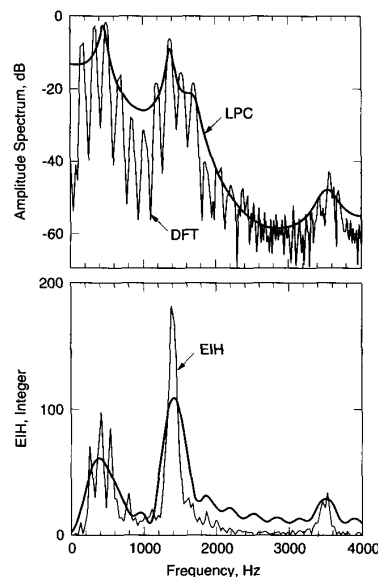


Fig. 17. A replication of Fig. 15, in a linear-frequency scale. The upper plot contains the same high-resolution Fourier magnitude spectrum and the same envelope fit, plotted over a linear-frequency/decibel scale. The lower plot shows the EIH representation with a linear bin allocation (128 bins over [0, 4000] Hz) and the corresponding envelope fit, achieved by computing the inverse DFT of the EIH and truncating the cepstral series at c_{25} .

nature of the bin allocation (all other parameters of the EIH configurations are identical). Using a linear-frequency scale turns out to be advantageous for recognition since only one quarter of the number of bins is used for the first 1000 Hz. Therefore, a cepstral fit of the EIH becomes effective in smoothing out individual harmonics and, in turn, the L_2 -norm is no longer sensitive to their location. (This is illustrated in Fig. 17 which replicates Fig. 15, in a linear-frequency scale. The upper plot contains the same high-resolution Fourier magnitude spectrum and the same envelope fit, plotted over a linear-frequency/decibel scale. The lower plot shows the EIH representation with a linear bin allocation (128 bins over [0, 4000] Hz), and the corresponding twentyfifth-order cepstral envelope fit. Obviously, the harmonic information is no longer presented in the envelope fit of the EIH). At this point, the reader may wonder why we prefer the ERB-rate scale over the linear-frequency scale, even though the latter results in a better fit to human performance. We study the EIH with an ERB bin allocation because it represents more realistically the actual organization of the auditory periphery. Working with this scale, although leaving us with an unresolved question (namely, what is the appropriate measure between two such EIH observations), provides the appropriate framework for understanding the strategy of the human observer in performing this task.

The second noteworthy source of errors is the way we measure the distance between two speech segments that extend over durations of, say, 50–100 ms. In the simulation of the 112AFC we used $D(O^i, O^j)$ of (6) to measure the distance between two sequences O^i and O^j , where $D(O^i, O^j)$ is the average distance between corresponding single observation

vectors in the two sequences, after the sequence O^j has been optimally warped on to the sequence O^i . But, recent study on the perceptual effects of modifications in the time-frequency domain [11] indicates that a human observer utilizes a much more complex strategy. Next, we briefly summarize that study.

B. The Tiling Study (After [11])

The study was aimed at understanding how auditory nerve activity is integrated over intervals of 50–150 ms, and over diphones in particular. From the outset, it was assumed that distinct peripheral auditory functions do exist which operate on different sections of the auditory nerve. Due to the tonotopic organization of the auditory nerve fibers, the nature of these auditory functions, as well as their accuracy of performance, can be inferred from psychophysical data on the perception of acoustic information contained in different time-frequency regions. Obtaining this psychophysical data was the subject of the study.

A database was used which covers the speech subspace associated with initial and with final diphones. It comprised two parts, Voiers' old DRT database for initial diphones (described in Section III), and a new database, also designed by Voiers and termed diagnostic alliteration test (DALT), for final diphones [34]. Tables I and II show the lists of words in the DRT and the DALT, respectively. Like the DRT database, the DALT database also consists of 96 pairs of confusable words spoken in isolation. Each word is of the CVC type, and words in a pair differ only in their **final** consonants. The diphones are equally distributed among the same six phonemic features and the same eight vowels used in the DRT database.

Modified versions of the database were generated by introducing well-defined distortions into pre-selected time-frequency regions. Figs. 18–20 illustrate the procedure that was used to generate the modifications. Although the illustrations (examples and figures) relate to modifications of initial diphones in the DRT database, their analogous counterparts are appropriate for illustrating modifications of final diphones in the DALT database. Fig. 18 shows the waveforms and the wideband spectrograms of the DRT word-pair *shock/mock* spoken by the same male speaker. Since the words differ only in their initial diphone, the main difference between the spectrograms is in the time-frequency region associated with those diphones, i.e. the region bounded by the bold lines containing the initial consonant (either [S] or [m]) and the left, coarticulated part of the vowel [a].

