Adequacy of auditory models to predict human internal representation of speech sounds

Oded Ghitza
AT&T Bell Laboratories, Acoustics Research Department, Murray Hill, New Jersey 07974

(Received 24 March 1992; revised 20 August 1992; accepted 13 December 1992)

A long-standing question that arises when studying a particular auditory model is how to evaluate its performance. More precisely, it is of interest to evaluate to what extent the model representation can describe the actual human internal representation. Here, this question is addressed in the context of speech perception. That is, given a speech representation based on the auditory system, to what extent can it preserve phonetic information that is perceptually relevant? To answer this question, a diagnostic system has been developed that simulates the psychophysical procedure used in the standard Diagnostic-Rhyme Test (DRT). In the psychophysical procedure, the subject has all the cognitive information needed for the discrimination task, *a priori*. Hence, errors in discrimination are due mainly to inaccuracies in the auditory representation of the stimulus. In the simulation, the human observer is replaced by an array of recognizers, one for each pair of words in the DRT database. The recognizer used [Ghitza and Sondhi, J. Acoust. Soc. Am. Suppl. 1 **87**, S107 (1990)] was designed to keep errors due to the recognition procedure to a minimum, so that the overall detected errors are due mainly to inaccuracies in the front-end representation. The system provides detailed diagnostics that show the error distributions among six phonetically distinctive features. Results are given for the behavior of two speech analysis methods, a representation based on the auditory system and one based on the Fourier power spectrum, in quiet and in a noisy environment. These results are compared with psychophysical results for the same database.

PACS numbers: 43.72.Ar, 43.71.An

INTRODUCTION

From a functional point of view, one can divide the auditory pathway into two parts, peripheral and central (Fig. 1). Using this partitioning, the input to the auditory periphery is the acoustic signal and its output is some "auditory representation" which, in turn, serves as the input to the central part. The processing principles of the central part are associated with cognition; i.e., higher-order sensations, understanding, decision making. On the other hand, the preprocessing that takes place in the auditory periphery is based on acoustic properties, aiming to provide an appropriate internal representation that is free from nonrelevant information. Morphologically, the boundary between the two parts may be drawn in the area of the primary auditory cortex. Hence, the periphery contains the outer, middle and inner ears as well as the neural centers composing the auditory brain stem and auditory midbrain. The cognitive element is the auditory cortex.

Current research activity in auditory modeling is mostly devoted, according to the partition of Fig. 1, to the study of the auditory periphery. Usually, the purpose of an auditory model is to provide a representation which is perceptually relevant. A long-standing question that arises when studying a particular auditory model is how to evaluate its performance. More precisely, it is of interest to measure to what extent the model representation can describe the actual human internal representation. In this study, we address this question in the context of speech perception. That is, given a speech representation based on the auditory system, to what extent can it preserve phonetic information that is perceptually relevant?

In the past, auditory models were quantitatively evaluated only in the context of speech recognition, serving as front-ends to automatic speech recognition systems (e.g., Ghitza, 1986). However, this kind of evaluation technique does not address the needs specified earlier. First, the measured performance is of the overall system, front end (auditory model) and back end (recognizer) combined. There is no way of separating the errors caused by the front end from those caused by the back end and, therefore, there is no clear picture of how well the auditory model performs. And second, there is no attempt to relate the performance of the auditory model to the performance of a human. Therefore, the question of how well the model representation predicts the internal human representation is not addressed.

In this study, we introduce a method that attempts to resolve these limitations. The proposed method comprises two parts. In the first part, data are collected on how accurate the human periphery is in representing phonetic distinctive features. To obtain these data a psychophysical experiment was identified, capable of measuring inaccuracies in the internal auditory representation of speech in isolation from the cognitive stages of speech perception. In the second part, the experimental procedure is simulated. The auditory periphery is represented by the auditory model under investigation, and the cognitive element by an

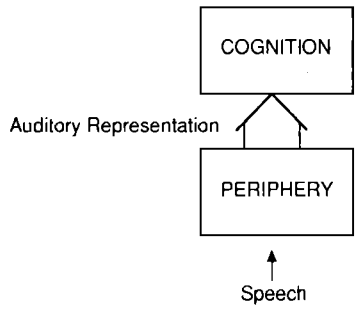

FIG. 1. The auditory pathway partitioning used in this study. According to this partitioning, the auditory pathway is divided into two parts, a peripheral part and a cognitive part. The input to the auditory periphery is the acoustic signal and its output is an "auditory representation" which, in turn, serves as the input to the cognitive part.

automatic speech recognition system especially designed to keep errors due to the decision process to a minimum. Error patterns are created that show the distribution of errors among six phonetically 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 representation is to the actual human representation. This comparison can also, in principle, provide diagnostic information on the ill-modeled parts of the auditory model. However, the study of the relationship between the diagnostic information and the parameters of the auditory model is beyond the scope of this paper.

The psychophysical experiment that we selected is the one used in the standard Diagnostic Rhyme Test (DRT), suggested by Voiers (1983). In general, the DRT test attempts to evaluate how well phonetic information is perceived by a human listener. The test is divided into two parts, collecting the psychophysical data and deriving an intelligibility score. For our purposes, only the first part, i.e., the data collection, is relevant. We discuss it in Sec. I. We show that in the DRT psychological procedure, errors in discrimination are due mainly to inaccuracies in the auditory representation of the stimulus. This is so because the subject is provided, *a priori*, with all the cognitive information needed for the discrimination task.

In Sec. II, we describe the simulation of the DRT procedure. The cognitive process is replaced by an array of recognizers, one for each pair of words in the DRT database. The prototype recognizer is described in Appendix A. It was designed to keep errors due to the recognition procedure to a minimum, so that the overall detected errors are due mainly to inaccuracies in the front-end representation.

The DRT simulation is not constrained to auditory models alone and can be used to evaluate any speech analysis method. In Sec. III, we diagnose the behavior of two speech analysis methods, a representation based on the auditory system and the Fourier power spectrum, in quiet and in a noisy environment. The results are compared with psychophysical results for the same database.

TABLE I. Stimulus words used in the DRT (borrowed from Voiers, 1983).

VOICING Voiced–Unvoiced	NASALITY Nasal–Oral	SUSTENTION 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
SIBILATION Sibilated–Unsibilated	**GRAVENESS** Grave–Acute	**COMPACTNESS** Compact–Diffuse
zee–thee	weed–reed	yield–wield
cheep–keep	peak–teak	key–tea
jilt–gilt	bid–did	hit–fit
sing–thing	fin–thin	gill–dill
juice–goose	moon–noon	coop–poop
chew–coo	pool–tool	you–rue
Joe–go	bowl–dole	ghost–boast
sole–thole	fore–thor	show–so
jest–guest	met–net	keg–peg
chair–care	pent–tent	yen–wren
jab–dab	bank–dank	gat–bat
sank–thank	fad–thad	shag–sag
jaws–gauze	fought–thought	yawl–wall
saw–thaw	bong–dong	caught–taught
jot–got	wad–rod	hop–fop
chop–cop	pot–tot	got–dot

I. THE DIAGNOSTIC RHYME TEST (DRT)

The specific Diagnostic Rhyme Test that we use was suggested by Voiers (1983) as a way of measuring the intelligibility of processed speech. The test is carefully controlled in both the contextual information presented to the listener and the psychophysical procedure.

