On the perceptual distance between speech segments

Oded Ghitza and M. Mohan Sondhi
Acoustics and Audio Communication Research, Bell Laboratories, Murray Hill, New Jersey 07974

(Received 5 June 1996; accepted for publication 2 August 1996)

For many tasks in speech signal processing it is of interest to develop an objective measure that correlates well with the perceptual distance between speech segments. (Speech segments are defined as pieces of a speech signal of duration 50–150 ms. For concreteness, a segment is considered to mean a diphone, i.e., a segment from the midpoint of one phoneme to the midpoint of the adjacent phoneme.) Such a distance metric would be useful for speech coding at low bit rates. Saving bits in those systems relies on a perceptual tolerance to acoustic perturbations from the original speech—perturbations whose effects typically last for several tens of milliseconds. Such a distance metric would also be useful for automatic speech recognition on the assumption that perceptual invariance to adverse signal conditions (e.g., noise, microphone, and channel distortions, room reverberation, etc.) and to phonemic variability (due to nonuniqueness of articulatory gestures) may provide a basis for robust performance. In this paper, attempts at defining such a metric will be described. The approach in addressing this question is twofold. First psychoacoustical experiments relevant to the perception of speech are conducted to measure the relative importance of various time-frequency ''tiles'' (one at a time) *when all other time-frequency information is present*. The psychophysical data are then used to derive rules for integrating the output of a model of auditory-nerve activity over time and frequency. [S0001-4966(97)03901-5]

PACS numbers: 43.72.Ar, 43.71.An, 43.71.Cq, 43.66.Ba [JS]

INTRODUCTION

This paper is concerned with the derivation of a quantitative measure of the perceptual distance between speech segments, where by ''speech segment'' we mean a piece of a speech signal of duration 50–150 ms—in particular, a diphone, i.e., a segment from the midpoint of one phoneme to the midpoint of the adjacent phoneme. In deriving such a measure, we will present a model of how the auditory system integrates auditory nerve activity over time and frequency. A measure of the perceptual distance would be of interest in its own right. It would also have practical applications. For instance, the influence of perturbations introduced by low bit rate coders may extend, in general, over segment length intervals. The design and evaluation of such coders should therefore benefit from the derivation of a perceptual distance of the type considered here. Also, we believe that such a perceptual distance would provide a robust measure for automatic speech recognition. This belief is based on the following reasoning. Human beings perform far better than any existing automatic speech recognizer, especially when the speech signal has been degraded due to variations in the transmission path, the presence of noise, articulatory changes induced by the noise (i.e., the Lombard effect), etc. Therefore, use of a distance measure based on perceptual dissimilarity may be expected to improve automatic speech recognition.

We approach the derivation of such a measure of distance in two main steps. First of all we conduct a psychoacoustical experiment relevant to the perception of speech. In traditional psychophysical experiments speech is rarely used as a test stimulus. Typically these experiments are concerned with (a) masking of steady-state signals by other steady-state signals (e.g., masking of tones by noise, noise by a tone, etc.); or (b) measurement of the just-noticeable difference (jnd) of some steady-state property (e.g., jnd for amplitude or frequency of a tone, jnd for formant frequencies or pitch, etc.). Speech, however, is a highly nonstationary signal. For processing speech signals, this nonstationarity is dealt with by partitioning the signal into contiguous ''frames'' (i.e., short time windows of about 20- to 30-ms duration). Within each frame the signal may be regarded as stationary. However, from frame to frame there is considerable nonstationarity. It is not clear how masking properties, jnds, etc., change due to this nonstationarity. Therefore the traditional studies cited above are of limited application to problems such as speech coding at low bit rates and automatic speech recognition. Not surprisingly, almost all progress in these areas has come from application of signal processing techniques, with little help from psychophysics.

This paper is aimed at improving this situation. In contrast to experiments of the type mentioned above, our experiment involves ''segment level'' properties, i.e., properties of the whole segment rather than those of individual frames. For concreteness we will consider diphones, although longer segments could be studied by similar methods. Our experiment is aimed at quantifying the relative importance of various time-frequency regions (which we call ''**tiles**'') for the perception of a given segment. To achieve this, we study the perceptual effects of modifying a selected tile **while at the same time leaving the information in all other time-frequency regions unchanged**. This experiment, which we call the ''tiling'' experiment is described in the next section.[1]

The second important step of our approach is to **simulate** the tiling experiment. The simulation depends upon a definition of distance between observation vectors based on the ensemble interval histogram (EIH). As discussed in

(Ghitza, 1994), the EIH is a functional model of how auditory-nerve firings are analyzed in the auditory periphery. The desired distance measure is derived by driving the performance of the simulated tiling experiment to mimic that of the human subjects.

The detailed description of these procedures given in the next few sections may be summarized as follows: The psychophysical paradigm used is the diagnostic rhyme test (DRT). The word pairs in the DRT are modified by interchanging judiciously selected time-frequency regions (tiles). This modified database is used in the standard DRT, and the error patterns induced by these changes are recorded. The same DRT is then simulated by an array of speech recognizers. These recognizers use a parametrized distance between EIH vector sequences derived from the speech waveform. The parameters of the distance metric are jointly optimized over relevant tiling conditions so as to mimic the error patterns of the human subjects. The optimal set of parameters then defines the desired perceptual distance metric. (Note that such an optimization can be performed with any choice of observation vectors. We chose EIH on the assumption that the optimized parameters will have relevance to human speech perception because, as mentioned above, the EIH is a functional model of the auditory periphery (Ghitza, 1994).

I. PSYCHOPHYSICS

The experiment used in our search for the perceptual distance is what we call the ''tiling'' experiment. It has been described in a recent publication (Ghitza, 1993a), so we will give only a brief description of it here. In this experiment we measured the relative importance of various time-frequency ''tiles'' by studying the perceptual effects of modifying these tiles one at a time, or by simultaneously modifying various combinations of these tiles. It is important to note that when a particular tile (or a combination of tiles) is modified, the information in the rest of the time-frequency plane is left unaltered. In this way we measure the perceptual importance of that tile (or combination) in the presence of all other time-frequency information.