Fig. 19 shows a diagram of the time-frequency region occupied by a spoken CVC word, like the word *shock* or *mock* of Fig. 18. The time-frequency region of interest (i.e., the initial diphone) is subdivided into 6 "tiles." The frequency boundaries (from the bottom up) are 0 Hz, 1000 Hz, 2500 Hz, and the highest frequency in the band, say 4000 Hz. The time landmarks are (from left to right) the beginning of the word ($t = 0$), the transition from the initial consonant to the vowel, and the mid-point of the vowel.

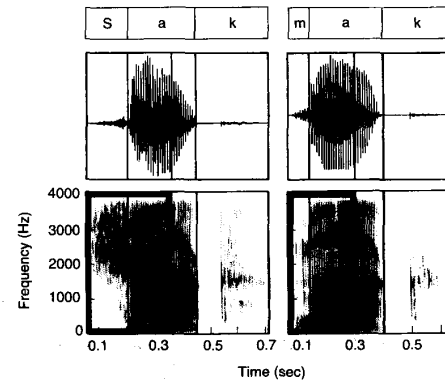


Fig. 18. The waveforms and the wideband spectrograms of the DRT word-pair *shock/mock* spoken by a male speaker. Also indicated are the boundaries between the phonemes. Since the words differ only in their initial diphone, the main difference between the spectrograms is in the time-frequency region associated with that diphone, i.e., the region bounded by the bold lines containing the initial consonant (either [S] or [m]) and the left, coarticulated part of the vowel [a].

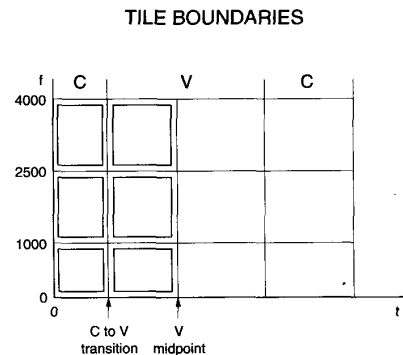


Fig. 19. A diagram of the time-frequency domain occupied by a spoken CVC word. The time-frequency region of the initial diphone is divided into six "tiles." The frequency boundaries (from the bottom up) are 0 Hz, 1000 Hz, 2500 Hz and 4000 Hz. The time landmarks are (from left to right) the beginning of the word ($t = 0$), the transition from the initial consonant to the vowel, and the midpoint of the vowel.

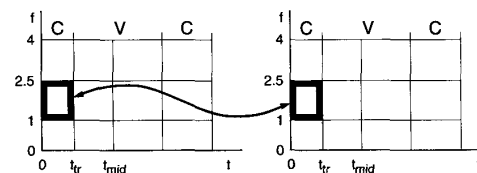


Fig. 20. A diagram of the time-frequency domain occupied by a prototype DRT word-pair, where the region corresponding to the initial diphones are divided into six tiles each. Illustrated is the interchange of the consonantal part of the second frequency band.

Fig. 20 shows a diagram of the time-frequency domain occupied by a prototype DRT word-pair, where the regions corresponding to the initial diphones are divided into six tiles each. To modify the database, an **interchange** operation was performed between the words in the pair. An interchange was defined as a "cut and paste" procedure, where a particular tile from one word was interchanged with the same tile of the opposite word. Once a particular interchange was chosen,

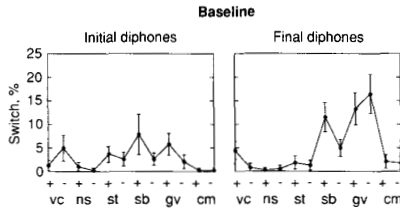


Fig. 21. The average human performance and the 95% confidence interval for the baseline (i.e., no interchange) versions of the DRT (left) and DALT (right). The abscissa is as in Fig. 12. The ordinate is termed "switch," and it represents the number of words in a phonemic category that, when played to the listener, were judged to be the opposite word in the word pair (i.e., the listener "switched" to the opposite category). The switch is represented in percents, relative to 16 (the total number of words per phonemic category).

it was applied to the entire database. The example of Fig. 20 illustrates the interchange of the consonantal part of the second frequency band. Of course, one can interchange more than one tile at a time.