Voiers database consists of 96 pairs of confusable words spoken in isolation by several male and female speakers. Each word is of the CVC type, and words in a pair differ only in their initial consonant. The words are equally distributed among six phonetic distinctive features, 16 word pairs per feature. The feature classification on which Voiers DRT test is based follows the binary system suggested by Jakobson *et al.* (1952). Table I (borrowed from Voiers, 1983) lists the distinctive features and their binary dimensions. It also shows the list of words in the DRT test. Selecting the features' (binary) values characterizes the mode of the speech production mechanism for producing the initial consonant in a given CVC word. The *voicing* feature characterizes the nature of the source, being periodic or nonperiodic. The *nasality* feature indicates the existence of a supplementary resonator. *Sustention* corresponds to the continuant-interrupted contrast of Jakobson

TABLE II. Consonant taxonomy used in construction of the DRT (borrowed from Voiers, 1983). Key: + = present, − = absent, ∘ = does not apply.

Features	m	n	v	ð	z	ʒ	ʒ̂	b	d	g	w	r	l	j	f	θ	s	ʃ	ʃ̂	p	t	k	h]
Voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−
Nasality	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Sustention	−	−	+	+	+	+	−	−	−	−	+	+	+	+	+	+	+	+	−	−	−	−	+
Sibilation	−	−	−	−	+	+	+	−	−	−	−	−	−	−	−	−	+	+	+	−	−	−	−
Graveness	+	−	+	−	−	∘	∘	+	−	∘	+	−	∘	∘	+	−	−	∘	∘	+	−	∘	∘
Compactness	−	−	−	−	−	+	+	−	−	+	−	−	∘	+	−	−	−	+	+	−	−	+	+

et al., and *sibilation* corresponds to their strident-mellow contrast. Finally, *graveness and compactness* represent broad resonant features of the speech sound, related to Miller and Nicely's place of articulation (Miller and Nicely, 1955). Table II (also borrowed from Voiers, 1983) shows the feature makeup of various consonants. In compiling the DRT word list, various criteria were observed in order to achieve balanced representation of each feature (Voiers, 1983). It should be noted, however, that some of the choices are controversial (Syrdal, 1987). In this study, we used Voiers's original DRT word list.

The psychophysical procedure in the DRT is also very carefully controlled. The listeners are well trained and are very familiar with the database, including the voice quality of the individual speakers. The experiment is a two-alternative, forced choice (2AFC) experiment. First, the subject is presented visually with a pair of rhymed words. Then, one word of the pair (selected in 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 semi confusion matrix, or as a distribution among the six phonetic distinctive features.

The controlled nature of both the database and the test procedure is the basis for our assumption that all cognitive information needed for the discrimination task is available to the listener prior to the aural presentation. 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 can be used for reference purposes, reflecting errors in the internal human auditory representation during the DRT discrimination task.

II. SIMULATING THE DRT

In the simulation, the peripheral part of the auditory pathway is modeled by the auditory model under investigation, 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.

A recognizer in the array represents one DRT word pair. For a test word (out of a given word pair), the recognizer makes a maximum-likelihood decision 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 phonetic sequence. Each state of the HMM word model represents one phonetic unit. The recognizer used is an HMM recognizer with time-varying states, suggested by Ghitza and Sondhi (1990,1993) and described in Appendix A. In this recognizer, state of the HMM represents one phonetic unit in terms of a time-varying mean sequence of ordered frames, a template, and a block covariance matrix that characterizes the intraframe statistical dependence within the phonetic unit. The particular phonetic unit that was selected is a diphone. In this way, the dynamic nature of coarticulation between two successive phonemes is represented more accurately, and the ability to discriminate between the initial consonants of the DRT words is improved.

The simulation procedure is described in Fig. 2. To further increase the accuracy of the recognition procedure,

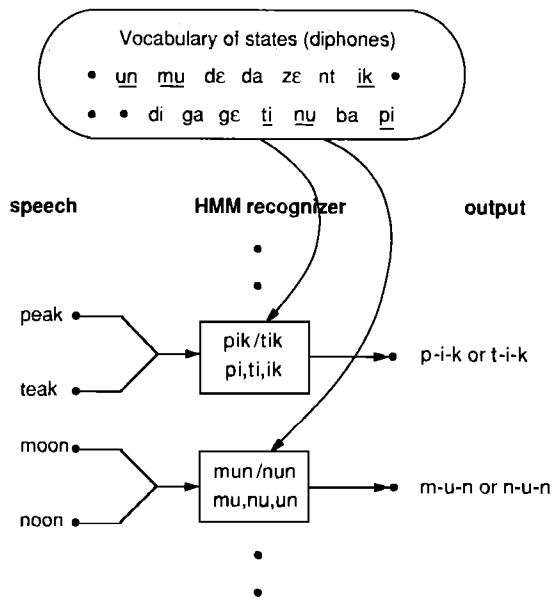

FIG. 2. An illustration of the DRT simulation procedure. To test the word pair "peak/teak," for example, the state models for the diphones pi, ti, and ik are drawn from the states vocabulary, along with the appropriate transition matrix that allows only the necessary transitions. The recognizer is then 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.

the simulation is done on a speaker-dependent basis. Every speaker in Voiers database 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 repetitions by hand. The vocabulary covers all the diphones that appear 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. This is in view of our assumption that the cognitive representation of a phonetic unit is universal and not context dependent.

For a given model to be evaluated, the diphones in the vocabulary are transformed to the appropriate representation domain, resulting in an inventory of state models. The words in the test repetition are also processed and represented in the same domain.

The testing phase is a simulation of the 2AFC paradigm. For testing a particular word pair, the recognizer is first loaded with the appropriate state models (drawn from the inventory) and transition matrices. This step simulates 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 can then be displayed in form of a semi-confusion matrix or as a distribution among the six distinctive features.

Figure 2 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. 2), 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."

As we mentioned before, the simulation is performed on a speaker-dependent basis to keep errors due to the recognition process to a minimum. The single-speaker test procedure (training and testing) is repeated for every speaker in the database. The overall data are then analyzed, to find similarities in the error patterns across speakers, signal conditions, and phonetic features.

III. EXPERIMENTAL RESULTS

A. Signal conditions

In this study, we used three male speakers, 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. Three "noisy" versions of the testing repetitions were created by adding white noise to the original ("clean") signals. The signal-to-noise ratio (SNR) levels were 30, 20, and 10 dB. The 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. Here, 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.