For the psychophysical paradigm we have chosen the DRT, which was first suggested by Voiers (1983), and which has been in extensive use for evaluating speech coders. In the DRT, Voiers uses 96 pairs of confusable words spoken by several male and female speakers. All the words are of the consonant-vowel-consonant (CVC) type, and the words in each pair differ only in the initial consonant. In an attempt to uniformly cover the speech subspace associated with initial diphones, the DRT database was designed such that the target diphones 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 Jakobson *et al.* (1952)[2] and the target consonants in each pair differ in the presence or absence of one of these dimensions. An explanation of these attributes, as well as the complete list of words, may be found in Ghitza (1993a).

The database is used in a very carefully controlled psychophysical procedure. The listeners are well trained and quite familiar with the database, including the voice quality

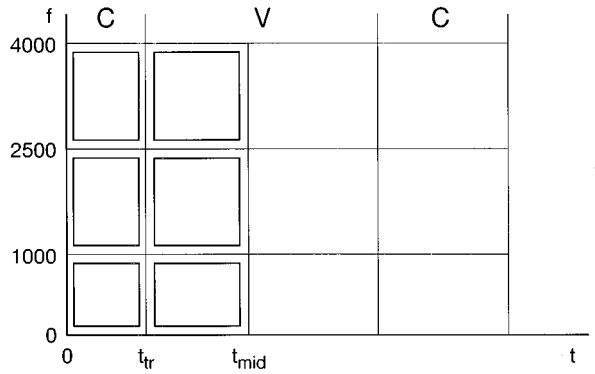

FIG. 1. Configuration of the six time-frequency tiles chosen for the tiling experiment.

of the individual speakers. As for the psychophysical procedure, a one-interval two-alternative forced-choice paradigm is used. A word pair is selected at random and displayed as text on a screen. One of the words in the pair (selected at random) is next presented aurally, and the subject is required to indicate which of the two words was heard. The procedure is repeated until all the words in the database have been presented. The errors made by the subjects are recorded.

For the tiling experiment the DRT was conducted on several distorted versions of Voiers' standard database. The details of the signal processing involved in creating those distortions may be found in Ghitza (1993a). Briefly, we divided the time-frequency plane into nonoverlapping regions called ''tiles'' that cover the target diphone in each pair of words in the DRT. Ideally, one should use many small tiles, but the experiments become increasingly time consuming and expensive with increasing number of tiles. From considerations of feasibility, we decided to use six tiles with the configuration shown in Fig. 1. The six regions were chosen on the basis of the following rough reasoning: On the time axis a break at the boundary between the C and V portions of the target diphone is an obvious choice. This boundary as well as the midpoints of the C and the V were hand labeled by a trained phonetician (see Ghitza, 1993a). On the frequency axis two breaks were selected. A break at 1 kHz is suggested by the known change in the properties of nerve firings at approximately this frequency (e.g., loss of synchrony beyond 1 kHz). A break at 2.5 kHz corresponds roughly to the upper limit of the second formant frequency (Peterson and Barney, 1952). We will call the resulting frequency regions as band-1 (0–1 kHz), band-2 (1–2.5 kHz), and band-3 (2.5–4 kHz).

Each distorted database was generated by interchanging a particular tile (or a combination of tiles) between the target diphones of each of the 96 pairs of words in the database. Such an interchange is illustrated in Fig. 2, in which the tile selected is the consonant part of the target diphone between 1 and 2.5 kHz. In a similar manner, a total of 14 distorted versions of the database were created. As described in Ghitza (1993a), special care was taken to minimize artifacts in the speech signals due to the interchange operation.

A DRT test was performed on the original database as well as on each of these distorted versions. The error for each

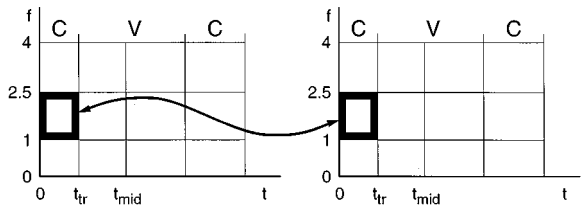

FIG. 2. Illustrating the interchange of tiles for a pair of words in the DRT database.

word pair, for each of these distortion conditions was recorded for each of three speakers and each of eight listeners. As described in Ghitza (1993a), these experiments demonstrated that perceptually, the interchange of the entire diphone in each band is far more dominant than the interchange of the consonant part or the vowel part alone. Therefore, to derive the parameters of the perceptual distance we used only the original, undistorted database and the distorted versions corresponding to the interchange of the entire diphone in band-1, band-2, and band-3, respectively.

These error patterns constitute the psychophysical data which we would like to mimic with a simulated DRT test. The simulated DRT described in the next section uses the same speech waveforms as were used in the psychophysical test. The optimization procedure used to drive the error patterns of the simulated test toward those of the human subjects is discussed in Sec. III.