The modified versions were tested using Voiers' psychophysical discrimination task (see Section III). The responses of the subjects were displayed as a distribution along Jakobson, Fant and Halle's phonemic dimensions, very much like the displays used in Fig. 12. Detailed discussion of the experimental results is beyond the scope of this paper (for that, see [11]). Highlights of the results are plotted in Figs. 21–23. Fig. 21 shows the average human performance and the associated 95% confidence interval (among the eight listeners) for the baseline versions (i.e., no interchange) of the DRT (left) and DALT (right). The abscissa is as in Fig. 12. The ordinate is termed "switch," and it represents the percentage of words in a phonemic category that, when played to the listener, were judged to be the opposite word in the word pair (i.e., the listener "switched" to the opposite category).

Fig. 22 shows the human response for three selected modifications of the DRT database: an interchange of the first frequency band, an interchange of the second frequency band and an interchange of the third frequency band of the entire diphone. The abscissa is as in Fig. 12. As for the ordinate, from the measured switches for the selected interchange condition we subtracted the number of switches for the baseline version (see Fig. 21, left). We termed the ordinate " Δ switch," since it represents the *additional* number of switches, relative to the baseline version, that occurred due to the particular interchange operation. The upper right plot shows the amount of Δ switch, in percent, when the first frequency band (i.e., [0, 1000] Hz) of the initial diphone is interchanged. The lower left plot is for the case when the second frequency band (i.e., [1000, 2500] Hz) of the initial diphone is interchanged, and the lower right plot is for the interchange of the third frequency band (i.e., [2500, 4000] Hz) of the initial diphone. Several observations are noteworthy. The phonemic dimensions *Voicing* and *Nasality* are sensitive to an interchange of the first frequency band of the diphone, *Graveness* and *Compactness* to an interchange of the second frequency band of the diphone, and *Sibilant* to the interchange of the third frequency band of the diphone. *Sustention* is equally sensitive to interchanges of the diphone information in the second and the third frequency bands. Such

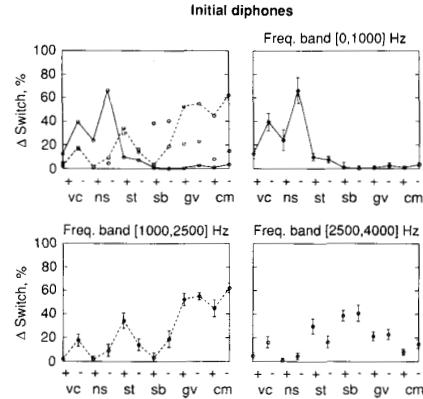


Fig. 22. Human performance under the interchange of each frequency band over the entire diphone, on the DRT database. The upper left plot is a summary of the other three plots, with the confidence-interval bars omitted. The abscissa is as in Fig. 12. The ordinate is termed " Δ switch," since it represents the *additional* number of switches, relative to the baseline version, that occurred due to the particular interchange operation. The upper right plot shows the amount of Δ switch, in percent, when the first frequency band of the initial diphone is interchanged. The lower left plot is for the interchange of the second frequency band, and the lower right plot is for the interchange of the third frequency band of the initial diphone. Notice that *Voicing* and *Nasality* are strongly correlated with the first frequency band of the diphone, *Graveness* and *Compactness* with the second frequency band of the diphone, and *Sibilant* with the third frequency band of the diphone.

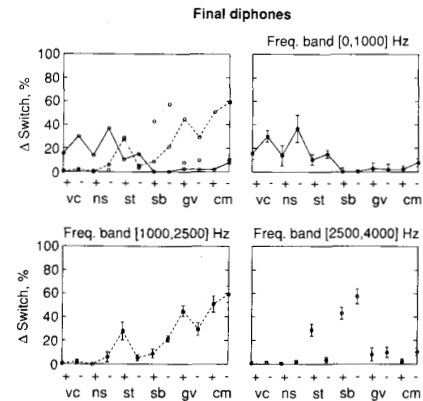


Fig. 23. Same as Fig. 22, for DALT.

a behavior is expected, since the acoustic manifestation of "interrupt" (which is one of the binary attributes of *Sustention*) is an abrupt temporal change, spread over a wide frequency range.)

Fig. 23 summarizes the human response for similar modifications of the DALT database. Somewhat surprisingly, the qualitative behavior is similar to that of Fig. 23. However, there are quantitative differences, i.e., in the numerical values of the Δ switch in various conditions.