The noisy versions were 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 recording tape 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 signal was amplified (or attenuated) by a gain factor that was calculated in advance, to maintain the desired global SNR.

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

B. Description of the analysis methods

We tested two speech representation methods, the Ensemble Interval Histogram (EIH) and the traditional Fourier power spectrum. The first representation is based on the auditory model suggested by Ghitza (1992). Both the EIH and the Fourier power spectrum contain information about the spectral envelope as well as about the spectral fine structure. The tests to be described here were performed by utilizing only spectral envelope variations. The spectral envelope variations were represented by a P-order truncated cepstral series. Since we are not considering the effect of signal intensity we set c_0 to 0.

The auditory model for the EIH representation is described in Appendix B. The model uses 165 filters (approximating the filters in a cat's cochlea), and five thresholds per filter. The EIH is computed once every 10 ms. The interval statistics at time t_0 are collected from all 825 (165×5) threshold detectors, using all simulated firing records which exist in the windows that end at time t_0 (see Fig. B3). The length of each window is 20/CF, where CF is the center frequency of the cochlear channel. Since the

levels are equally distributed on a logarithmic scale, the EIH is treated as a log "spectrum." Hence, a "cepstral" representation of the EIH can be obtained by computing the inverse DFT of the EIH. For the EIH, an appropriate envelope fit is achieved by truncating the cepstral series at c_{25}. This order of fit was required because of the larger dynamic range of the EIH (compared to that of the Fourier power spectrum). The first "cepstral" coefficient (c_0), which can be used as an estimate of the loudness at time t_0, was set to 0.

For the Fourier power spectrum, an 11th-order cepstral representation was computed every 10 ms. 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 ten coefficients were used. In this way, the envelope is normalized in the sense that the average value of the LPC log spectrum is 0.

C. Results

The raw data that summarize the outcome of one experimental run are organized in the form of a matrix with 12 rows and 16 columns. The rows stand for the phonetic features (six dimensions times two values per dimension—attribute present and attribute absent) and the columns represent the words 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 indicates the number of listeners who made a mistake in identifying the corresponding word. A matrix element can be any integer between 0 and 8, where 8 is the number of listeners participating in the test. In order to have raw-data matrices of the same nature for both the psychophysical and the simulated procedures, we transformed the psychophysical results into a binary form. A threshold of 3 was arbitrarily specified and any matrix element less than the threshold value was set to 0. Otherwise, it was set to 1. This is to say that in the psychophysical procedure, an error is deemed to have occurred only if more than two listeners made the error.

Let us define three main variables, the *analyzer*, the *speaker* and the SNR. In our case, we have three analyzers (Human, EIH, Fourier power spectrum), three speakers and three levels of signal-to-noise ratio. An experiment was run for every combination of those variables, yielding 27 raw-data matrices.

In analyzing the data, we first compute (for every matrix) the average error per row (feature) which equals the number of 1's in a row divided by 16. The outcome of every experimental condition is now reduced to a 12-dimensional error vector, representing the average error, over words, for each of the 12 phonetic attributes for this condition. We now average these error vectors across speakers to create an average error vector for every combination of analyzer and SNR.

Figures 3–8 show the resulting error patterns, displayed in six different ways. Every figure contains four plots, where the left-upper plot is a summary of the other three plots, excluding the standard-error bars. 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.

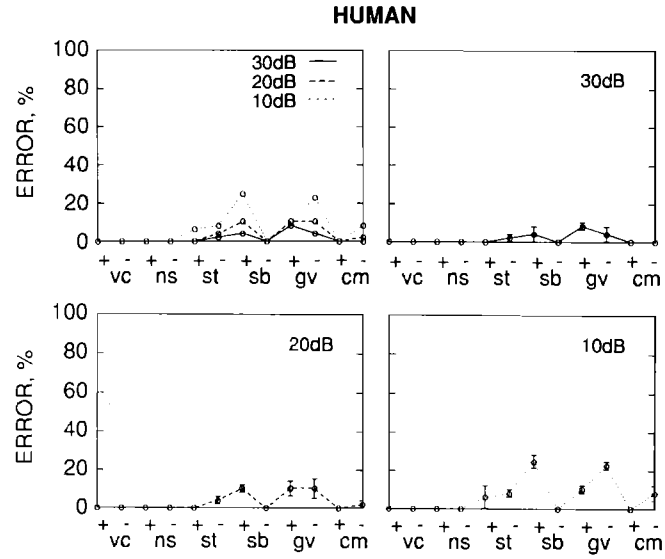

FIG. 3. Distribution of errors made by the human listener in three values of signal-to-noise ratio, processed from the responses of eight listeners. The left-upper plot is a summary of the other three plots, excluding the standard-error bars. The abscissa of every plot indicates the six phonetic features: "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).

In Figs. 3–5, the error distributions are displayed separately for every analyzer, with the SNR as a parameter. We see that although the volume of the errors increases

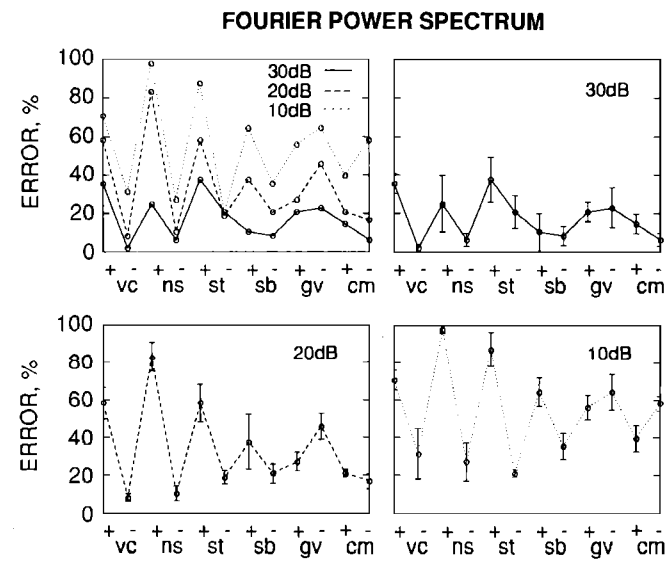

FIG. 4. As in Fig. 3, for the Fourier power spectrum. Only spectral envelope variations (represented by the 11th-order truncated cepstral series) are utilized.