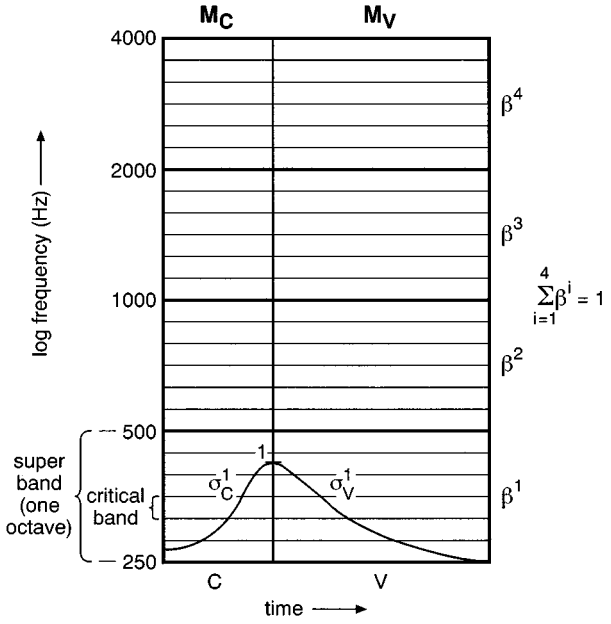

FIG. 3. A schematic diagram describing the parameters used in the perceptual distance metric. The matrices $\mathbf{M}_C$ and $\mathbf{M}_V$ capture the intuitive notion of ''lateral interaction'' between neighboring critical bands. The γ_n^i's weight the vector distance near the transition point more heavily than vector distances further away from it. The β^i's reflect the relative importance of the ith superband, in the presence of normal activity in all other superbands, and are subject to the constraint $\Sigma_{i=1}^4 \beta^i = 1$.

II. SIMULATION

The method of simulating the DRT has been described in Ghitza (1993b), and the reader is referred to that article for details. Recall that the DRT is a one-interval, two-alternative forced-choice experiment. At each stage the subject knows which pair of words has been selected and that one of them will be presented at random. The subject must make a decision and indicate which word was heard. We therefore postulate that for each of these binary decisions, the subject is able to retrieve from memory a recognizer optimized for that pair of words.

In view of this postulate, we **simulate** the DRT by replacing the human subject by an array of recognizers (one for each pair of words in the database). The particular type of speech recognizer that we use in the simulation has also been described in a recent article (Ghitza and Sondhi, 1993), so we will not describe it in detail here. Suffice it to mention that the recognizers utilize hidden Markov models with **nonstationary** states, where each state is a template of a diphone. When used in the DRT, each recognizer in the array reduces to a binary recognizer for a pair of initial diphones, since the second diphone of the CVC is identical for the two words in each pair.[3] Thus correct recognition occurs if and only if the initial diphone of the test utterance is closer to the initial diphone of the correct word model than to the initial diphone of the other word model of the pair. The word models were derived as follows: Every speaker in the DRT database provides two repetitions of each word. We assign one of these repetitions to be the ''training'' database and the other to be the test database. The set of word models is obtained from the training database by hand segmentation. If more data were available for training, more accurate word models could be derived. We believe, however, that by restricting our experiment to a speaker-dependent mode we reduce acoustic variability, so that word models derived from just one repetition are accurate enough.

The errors made in this simulation are entirely governed by the definition of distance between the test diphone and the model diphone. The parametric form chosen for this distance is therefore of crucial importance for the successful derivation of the perceptual metric. Let us discuss briefly the parametrization that we have chosen, and some of the considerations which led us to this choice. For this discussion, it will be helpful to refer to Fig. 3.

We begin by defining a diphone as a sequence of feature vectors—one for each frame of the speech signal—roughly 100 frames per second. Our choice of feature vector is a 24-dimensional EIH vector, with the histogram bins allocated on the ERB scale. (ERB stands for equivalent rectangular bandwidth, which is the bandwidth of a hypothetical rectangular filter that approximates the critical band of the human auditory filters. See detailed definition in Ghitza, 1994.) As mentioned earlier, the EIH is a functional model of how auditory-nerve firings are analyzed in the auditory periphery (Ghitza, 1994).

Let $\mathbf{x}_k$, $k=1,2,\ldots,K$ be the EIH vectors of some given test sequence $\mathbf{X}$. We need to define the distance of $\mathbf{X}$ from a template (or state) sequence $\mathbf{S}$ with EIH vectors $\mathbf{s}_n$, $n=1,2,\ldots,N$. The length N of the template sequence is in general, different for different templates. The length K of the test sequence is arbitrary within some broad range of values.

Note that each component of the EIH vector functionally represents the activity in a local region of the frequency spectrum (roughly a critical band). We postulate that these components are processed in groups—or **superbands**—to arrive at a distance within each group, and then the individual distances are combined to give the overall distance between the EIH vector sequences **X** and **S**. In the auditory periphery, presumably, there is a continuum of overlapping superbands covering the frequency spectrum. In our functional model we replace this continuum by four nonoverlapping superbands—roughly one octave wide. Each superband consists of six ERB bins of the whole EIH vector.

In order to define a distance between **X** and **S** we first time align the two sequences by warping **X** onto **S**. For this we use the usual dynamic time warp (DTW) algorithm,

$$\Delta^2(\mathbf{S},\mathbf{X}) = \frac{1}{N} \min_{k(n)} \sum_{n=1}^{N} d^2(\mathbf{s}_n, \mathbf{x}_{k(n)}), \tag{1}$$

where $k(n)$ is the time warp [with $k(1)=1$ and $k(N)=K$], and d is the Euclidean distance between $\mathbf{s}_n$ and $\mathbf{x}_{k(n)}$. Let $\tilde{\mathbf{x}}_n$, $n=1,2,\ldots,N$, be the EIH vectors of **X** after this alignment. We next define the distance between individual vectors $\tilde{\mathbf{x}}_n$ and $\mathbf{s}_n$. Let $\mathbf{s}_n^i$ and $\tilde{\mathbf{x}}_n^i$ be the ith subvectors of $\mathbf{s}_n$ and $\tilde{\mathbf{x}}_n$, respectively, representing the ith superband.[4] Let $\mathbf{m}_n^i$ be a 6×6 matrix defined for the ith superband, for the time index n of the template sequence **S**. Define $\hat{\mathbf{s}}_n^i$ and $\hat{\mathbf{x}}_n^i$ as the unit length vectors

$$\hat{\mathbf{x}}_n^i = \frac{\mathbf{m}_n^i \tilde{\mathbf{x}}_n^i}{\|\mathbf{m}_n^i \tilde{\mathbf{x}}_n^i\|}, \tag{2a}$$

$$\hat{\mathbf{s}}_n^i = \frac{\mathbf{m}_n^i \mathbf{s}_n^i}{\|\mathbf{m}_n^i \mathbf{s}_n^i\|}, \tag{2b}$$

where $\|\cdot\|$ denotes the Euclidean norm, or length, of a vector. We next define the distance between the subvectors $\tilde{\mathbf{x}}_n^i$ and $\mathbf{s}_n^i$ by the relation

$$d(\mathbf{s}_n^i, \tilde{\mathbf{x}}_n^i) = \|\hat{\mathbf{s}}_n^i - \hat{\mathbf{x}}_n^i\|. \tag{3}$$