Summarizing the results of this study, the following important observations have been noted: 1) the human observer uses different time-frequency tiles for different phoneme-discrimination tasks, 2) there is a direct mapping between phonemic/articulatory features and particular time-frequency tiles, and 3) tiles associated with diphones are far more

important than tiles associated with the consonant part or the vowel part alone. We suggest that these observations should serve as a guideline to formulate a proper measure, to replace D of (6) (and d of (5)).

VI. SUMMARY

In this paper we evaluated the adequacy of current peripheral auditory models to mimic human performance, in the context of speech recognition and speech coding. Models with such a capability are of interest since they can provide a basis for realizing effective speech processing systems. We first described our auditory model, then we outlined a methodology for a quantitative evaluation of the model accuracy, and finally we provided a critique of the current status of the model.

We focused on a particular auditory model, the EIH, that was developed using a "bottom-up" approach. Its first stage uses knowledge about the functional operation of pre-auditory-nerve elements of the auditory periphery to model, in considerable detail, human auditory nerve firing activity. In order to achieve human performance, the simulated auditory nerve firing patterns would have to be processed according to principles derived from properties of post auditory-nerve nuclei. Unfortunately, such information is not available at present. Our approach, therefore, was to process the simulated auditory nerve firing patterns according to principles that are motivated by observed properties of the actual auditory nerve response. The resulting representation, the EIH, is a measure of the spatial (tonotopic) extent of coherent activity across the simulated auditory nerve and it signifies different physical properties of the acoustic stimulus depending on frequency, from synchrony at low frequencies to instantaneous rate at high CF's. As such, EIH differs markedly from traditional speech representation methods and, consequently, exhibits quite different properties. The premise is that these properties accurately reflect properties of the internal human representation of speech.

The extent of this accuracy was measured in a quantitative manner by comparing the performance of a system that uses the EIH representation to that of a human listener while performing the same task. This methodology was used in tasks related to speech recognition and to speech coding. In general, performance was improved by replacing a cepstral speech representation by the auditory model, but was still short of achieving human performance. The evaluation process indicated that the shortcomings occur particularly in tasks related to low bit-rate coding and to speech recognition. Since schemes for low bit-rate coding rely on signal manipulations that spread over durations of several tens of ms, and since schemes for speech recognition rely on phonemic/articulatory information that extend over similar time intervals, it was concluded that the shortcomings are due mainly to a lack of perceptually related integration rules over durations of 50–100 ms. The results of the "tiling" experiment confirmed human usage of such integration, with different integration rules for different time-frequency regions depending on the phoneme-discrimination task. From the perspective of representation of speech, we conclude that "snap-shot" representations (i.e.,

representations that are local in time and frequency) provided by current auditory models are inappropriate, and that a representation based on integration over time intervals of 50–100 ms is needed. Also needed is a formulation of a proper measure of the difference between two speech segments of that duration.

ACKNOWLEDGMENT

The author wishes to thank J. L. Goldstein for providing the parameters of the MBPNL cochlear model, M. M. Sondhi for stimulating discussions throughout this work, and D. A. Berkley, A. L. Gorin, W. B. Kleijn, M. A. Randolph, J. Schroeter, and M. Slaney, for reviewing earlier versions of the paper.