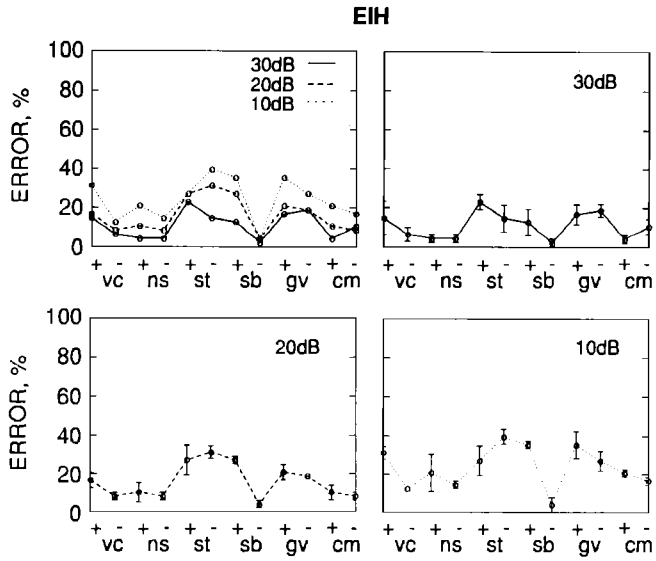

FIG. 5. As in Fig. 3, for the EIH. Only EIH envelope variations (represented by the 25th-order truncated "cepstral" series) are utilized.

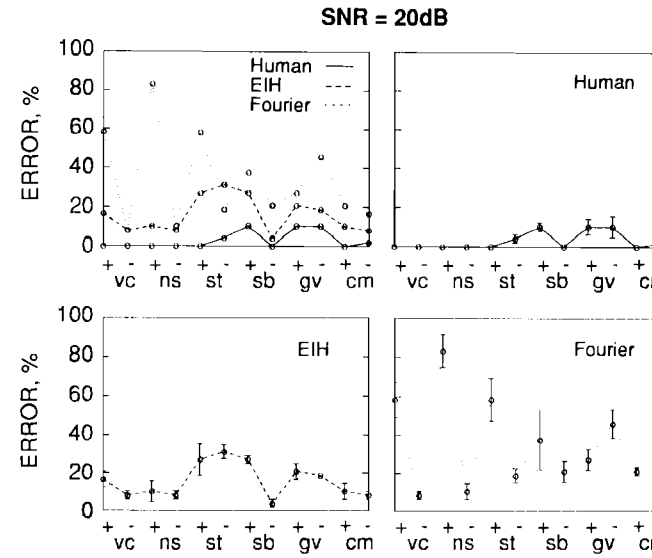

FIG. 7. As in Fig. 6, for SNR of 20 dB.

with the increase of noise level, the error patterns for every analyzer remain similar. The number of errors made by the human is much lower than the number of errors made by the machine, for both analyzers.

In Figs. 6–8, the error distributions are displayed separately for every SNR condition, with the analyzer as a parameter. Every analyzer exhibits a characteristic error distribution. The error distributions are substantially different from each other. Moreover, as expected from Figs. 3–5, the differences between the error patterns are consistent across all noise levels that were examined. Three points are noteworthy. First, the human observer performs much better than the EIH and Fourier power spectrum, and is very robust to noise. Second, the errors made by the Fourier power spectrum analyzer are mainly in the presence of voicing, nasality, sustention, and sibilation. And third, EIH is more robust to noise than the Fourier power spectrum, in agreement with previous reports (e.g., Ghitza, 1992).

It is possible to argue that the comparison of error distributions due to the EIH and the Fourier power spectrum might be biased because the recognizer may be matched better to one analyzer than the other. We tried to eliminate this kind of bias and to ensure that errors are due to the properties of the analyzer. In the training procedure, the origin of a particular state (diphone) was the same across all representation methods, and the segmentation by hand was done in the time domain; The same initial, middle, and final time instants for a given phoneme were used by both front-ends. The structure of the recognizer, as well as the transition matrix, were also fixed, irrespective of the

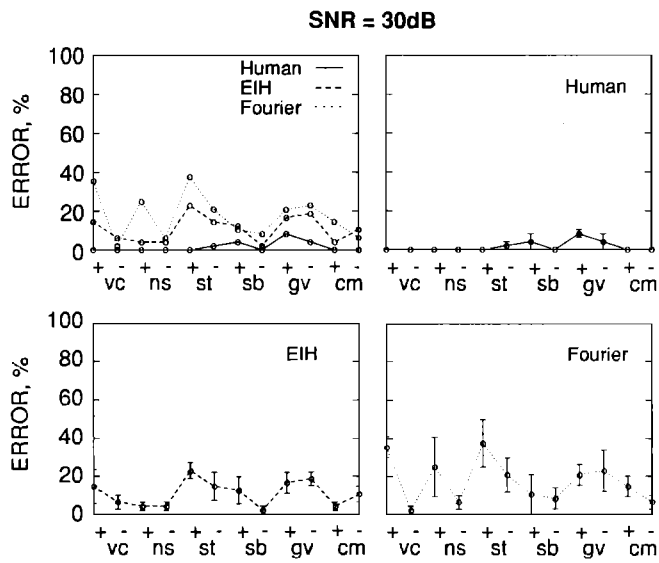

FIG. 6. Comparing error distributions of the human listener, the Fourier power spectrum and the EIH, in a 30-dB signal-to-noise ratio. Figure legend is as in Fig. 3.

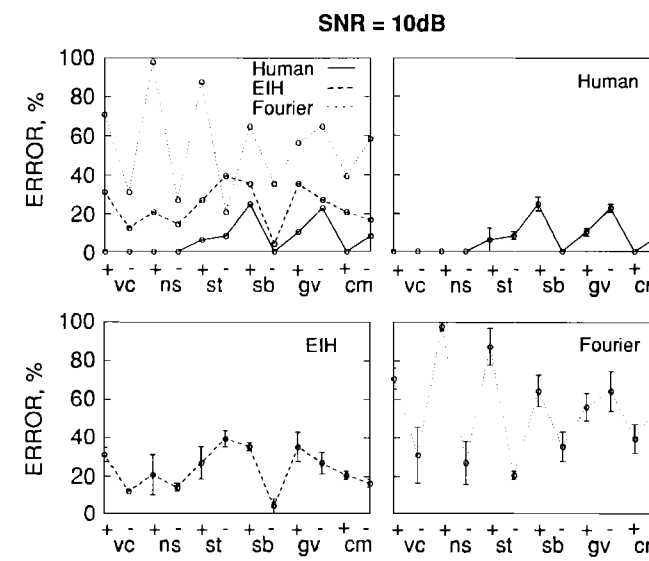

FIG. 8. As in Fig. 6, for SNR of 10 dB.

representation method. Therefore, the error differences are indeed due to the characteristics of the analyzers.

IV. DISCUSSION

In previous sections, we outlined a method for evaluating how adequate an auditory model is in predicting the internal human representation of speech sounds. In this section, we shall discuss arguments that were considered in designing the main two components of the system, the psychophysical procedure and the simulation system.