With this definition of distance between vectors, the distance between the sequences within a superband, $\mathbf{X}^i$ and $\mathbf{S}^i$ is defined as

$$D^2(\mathbf{S}^i,\mathbf{X}^i) = \frac{1}{N} \sum_{n=1}^{N} \gamma_n^i d^2(\mathbf{s}_n^i, \tilde{\mathbf{x}}_n^i). \tag{4}$$

Here, $d^2(\mathbf{s}_n^i, \tilde{\mathbf{x}}_n^i)$ is weighted by the factor γ_n^i depending upon position along the template sequence.

In general, the distance D between the EIH sequences **S** and **X** can be any function of the $D^2(\mathbf{S}^i,\mathbf{X}^i)$'s. For the present we assume that they are linearly combined. Thus

$$D^2(\mathbf{S},\mathbf{X}) = \sum_{i=1}^{4} \beta^i D^2(\mathbf{S}^i,\mathbf{X}^i), \tag{5}$$

where $D^2(\mathbf{S}^i,\mathbf{X}^i)$ is as defined in Eq. (4), and the β^i are subject to the constraint $\Sigma_{i=1}^4 \beta^i = 1$. The β^i's reflect the relative importance of the ith superband, in the presence of normal activity in all other superbands. (See Fig. 3.)

With vector distance defined as in Eqs. (2) and (3), the matrices $\mathbf{m}_n^i$ may be regarded as the submatrices of a 4×4 block diagonal matrix $\mathbf{M}_n$, in which each block has dimension 6×6. The entries in the matrices $\mathbf{M}_n$, $n=1,2,\ldots,N$, are part of the set of parameters to be determined by optimization as discussed in the next section. We have allowed the matrices $\mathbf{M}_n$ here to depend arbitrarily on the time index n of the template. With this generality, however, the number of parameters to be optimized becomes too large. We therefore restrict $\mathbf{M}_n$ to two possibilities: $\mathbf{M}_n = \mathbf{M}_C$ if the index n is in the consonant portion of the template, and $\mathbf{M}_n = \mathbf{M}_V$ if it is in the vowel portion. If the matrices $\mathbf{m}_n^i$ are chosen to be diagonal, then they serve to specify the relative importance of different components of the EIH vector (or, essentially, different critical bands). With a more general structure, they can capture the intuitive notion of ''lateral interaction.'' That is, the notion that the output of a channel might be influenced by the activity in neighboring channels. In our study we chose the $\mathbf{m}_n^i$ matrices to be tridiagonal.

The γ_n^i are introduced because we believe that the vectors near the transition point are more important for recognizing the diphone than vectors further away from it. With this in mind, we specify the function γ_n^i with just two parameters, $\sigma^i \equiv (\sigma_C^i, \sigma_V^i)$. These are the variances of two Gaussian curves with peaks at the transition point—one for the consonant part and one for the vowel part. (See Fig. 3.)

In principle, the set of parameters that define the distance D in Eq. (5) (e.g., the matrices $\mathbf{M}_C$ and $\mathbf{M}_V$ and the parameters σ^i and β^i) should be allowed to be different for different diphones. This is again not feasible because of the number of parameters involved. Note that the total number of diphones is on the order of 2000 in English. In the DRT database alone, the number of diphones is 192. Unique matrices for each diphone would require an enormous number of parameters. We therefore restrict the number of parameters by using the same sets for diphones with ''similar'' properties. At present we group together consonants into seven categories according to **manner** of articulation (voiced and unvoiced stop, voiced and unvoiced fricative, nasal, glide, and affricate). The vowels are grouped into four categories according to the location of the constriction (low back, high back, low front, and high front). This gives us 28 classes of diphones, and we assign a parameter set to each such class.

In summary, the distance of a test segment **X** from a diphone template **S** is derived as follows: Depending on the template **S**, choose the appropriate parameter set $\mathbf{M}_C$, $\mathbf{M}_V$, σ^i, and β^i. Then compute the distance according to Eqs. (1)–(5). For a given specification of all the parameters, the definition of D gives us a parametrized distance which depends on the template (or state). The entire set of parameters is optimized to best mimic human performance as described in the next section.

Finally, let us note that the parametrization described above is not necessarily optimal. Indeed we believe it can be improved in several ways. Allowing a greater range of choices for the matrices $\mathbf{M}_C$ and $\mathbf{M}_V$ as well as allowing the submatrices $\mathbf{m}_n^i$ to be full matrices (rather than tridiagonal)

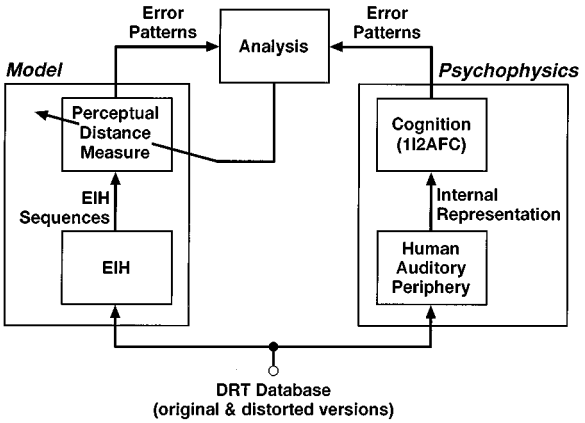

FIG. 4. A schematic diagram describing the optimization procedure. The parameters of the perceptual distance measure are iteratively adjusted to match the error patterns produced by the machine to those of the human subjects, jointly over several tiling conditions. In the box marked ''cognition,'' the abbreviation 1I2AFC stands for the ''one-interval two-alternative forced-choice'' paradigm.