REFERENCES

- [1] J. B. Allen, "Cochlear modeling," *IEEE-ASSP Mag.*, pp. 3–29, 1985.
- [2] J.-H. Chen, R. V. Cox, Y.-C. Lin, N. Jayant, and M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE Trans. Select. Areas Communicat.*, vol. 10, pp. 830–849, 1992.
- [3] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: I. Vowel-like sounds," *J. Acoust. Soc. America*, vol. 75, pp. 866–878, 1984.
- [4] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: III. Voiceless fricative consonants," *J. Acoust. Soc. America*, vol. 75, pp. 887–896, 1984.
- [5] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics," *J. Acoust. Soc. America*, vol. 75, pp. 897–907, 1984.
- [6] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: V. Vowels in background noise," *J. Acoust. Soc. America*, vol. 75, pp. 908–918, 1984.
- [7] B. Delgutte, "Analysis of french stop consonants using a model of the peripheral auditory system," in *Invariance and variability in speech processes*, J. S. Perkell and D. H. Klatt, Eds. Hillsdale, NJ: Lawrence-Erlbaum, 1986, pp. 163–177.
- [8] L. Deng, C. D. Geisler, and S. Greenberg, "A composite model of the auditory periphery for the processing of speech," *J. Phonetics*, pp. 93–108, 1988.
- [9] O. Ghitza, "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 453–485.
- [10] O. Ghitza, "Adequacy of auditory models to predict internal human representation of speech sounds," *J. Acoust. Soc. America*, vol. 93, no. 4, pp. 2160–2171, 1993.
- [11] O. Ghitza, "Processing of spoken CVC's in the auditory periphery: I. Psychophysics," *J. Acoust. Soc. America*, vol. 94, no. 5, pp. 2507–2516, 1993.
- [12] O. Ghitza and M. M. Sondhi, "Hidden Markov models with templates as nonstationary states: an application to speech recognition," *Comput. Speech, Language*, vol. 7, no. 2, pp. 101–120, 1993.
- [13] J. L. Goldstein, "Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering," *Hearing Res.*, vol. 49, pp. 39–60, 1990.
- [14] J. L. Goldstein and P. Srulovicz, "Auditory nerve spike intervals as an adequate basis for aural spectrum analysis," in *Psychophysics and Physiology of Hearing*, E. F. Evans and J. P. Wilson, Eds. London: Academic, 1977, pp. 337–346.
- [15] D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. America*, vol. 87, no. 6, pp. 2592–2605, 1990.
- [16] R. Jakobson, C. G. M. Fant, and M. Halle, "Preliminaries to speech analysis: the distinctive features and their correlates," Acoustic Laboratory, M.I.T., Cambridge, MA, Tech. Rep. 13, 1952.
- [17] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- [18] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J. Acoust. Soc. America*, vol. 68, pp. 1115–1122, 1980.
- [19] N. Y. S. Kiang, and W. T. Peake, "Physics and physiology of hearing," in *Stevens Handbook of Experimental Psychology*, second ed. New York: Wiley, 1988, pp. 277–326.

- [20] N. Kitawaki, "Quality assessment of coded speech," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 357-386.
- [21] P. Kroon and B. S. Atal, "Predictive coding of speech using analysis-by-synthesis techniques," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 141-164.
- [22] M. C. Liberman, "Auditory-nerve response from cats raised in a low-noise chamber," *J. Acoust. Soc. America*, vol. 63, pp. 442-455, 1978.
- [23] M. C. Liberman, "Single-neuron labeling in the cat auditory nerve," *Sci.*, vol. 216, pp. 1239-1241, 1982.
- [24] R. F. Lyon, "Experiments with a computational model of the cochlea," *Int. Conf. Acoust. Speech, Signal Processing, ICASSP'86*, vol. 3, 1986, pp. 1975-1978.
- [25] B. C. J. Moore, *Frequency Selectivity in Hearing*. London: Academic, 1986.
- [26] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," *J. Acoust. Soc. America*, vol. 74, pp. 750-753, 1983.
- [27] J. O. Pickles, *An Introduction to the Physiology of Hearing*. London: Academic, 1982.
- [28] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice Hall, 1988.
- [29] M. B. Sachs and E. D. Young, "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," *J. Acoust. Soc. America*, vol. 66, pp. 470-479, 1979.
- [30] M. B. Sachs, C. C. Blackburn, and E. D. Young, "Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus," *J. Phonetics*, vol. 16, pp. 37-54, 1988.
- [31] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, pp. 55-76, 1988.
- [32] S. Shamma, "The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives," *J. Phonetics*, vol. 16, pp. 77-92, 1988.
- [33] W. D. Voiers, "Evaluating processed speech using the Diagnostic Rhyme Test," *Speech Technol.*, vol. 1, no. 4, pp. 30-39, 1983.
- [34] W. D. Voiers, "Effects of noise on the discriminability of distinctive features in normal and whispered speech," *J. Acoust. Soc. America*, vol. 90, no. 4, pt. 2, Abst. 6SP10, Oct. 1991.
- [35] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Trans. Select. Areas Commun.*, vol. 10, pp. 819-830, 1992.
- [36] J. G. Wilpon and L. R. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 587-594, 1985.
- [37] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1381-1403, 1979.



Oded Ghitza (SM'93) was born in Haifa, Israel on September 24, 1948. He received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1975, 1977, and 1983, respectively.

From 1980 to 1984 he was with the Signal Corps Research Laboratory of the Israeli Defense Forces as a consultant. During 1984 and 1985 he was a Bantrell post-doctoral fellow at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, and a consultant at the Lincoln Laboratory Speech Systems Technology group, Lexington, MA. Since 1985, he has been with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ, where his research is aimed at modeling auditory physiology and perception and at creating signal analysis methods for speech recognition, coding, and evaluation.