A. Selecting the psychophysical procedure

To evaluate the performance of the auditory model, psychophysical data are needed on the accuracy of the human periphery in representing speech sounds. To obtain such data, an experimental procedure should be identified, capable of measuring human responses that are due only to inaccuracies in the periphery. The highly controlled 2AFC paradigm of the DRT addresses this concern. The listener, well trained and very familiar with the database, is provided with all the information required for the discrimination task *a priori*. He is first presented, visually, with the word pair. Then, one of these words is presented aurally and he has to indicate which of the words appearing in the visual display was played. Our assumption is that if, under those conditions, an error still occurs, it is due mainly to inaccuracy in the internal representation of the stimulus. If our assumption is correct, this psychophysical procedure has the property of measuring errors originated in the periphery—separated from errors made during the cognitive process.

To further demonstrate the suitability of the 2AFC paradigm to our needs, let us examine an alternative psychophysical procedure used by Miller and Nicely (1955) to measure perceptual confusions among the 16 English consonants. CV stimuli were used, with the consonants followed by the vowel [a] (as in father). The subject was first presented aurally with the stimulus and then was required to indicate which of the 16 consonants was played. Clearly, in this case, errors made by the subject are due to inaccuracies in the cognitive process as well.

B. Designing the simulation system

Using a 2AFC paradigm turns out to also be necessary for a suitable design of the simulation system. The system should detect errors due to the auditory model under investigation, as distinct from errors generated by the back end of the system. Ideally, one would like to have an error-free back end. In reality, however, the number of errors due to the back end cannot be reduced to zero even when simulating a 2AFC paradigm. Therefore, we took the approach of reducing the number of errors due to the back end to a minimum. This was done by: (1) increasing the discrimination power of the recognizer. An HMM with time-varying states was introduced, where a state model is defined by a time-varying mean sequence of ordered frames, representing a diphone, and a block covariance matrix that characterizes the intraframe statistical dependence within the diphone. (2) Specifying a strict state transition matrix such that only the necessary transitions defined by the tested word pair are allowed. Finally, (3) simulating the psychophysical procedure on a speaker-dependent basis, and averaging results across speakers.

To illustrate how these design criteria reduce the number of back-end errors, let us reexamine the example of testing the word pair "peak/teak," given in Sec. II. The recognizer consists of three states (corresponding to the diphones pi, ti, and ik) and a transition matrix that allows only two transitions (from pi to ik or from ti to ik) with equal probabilities. Hence, the recognizer is designed to make a maximum-likelihood decision between two HMM word models, "peak" and "teak." The word models are made by concatenating the best DTW versions of the state models pi and ik (for "peak") and ti and ik (for "teak"), to match the actual input word ("peak" or "teak"). Both word models utilize the same state model for the final, VC, part of the word pair (ik in our example). Hence, the error in representing this part of the actual input word by either word model is identical. Therefore, the maximum-likelihood decision between the two word models considers only the degree of accuracy by which the relevant CV part of the word is represented.

For comparison let us examine the performance of two alternative designs for the back end, based upon two approaches that are commonly used in the case of isolated word recognition. The two alternatives are the traditional HMM recognizer and the whole-word DTW recognizer. A system based upon the traditional HMM differs from our system (i.e., HMM with time-varying states) only in the way the state model is defined. A state model of a traditional HMM can be viewed as a quantized version of a state model in the HMM with time-varying states, where the degree of quantization depends on the number of substates in the state model. Therefore, the performance of the HMM with time-varying states can be viewed as the upper bound for the performance of the traditional HMM. If the DTW approach is to be used, two whole-word template models should be created for the word pair under testing (in our example, one template model for "peak" and one template model for "teak"). Hence, the simulation system should be designed to make a minimum distance decision between the two whole-word models and the actual input word. This decision, however, will depend on errors that are accumulated over the entire optimal paths (in the DTW sense) that map the input word, CV and VC combined, to the template word model. This is in contrast to the decision rule used by the HMM with time-varying states, where the discrimination between the word models depends only upon the distance between the relevant CV part of the input word and the corresponding, time-varying, CV states.

C. Correlates between phonetic features and perceptual dimensions

As discussed in Sec. I, the word pairs in Voiers database were chosen to equally cover the six phonetic distinctive features suggested by Jakobson *et al.* (1952). Hence,

the human performance, as well as the performance of the auditory model under investigation, are measured in terms of the error distributions for each of these phonetic features. But, to better suit our purpose, it would be desirable to display the errors along explicit perceptual dimensions, instead. Finding perceptual correlates to Jakobson *et al.*'s features is beyond the scope of this paper. Nevertheless, indirect evidence exists for such a correlation. Using multidimensional scaling techniques, Shepard (1972) and Wish and Carroll (1974), demonstrated that the place and manner features do relate to some specific perceptual dimensions. In view of the relationship between these features and Jakobson *et al.*'s features, Voiers database might be implicitly specified over perceptual dimensions. One additional point is noteworthy. Shepard, as well as Wish and Carroll based their studies on Miller and Nicely's experimental data (1955) which reflects the overall auditory responses—peripheral and central combined. Hence, the perceptual dimensions they proposed are associated with the central parts of the auditory pathway as well. On the other hand, errors measured during the DRT experimental procedure occur at the peripheral level. We may conclude, therefore, that error patterns produced by the DRT procedure (e.g., Figs. 3–8) display inaccuracies in the human peripheral representation along relevant, cognitive dimensions.

D. Training the simulation system

In the current study, we considered the case of noisy speech signals (Figs. 3–8). Nevertheless, we trained the recognition system using the clean version of the training database. Since the training phase is associated with mimicking the (error-free) human cognitive element, a question can be asked: Is training by using clean speech suitable for an appropriate mimic, or should the training be under noisy conditions similar to those of the testing database? To reduce the back-end errors to a minimum, it is preferred to train the recognizer under noisy conditions (e.g., Juang, 1991, Fig. 1). Doing so, however, implies an underlying assumption that the decision strategy of the human listener involves an adaptation process in which the "internal states" that represent the basic speech units change with changes in the environmental conditions. Training in quiet, on the other hand, implies that the cognitive states of the basic speech units remain unchanged and that the auditory periphery is capable of producing representations that remain stable under variations in the signal conditions.

E. Covering the perceptual space of speech

Finally, it is important to note that the use of Voiers database (which contrasts only the initial consonant in a list of CVC word pairs) limits the range of acoustic ambiguities tested. Complementary diagnostic information should be obtained from additional databases, for example a database that contrasts the final consonants in a list of CVC word pairs (Voiers, 1991), or middle consonants in a list of VCV word pairs (e.g., Schmidt-Nielsen, 1983). At this point, therefore, our study should only be regarded as a demonstration of the diagnostic capabilities offered by this approach.