are obvious possible improvements in the definition of the vector distance d. A more promising improvement is a generalization of the form of the segment distance D in the following manner: Let $\mathbf{V}^s$ be the $(24N)$-dimensional vector obtained by concatenating the N vectors $\mathbf{s}_n$ in the template sequence. Similarly let $\tilde{\mathbf{V}}^x_\kappa$ be the test sequence after warping on to the template sequence with a mapping $\kappa \equiv k(n)$. Then we can define the distance between $\mathbf{S}$ and $\mathbf{X}$ to be given by

$$D^2(\mathbf{S},\mathbf{X}) = \min_\kappa \mathbf{V}^{s\prime}\mathbf{\Phi}\tilde{\mathbf{V}}^x_\kappa, \tag{6}$$

where $\mathbf{\Phi}$ is a positive definite matrix and $'$ denotes matrix transpose. Here $\mathbf{\Phi}$ can be regarded as a block matrix with N^2 blocks, each block being a 24×24 matrix. Then the D of Eq. (5) is a special case of the D of Eq. (6) in which $\mathbf{\Phi}$ is a block diagonal matrix. With a full matrix $\mathbf{\Phi}$, we again have the problem of a large number of parameters to be estimated. As a first step, therefore, we might just generalize to a block tridiagonal matrix.

III. OPTIMIZATION

A schematic description of the optimization procedure is shown in Fig. 4. The right-hand side depicts the psychophysical data collected in the tiling experiment. The left-hand side shows the outputs of the simulated tiling experiment. The parameters of the simulation are iteratively adjusted to best mimic the psychophysical data.

Let $\boldsymbol{\theta}$ denote the parameter set which goes into the definition of D in Eq. (5), i.e., the $\mathbf{M}_C$'s, $\mathbf{M}_V$'s, σ^j's and β^j's. These are the adjustable parameters. In addition we have the template sequences $\mathbf{S}_j$, $j=1,192$—one for each initial diphone in the (undistorted) training database. These are kept fixed throughout the optimization procedure. For a given set of values for the parameters $\boldsymbol{\theta}$ (and the fixed templates) we define a cost function C which quantifies how badly the simulation performs when compared to the psychophysical data. Once C is defined, we use an optimization program to iteratively adjust the parameters $\boldsymbol{\theta}$ in order to minimize C. The program we use is a variant of Newton's method (Gay, 1983).

To complete the description of our optimization procedure, let us now indicate the definition of the cost function C. The data we are attempting to mimic are the responses of each of eight listeners to presentations of each of the words in the database spoken by each of three speakers, and distorted by each of K tiling conditions. (As mentioned in Sec. II, for optimization we chose only the distortions corresponding to the interchange of the entire diphone in band-1, band-2, and band-3, respectively. This gives $K=3$.) To give a concise definition for C it is convenient to define a set of indices.

Let:

p denote the word *p*air in the database $(1\leqslant p\leqslant 96)$,

w denote the *w*ord within a pair $(1\leqslant w\leqslant 2)$,

k denote the tiling condition $(0\leqslant k\leqslant K)$, with 0 for the undistorted database,

s denote the *s*peaker $(1\leqslant s\leqslant 3)$,

l denote the *l*istener $(1\leqslant l\leqslant 8)$.

Let x_{skpw} denote the speech signal corresponding to a chosen set of indices. Then the human response is a binary number, $h_l(x_{skpw})$, for each selection of the indices l,s,k,p,w in the range given above for each index. That is, h is 0 if listener l identified x_{skpw} correctly, and 1 otherwise. For a given set of parameters, $\boldsymbol{\theta}$, let $m(x_{skpw},\boldsymbol{\theta})$ denote the machine's response.

With these definitions, the cost function that we minimize is defined as[5]

$$C=\sum_{s=1}^{3}\sum_{k=0}^{K}\sum_{p=1}^{96}\sum_{w=1}^{2}\sum_{l=1}^{8}[h_l(x_{skpw})-m(x_{skpw},\boldsymbol{\theta})]^2. \tag{7}$$

The optimal solution $\boldsymbol{\theta}^*$ represents the parameters of the perceptual distance which provide the best mimic, jointly over all K tiling conditions. The accuracy of $\boldsymbol{\theta}^*$ depends upon the amount of data used for the optimization. In our database the number of tokens per **manner** class ranged between 36 and 165; the number of tokens per vowel category was 144.

IV. RESULTS

In terms of evaluating the validity of our approach, two questions come to mind. First, how closely can the machine error patterns be made to match human error patterns? And second, how does the performance of the ''optimal'' metric—derived by optimizing on ''tiling'' type of distortions—generalize to other kinds of distortions?

Figures 5(a)–(d) and 6(a)–(d) present results in an attempt to answer the first question. In all these figures we present error patterns by plotting the error rates for each of the Jakobson–Fant-Halle dimensions.[6] For each of these dimensions we plot two error rates.[7] At the abscissa marked ''+'' we plot the error rates for the subset of words in which the attribute is present, and at the abscissa marked ''−'' we plot the error rates for the words in which the attribute is absent. In the top panel of every figure the error rates for human subjects (solid line) are compared to those for the

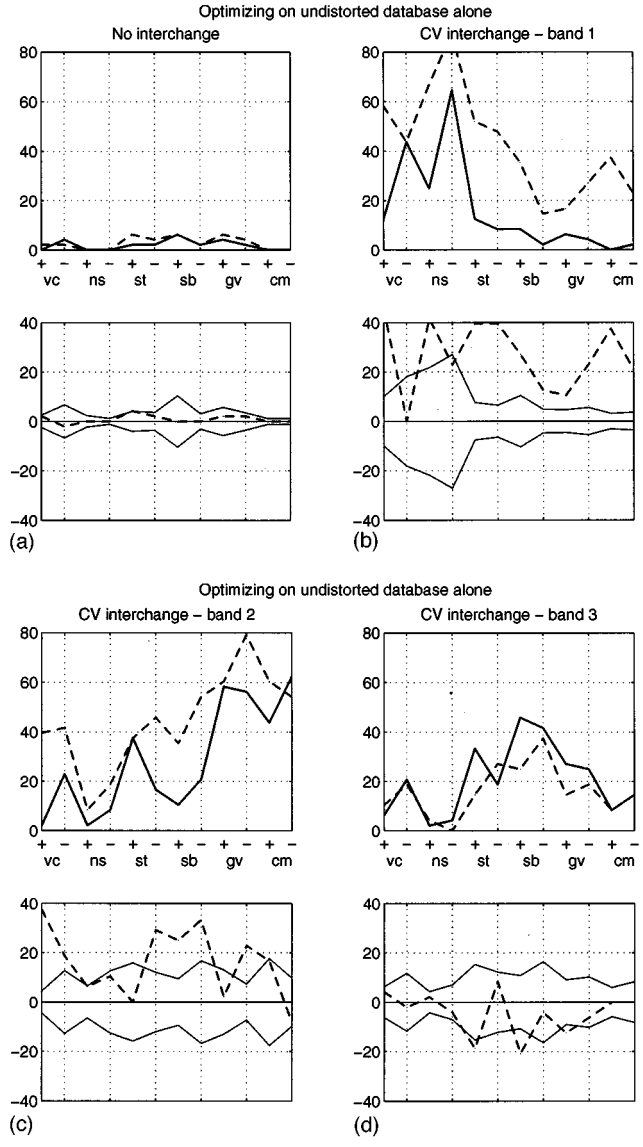

FIG. 5. Optimizing on undistorted database alone. *Top panels:* Mean human (solid line) and machine (dashed line) performance on the DRT database. The mean human performance is derived across three speakers and eight subjects. The abscissa of every plot indicates the six phonemic categories: ''**vc**'' 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 ordinate represents the number of words in the 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 as a percentage (relative to 16 which is the total number of words per phonemic category). *Bottom panels:* Difference between mean human performance and the machine performance (dashed line) compared to the human standard deviation (solid lines). The plots are for the original database (a) and for the three tiled versions obtained by interchanging bands 1, 2, and 3 of the entire diphone [(b), (c), and (d), respectively].

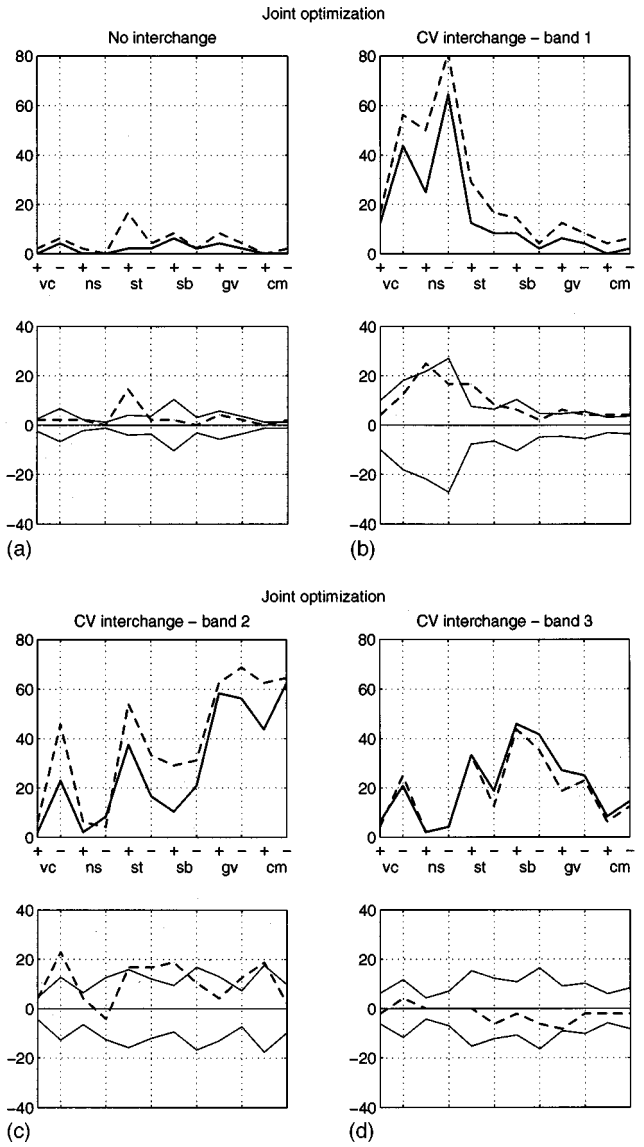

FIG. 6. Same format as in Fig. 5, but for joint optimization.

machine (dashed line). In the bottom panel the dashed line shows the error rate for the machine minus the error rate for the human subjects. Also plotted for comparison are two solid lines representing ± one standard deviation of the error rate for human subjects.

We first tested if the model structure is flexible enough for the purpose of mimicking the human performance for one tiling condition alone—say, the undistorted, original, DRT database. For this test we optimized the parameters of the distance metric, $\boldsymbol{\theta}$, on just the tiling condition 0. As it is seen in Fig. 5(a), the model can mimic the performance of the human subjects quite well—the difference between machine and human performance (bottom panel) is within one standard deviation for all the dimensions. However, this model (whose parameters were derived by optimizing on the undistorted database alone) fails to mimic human performance for other tiling conditions. Figure 5(b) and (c) shows that the machine makes significantly more errors than human subjects for the tiling conditions in which band-1 or band-2 is interchanged for the entire diphone. To arrive at a single model that is able to mimic human performance under different tile interchanges, $\boldsymbol{\theta}$ should be **jointly optimized** over several tiling conditions.