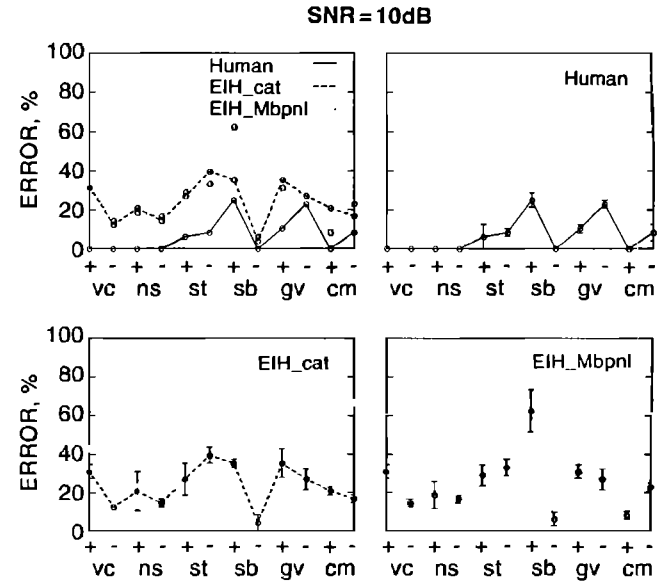

FIG. 9. Comparison of distribution errors made by the human listener, the EIH with the cat filters (EIH_cat), and the EIH with the Mppnl (EIH_Mbpnl), in a 10-dB signal-to-noise ratio.

V. SUMMARY

In this paper, we outlined a method for evaluating how adequate a speech analysis method is in predicting the actual human representation of speech sounds. In addition to measuring the overall error rate, the method provides detailed diagnostics that show the error distributions among six phonetically distinctive features.

To demonstrate the power of the suggested evaluation method, we considered the behavior of two speech analysis methods, a representation based on the auditory system (the EIH representation) and the Fourier power spectrum, in quiet and in a noisy environment. The results were compared with psychophysical results for the same database. The results show that the overall number of errors made by the machine (the EIH or the Fourier power spectrum) are far greater than the overall number of errors made by a human, at all noise levels that were tested. Further, the errors made by the human listener are distributed in a different way compared to the errors made by the machines, and that the distributions of errors made by the two analyzers are also quite different from each other.

ACKNOWLEDGMENTS

I wish to thank B. S. Atal, J. P. Olive, and M. M Sondhi for stimulating discussions throughout this work, and to J. P. van Santen for suggestions concerning the statistical analysis of the data.

ADDENDUM

Since this paper was submitted for publication the proposed method was used to evaluate other auditory models. Of a unique interest is an EIH model similar to the one discussed in Appendix B, but with different cochlear filters. Instead of the filters of Fig. B2 (derived from tuning curves of cats) we now use a phenomenological model of the human cochlea suggested by Goldstein [J. Goldstein, Hear. Res. **49**, 39–60 (1990)]. The model is termed Mbpnl, for "Multi bandpass nonlinear" processor. We use 190 Mbpnl channels distributed from 200 to 7000 Hz according to the frequency position suggested by Greenwood [D. Greenwood, J. Acoust. Soc. Am. **87**, 2592–2605 (1990)]. The filters operate in the time domain and change their gain and bandwidth with changes in the input intensity, in accordance with psychophysical behavior.

Figure 9 shows a comparison of distribution of errors made by the human listener, the EIH with the cat filters (EIH_cat) and the EIH with the Mbpnl filters (EIH_Mbpnl), in a 10-dB signal-to-noise ratio. (See Fig. 3 caption for abbreviation index.) EIH_cat and EIH_Mbpnl demonstrate very similar performance in all dimensions except sb+ and cm+. Although the overall number of errors of the two EIHs is almost the same, the error distribution of EIH_Mbpnl is closer in shape to the error distribution of the human observer. Note, however, that the overall number of errors of both EIHs is still much higher than that of the human.

APPENDIX A

This Appendix introduces an extension to the traditional application of hidden Markov models (HMMs) to speech recognition. The new concept was developed by M. Mohan Sondhi and the author (Ghitza and Sondhi, 1990,1993) and was used in the DRT simulations, as described in Sec. II.

1. Templates as states in a HMM with nonstationary states

Consider a Markov chain with N states, $Q \equiv [q_1, q_2, \ldots, q_N]$, and associated transition probability matrix $A \equiv [a_{ij},\ 1 \leqslant i,j \leqslant N]$. If S_k denotes the state of the Markov chain at time instant k, then by definition $a_{ij} = \text{prob}(S_{k+1} = q_j | S_k = q_i)$. A hidden Markov model (HMM) based on this Markov chain generates a random sequence of observation vectors $o_1, o_2, \ldots, o_k, \ldots$, whose statistical properties change as the state of the underlying Markov chain changes.

In almost all applications of HMMs to speech recognition, the probability distribution of the observation o_k, generated at time instant k, is assumed to depend only on the state $S_k \in Q$, in which it is generated. Hence, the observations generated in any given state are independent and identically distributed (i.i.d.). Thus, if the *sequence* of observations $O \equiv [o_t, o_{t+1}, \ldots, o_{t+K}]$ is generated in some state q (i.e., if $S_t = S_{t+1} = \cdots = S_{t+K} = q$), then the assumption is that the probability of that sequence has the form

$$P(O) = \prod_{k=t}^{t+K} p(o_k | q). \tag{1}$$

The state-dependent probability distribution $p(o|q)$ can take a variety of forms. If the observations are d-dimensional vectors of continuously distributed components, the distribution is usually assumed to be a d-dimensional Gaussian distribution (or a mixture of such distributions).

Some more general models have been considered in the literature (although not widely used). Thus Bahl *et al.* (1983) assumed that o_k depends on S_k as well as on S_{k-1}. Wellekens (1987) allowed o_k to depend on S_k, S_{k-1}, and o_{k-1}, i.e., on the previous observation as well.

Even with these generalizations, a sequence of observations generated in a given state is a segment of a stationary time-discrete random process. In certain situations (e.g., when the spoken word "eight" is represented by a five-state HMM), this assumption of stationarity is reasonable. If, however, the state is to represent a plosive, or a long segment of speech (tens of milliseconds) the assumption is clearly invalid. To the best of our knowledge, no one has considered HMMs in which the states are nonstationary, i.e., in which the probability of an observation sequence depends *explicitly* on the time index, k. It is this extension that is the subject of the HMM with nonstationary states.

Our motivation for studying such a model comes from the application of HMMs to speech recognition in terms of subword units. Such HMMs are of interest in large-vocabulary recognition, as well as in other applications where a decoding in terms of subword units is desirable. Specifically, consider the HMM "phonetic decoder" presented by Levinson (1987), in which each state represents a (variable-duration) phone. With this choice of subword units, the model has about 50 states, each specified by a duration probability density, and a probability density for the observations. Successive observations in a state are assumed, as above, to be i.i.d. Let us consider the problem faced by this model in representing the spoken word "gob" whose spectrogram is shown in the left side of Fig. A1. Shown below the spectrogram is an approximate phonetic transcription. It is clear that if the phone [a], say, is represented by a state in the HMM, that state must be nonstationary. (In fluent speech, such nonstationary states are the rule, and "steady" states the rare exception.) To represent such a state by time-averaged statistical properties is a gross approximation. Another unsatisfactory feature is that because of the i.i.d. assumption, the probability assigned to a set of observations is independent of the order in which the observations occur. Thus, for instance, reversing the direction of the formant transitions leaves the probability unchanged.