In Fig. 6(a)–(d) we show the same comparisons as in Fig. 5(a)–(d), except that now the parameters are optimized jointly over four tiling conditions: undistorted database and the tilings in which bands 1, 2, 3, respectively, are inter-

TABLE I. Experiment with the DRT database degraded by additive Gaussian white noise. The entries are the errors, summed over all the Jakobson–Fant–Halle dimensions, in percent.

	Clean	30 dB	20 dB	10 dB
Human	3	2	3	7
EIH (perceptual)	5	8	13	24
EIH (L_2)	18	17	21	27
MEL-CEP (L_2)	11	16	25	38

changed over the entire target diphone. It is seen from Fig. 6(a) that machine performance on the undistorted database is slightly worse than that in Fig. 5(a), and the machine performance for sustention is more than a standard deviation away from human performance. However, in exchange for this small deterioration, the performance for the other tiling conditions is now much closer to human.

As for the question of how the optimal metric generalizes, we ran a simulated DRT experiment on the DRT database degraded by additive Gaussian white noise. We used three different definitions of $d(\cdot,\cdot)$ (a) the observation vectors were 13th-order Mel-Cepstrum (MEL-CEP) and d was the L_2 (i.e., Euclidean) distance; (b) the observation vectors were EIH and d was the L_2 distance; and (c) the observation vectors were EIH and the distance metric was the optimized perceptual metric derived above. Table I shows the results for those three DRT simulations and for the human subjects, as a function of SNR. The entries are the errors, summed over all the Jakobson–Fant–Halle dimensions, in percent. From Table I we conclude that although the machine performance using EIH with perceptual metric does not match human performance, it is superior to the performance using EIH with L_2 metric (and also to the performance with L_2 norm between MEL-CEP vectors). Figure 7 shows the detailed distribution of these errors along the Jakobson–Fant–Halle dimensions. The detailed distributions are shown for the human subjects, the EIH with the perceptual distance and the MEL-CEP with the L_2 distance. The results for the EIH with L_2 distance were omitted in order not to clutter the figure. The figure demonstrates that the pattern of error distribution for the perceptual distance generally follows the pattern for human subjects.

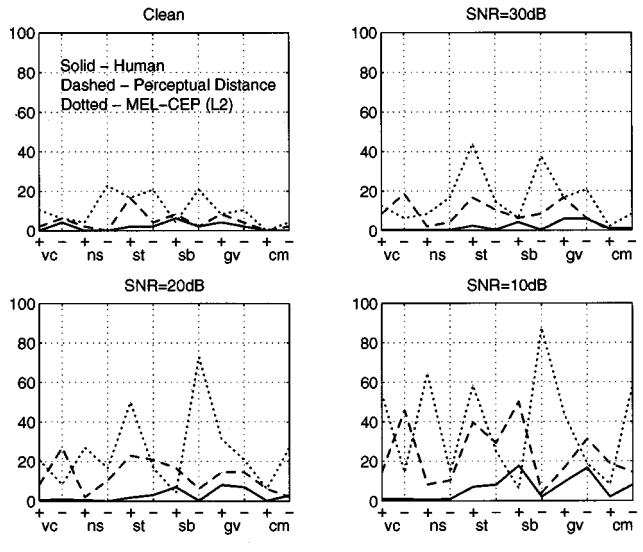

FIG. 7. Mean human performance (solid line), machine performance with the perceptual distance (dashed line) and machine performance with 13th order Mel-Cepstrum and L_2 distance (dotted line) for the DRT database in the presence of additive Gaussian white noise. The axes are as described in Fig. 5, for various SNR. (a) Clean speech, (b) SNR=30 dB, (c) SNR=20 dB, (d) SNR=10 dB.

V. DISCUSSION

In the preceding sections we have presented a method for deriving a perception-based measure of distance between speech segments. The segments we chose to investigate are diphones, although longer segments could be studied in a similar manner.

In our model, the template of a diphone is represented as a high-resolution sequence of EIH vectors (one vector every 10 ms). This template represents the articulatory gesture while moving from C to V, in terms of the time course of the EIH vectors. It may be thought of as the pattern of the diphone that is stored in memory during the early stages of language acquisition. Note that the template implicitly contains information about the **place** of articulation of the consonant. An unknown ''input'' diphone is compared to a template by first time warping it to the template and then computing a distance between the aligned sequences. This distance is expressed in terms of a set of parameters $\boldsymbol{\theta}$ which are allowed to depend upon the template. These parameters quantify the perceptual deviation from the diphone template. In order to keep the number of parameters manageable, we group the consonants into seven groups and the vowels into four groups, and assign the same parameters to the consonants and vowels within the same group. For consonants, we postulate that the parameters depend upon the **manner** of production (voiced and unvoiced stop, voiced and unvoiced fricative, nasal, glide, and affricate). The **vowels** are grouped according to the location of the constriction (front high, front low, back high, back low). In this way all C–V diphones are grouped into 28 different classes. Note that this grouping is only for the parameters $\boldsymbol{\theta}$ (that weight different time-frequency regions according to their relative perceptual importance). The templates themselves are not grouped.[8] Note also that two diphones, say /**ba**/ and /**da**/, whose consonants belong to the same **manner** class are assigned the same $\boldsymbol{\theta}$. The information about their different places of articulation is implicitly contained in the templates of the two diphones.

In deciding upon a structure for the distance we postulate that the auditory periphery processes the input in parallel frequency ''superbands'' (about an octave wide) and produces a distance in each such band. In our functional model we take four contiguous superbands, although in the auditory periphery there is presumably a continuum of overlapping bands. The distances from all superbands can be combined in many ways, providing the overall distance between the diphones. Here, we combine them linearly.