The way this nonstationarity has been dealt with in the past is by representing the transient state as a concatenation of two or more substates. Thus the nonstationary state is approximated by a sequence of piecewise stationary states. In principle, any transient state can be approximated this way by a sufficiently fine subdivision. We pro-

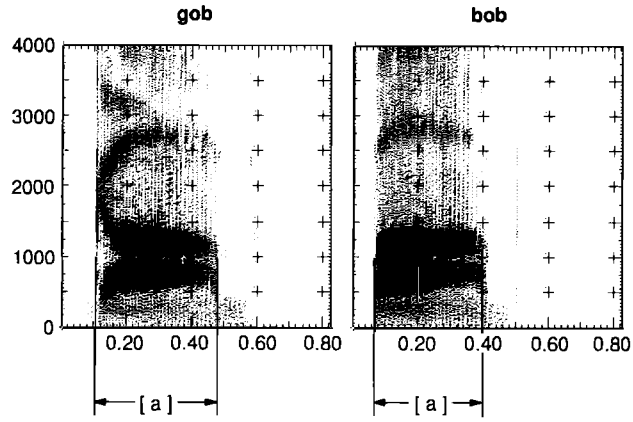

FIG. A1. Spectrograms of the phoneme [a] inside the words "bob" and "gob." It is clear that if the phoneme is represented by a state in an HMM, that state must be nonstationary, to reflect the time variation in the formant locations. To represent such a state by time-averaged statistical properties is a gross approximation.

pose an alternative point of view in which the entire subword unit is regarded as a single *nonstationary* state.

A moment's reflection shows that a diphone is the smallest subword unit for which such an HMM with nonstationary states makes sense. This is because of coarticulation. The spectral trajectory of, say, the vowel [a] is quite different in the CV syllable /bo/ from that in the syllable /go/, as shown in Fig. A1. Clearly, it would defeat our purpose if we would consider all occurrences of [a], regardless of context, to belong to the same ensemble. From the very outset, therefore, we consider states to represent diphones. (It is, of course, possible to consider even more complicated subword units. However, we have not done that.)

The structure of our HMM is similar to that of the variable duration HMM described by Levinson (1987). The main difference is, of course, in the definition of a state, and in the manner in which a probability is assigned to a sequence generated in a given state.

As mentioned before, we have chosen the states to be diphones. Assuming there are about 50 phonemes in English, we expect the number of states, N, to be on the order of about 2000.

The dwell time in a state of a conventional HMM is exponentially distributed. As this is not, in general, a good approximation to the duration distribution, we modify the HMM as in Levinson (1987). Thus the $N \times N$ state transition matrix A is constrained to have its diagonal elements $a_{ii}=0$, for all i, and the dwell time in a state is governed by a state-dependent probability distribution of durations.

The definition of a state is in terms of a template (or typical sequence of observations) and a probability distribution of the deviations from the template.

APPENDIX B

This Appendix briefly describes the auditory model that was used in the experimental part of this study. A detailed description of the model can be found in Ghitza (1992).

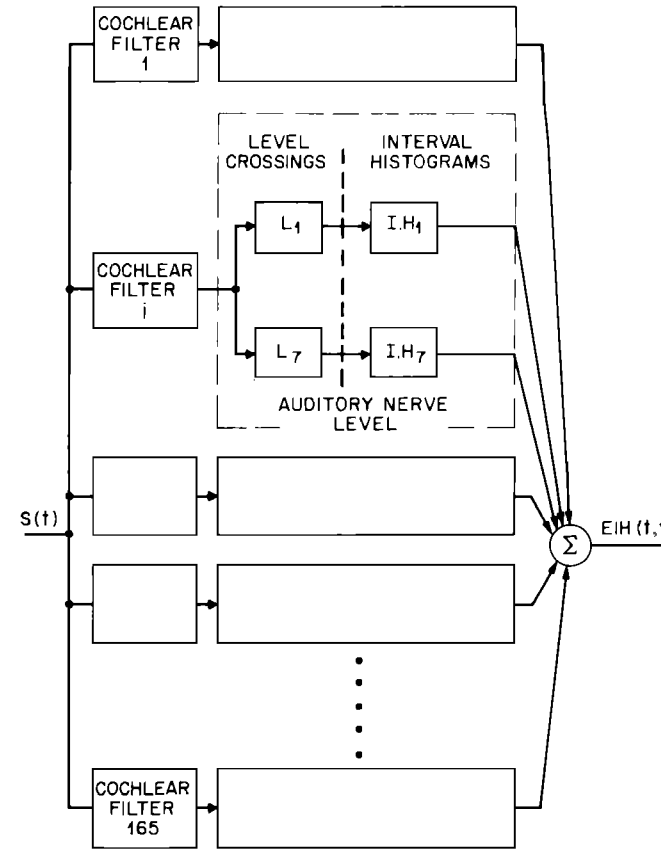

FIG. B1. The ensemble interval histogram (EIH) computational model. The cochlear component consists of 165 filters (channels), whose center frequencies are equally spaced on a log-frequency scale, between 150 and 7000 Hz. The level-crossings are measured at positive threshold crossings. The positive-threshold levels are pseudorandomly distributed, on a log scale, over the dynamic range of the signal. The multidimensional point process derived from the five level-crossing detectors simulates the auditory-nerve firing patterns. The interval histogram is created by distributing the inverse of the detected intervals across 128 bins equally spaced on a linear frequency scale, between 0 and 4000 Hz. Only the most recent intervals are used in the computation (an interval is defined as the time between two adjacent positive-going level crossings). The ensemble histogram is the sum of the corresponding histogram bins over all of the simulated fibers in the array.

1. The ensemble interval histogram (EIH) model

The model is schematically illustrated in Fig. B1. Its first stage represents the auditory periphery up through the level of the auditory nerve. The mechanical motion of the basilar membrane is sampled by 165 inner hair cell (IHC) channels, equally spaced along a log-frequency axis between 150 and 7000 Hz. The corresponding cochlear filters have been simulated in detail, using actual neural tuning curves for cats collected by M. C. Liberman (unpublished). The amplitude responses of 28 filters (one every six) are shown in Fig. B2. Their phase characteristic is minimum phase and their relative gain, measured at their center frequencies, reflects the cat's middle ear transfer function. The ensemble of nerve fibers innervating a single IHC is simulated by 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 cross-

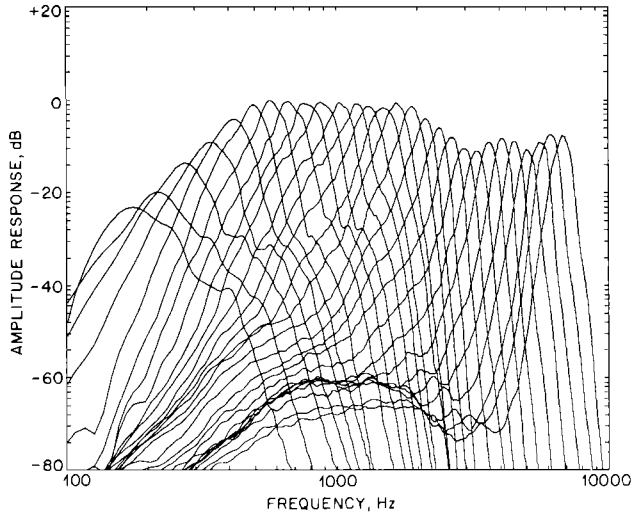

FIG. B2. The amplitude response of 28 (one every six) simulated cochlear filters, plotted in a logarithmic frequency/decibel scale. The filters' amplitude response is the actual neural tuning curves for cats collected by M. C. Liberman (unpublished data). Note the high degree of overlap among filters.

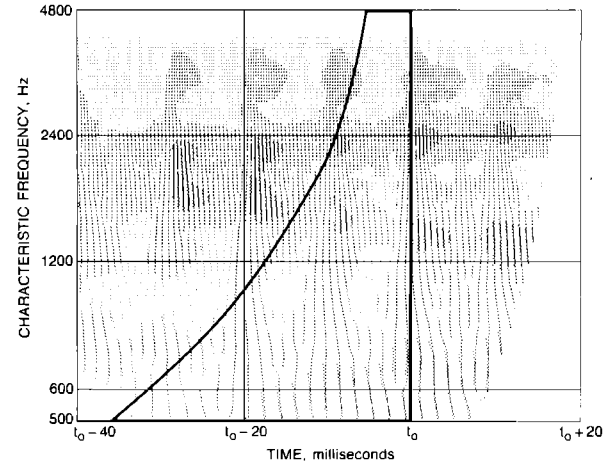

FIG. B3. Simulated auditory-nerve activity for the first 60 ms inside the vowel [a] in the word "gob." The abscissa represents time and the ordinate represents the characteristic frequency of the IHC channels. Note the logarithmic scale of the characteristic frequency, which represents the place-to-frequency 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. 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 center frequency (CF). The length of each window is 20 times 1/CF, where CF is the center frequency of the cochlear channel.

ing. The detectors are pseudorandomly distributed across a range of positive levels. The values assigned to the level j of every filter is a random Gaussian variable, with a mean L_j and a standard-deviation $\sigma=0.2L_j$. The mean values $L_1,...,L_j,..,L_5$ are uniformly distributed on a log scale over the amplitude range characteristic of speech sounds. The random nature of the values of the jth level across the cochlear filter array reflects the fact that the diameters and the synapse-connection size of the fibers that innervate the same side of different IHCs along the cochlear partition incorporate a certain amount of intrinsic variability (which is characteristic of most physiological systems).

The output of the level-crossing detectors represent the discharge activity of an ensemble of auditory-nerve fibers. Figure B3 shows simulated auditory-nerve activity for the first 60 ms inside the vowel [a] 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. B3. However, for large signal magnitudes, several levels will be activated, creating a "darker" area of activity.

The level-crossing patterns represent the auditory-nerve activity which serves, in turn, as the input to a second, more central stage of neural processing. It is assumed that: (1) neural circuits beyond the auditory nerve have a place-independent structure, and (2) these circuits operate on detailed timing information conveyed in the auditory-nerve fibers, irrespective of their tonotopic place of origin in the cochlear partition. Following these assumptions, a representation of the timing information is described as an ensemble interval histogram (EIH). Conceptually, the EIH is a measure of the spatial (tonotopic) extent of coherent neural activity across the simulated auditory nerve. Mathematically, it is the short-term probability density function of the reciprocal of the intervals between successive firings, measured over the entire simulated auditory nerve in a CF-dependent time-frequency zone (Fig. B3). The model belongs to the temporal-nonplace category. Note, however, that tonotopic information is present *implicitly*, since the information conveyed by each fiber is by itself place dependent due the tuning characteristics of the basilar membrane.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (**1983**). "A maximum likelihood approach to continuous speech recognition," IEEE Trans. PAMI **PAMI-5** (2), 179–190.

Ghitza, O. (**1986**). "Auditory nerve representation as a front-end for speech recognition in a noisy environment," Comput. Speech Lang. **1**, 109–130.

Ghitza, O. (**1988**). "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment," J. Phon. **16**, 109–124.

Ghitza, O. (**1992**). "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sondhi (Dekker, New York), pp. 453–485.

Ghitza, O., and Sondhi, M. M. (**1990**). "Templates as states in a hidden Markov model," J. Acoust. Soc. Am. Suppl. 1 **87**, S107 (1990).

Ghitza, O., and Sondhi, M. M. (**1993**). "Hidden Markov models with templates as nonstationary states: An application to speech recognition," Comput. Speech Lang. (in press.)

Goldstein, J. (1990). "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," Hear. Res. **49**, 39–60.

Greenwood, D. (1990). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Jakobson, R., Fant, C. G. M., and Halle, M. (1952). "Preliminaries to speech analysis: the distinctive features and their correlates," Tech. Rep. No. 13, Acoustic Laboratory, M.I.T., Cambridge, MA.

Juang, B. H. (1991). "Speech recognition in adverse environments," Comput. Speech Lang. **5**, 275–294.

Levinson, S. E. (1987). "Continuous speech recognition by means of acoustic/phonetic classification obtained from a hidden Markov model," Int. Conf. Acoust. Speech Signal Process. ICASSP '87, 93–96.

Liberman, M. C. (unpublished).

Miller, G. A., and Nicely, P. E. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Schmidt-Nielsen, A. (1983). "Intelligibility of VCV segments excised from connected speech," J. Acoust. Soc. Am. **74**, 726–738.

Shepard, R. N. (1972). "Psychological representation of speech sounds," in *Human Communication: A Unified View*, edited by E. E. David and P. B. Denes (McGraw-Hill, New York), pp. 67–113.

Syrdal, A. K. (1987). "Methods for a detailed analysis of Dynastat DRT results," AT&T Bell Laboratories internal Memorandum.

Voiers, W. D. (1983). "Evaluating processed speech using the Diagnostic Rhyme Test," Speech Technol. **1**(4), 30–39.

Voiers, W. D. (1991). "Effects of noise on the discriminability of distinctive features in normal and whispered speech," J. Acoust. Soc. Am. **90**, 2327(A) (1991).

Wellekens, C. J. (1987). "Explicit time correlation in hidden Markov models for speech recognition," Int. Conf. Acoust. Speech Signal Process. ICASSP '87, 384–386.

Wish, M., and Carroll, J. D. (1974). "Applications of individual differences scaling to studies of human perception and judgment," in *Handbook of Perception, Vol. II*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), pp. 449–491.