Throughout this study, we used the Jakobson–Fant–Halle feature space. These dimensions were used by Voiers to structure the DRT database, and we present our results along the same dimensions. It is worth noting, however, that

our aim is to derive a distance metric that is completely independent of the distinctive feature set. For this reason, $\boldsymbol{\theta}^*$ was computed by optimizing the cost function defined in Eq. (7). This cost function accumulates the contributions of the individual words in the database without regard to their distinctive features. However, the exact pairwise comparisons made will have some influence on the values of the optimal parameters.

The most important, and we believe novel, aspect of our work is the fact that we derive the distance measure on the basis of **perceptual** dissimilarity. We do that by mimicking human performance in the DRT framework, using tiling type of distortions. In this restricted task, at least, the metric performs significantly better than others that we have tried.

As a final note, we speculate that the perceptual distance derived here may be used to define a jnd for diphones (or phonemes). This jnd may be defined as a change for which the perceptual distance attains a threshold value.

ACKNOWLEDGMENTS

We would like to express our thanks to Ben Gold and the anonymous reviewers for many helpful suggestions to improve the manuscript.

[1]As mentioned above, psychophysical experiments dealing with speech perception are rare. Some experiments reported in the literature that have some relevance to the present paper are those of Fletcher (1953), Miller and Nicely (1955), Houtgast and Steeneken (1985), and Drullman *et al.* (1994). In Fletcher's experiments (Fletcher, 1953) subjects had to respond to stimuli that contained only parts of the speech signal (e.g., low-pass or high-pass filtered speech). Miller and Nicely (1955) studied the effect of filtering and additive noise on the confusion matrices for various phonemes. The experiments of Houtgast and Steeneken (1985) and Drullman *et al.* (1994) are concerned with the effects of filtering the speech envelope in contiguous frequency bands. Our experiments differ from all these in that we study the effects of modifying selected time-frequency regions of a speech signal while leaving the rest of the signal unchanged.

[2]The six Jakobson–Fant–Halle dimensions are *voicing*, *nasality*, *sustention*, *sibilation*, *graveness*, and *compactness*. The *voicing* (**vc**) feature characterizes the nature of the source, being periodic or nonperiodic. The *nasality* (**ns**) feature indicates the existence of a parallel resonator representing the nasal cavity. The terms *sustention* (**st**) and *sibilation* (**sb**) are due to Voiers. They correspond, respectively, to the continuant-interrupted and strident-mellow contrasts of Jakobson *et al.* (1952). Finally *graveness* (**gv**) and *compactness* (**cm**) represent broad resonance features of the speech sound, related to place of articulation.

[3]We are assuming that the effects of coarticulation due to the initial consonant do not extend beyond the midpoint of the vowel. This appears to be an accurate assumption, at least for the DRT database.

[4]Throughout the paper we used subscripts to indicate time index of a template, and superscripts to indicate superbands.

[5]If m in Eq. (7) is chosen to be a binary number, like the human responses, then C would be a discontinuous function of m which could be difficult to optimize. We therefore make $m(x_{skpw}, \boldsymbol{\theta})$ a real number between 0 and 1, whose value depends upon the distances of x_{skpw} from the two templates for the pair of words p. If d_{cor} and d_{inc} are the distances from the correct and incorrect templates, respectively, then we choose $m(x_{skpw}, \boldsymbol{\theta}) = [1+\arctan \alpha(d_{\text{inc}} - d_{\text{cor}})]/2$. The exact value of α is not very critical. The important property is that if $d_{\text{cor}} \gg d_{\text{inc}}$ then m goes to 0 and if $d_{\text{cor}} \ll d_{\text{inc}}$ then it goes to 1.

[6]The errors could be presented in other ways, e.g., along place-manner dimensions, or in the form of a confusion matrix. Note, however, that a confusion matrix format is inappropriate here because the psychophysical paradigm is a two-alternative forced-choice, and also because many binary comparisons are missing in the database. As to the choice of distinctive features, we chose the Jakobson–Fant–Halle dimensions because (a) Voiers' DRT database is organized along those dimensions and, (b) because those dimensions reflect acoustic properties in time and frequency (Jakobson *et al.* 1952).

[7]Note that here, we use the notion of ''error rate'' in the context of DRT, i.e., a binary decision paradigm: an occurrence of an error means that the listener ''switched'' to the opposite category.

[8]It may be argued that the articulatory gestures are quite similar for C–V diphones in which the vowel is the same and the place of articulation of the consonant is the same—e.g., /**ma**/ and /**ba**/. However, the corresponding spectra and EIH vectors are still quite distinct. Hence grouping of the templates themselves is not justified.

Drullman, R., Festen, F. M., and Plomp, P. (**1994**), ''Effect of temporal envelope smearing on speech reception,'' J. Acoust. Soc. Am. **95**, 1053–1064.

Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Krieger, Huntington, NY).

Gay, D. M. (**1983**). ''Algorithm 611—Subroutines for unconstrained minimization using a model/trust-region approach,'' ACM Trans. Math. Software **9**, 503–524.

Ghitza, O. (**1994**). ''Auditory models and human performance in tasks related to speech recognition and speech coding,'' IEEE Trans. Speech Audio **2**(1), 115–132.

Ghitza, O. (**1993a**). ''Processing of spoken CVCs in the auditory periphery. I. Psychophysics,'' J. Acoust. Soc. Am. **94**, 2507–2516.

Ghitza, O. (**1993b**). ''Adequacy of auditory models to predict internal human representation of speech sounds.'' J. Acoust. Soc. Am. **93**, 2160–2171.

Ghitza, O., and Sondhi, M. M. (**1993**). ''Hidden Markov Models with Templates as Nonsationary States: An Application to Speech Recognition,'' Comput. Speech Lang. **7**(2), 101–119.

Houtgast, T., and Steeneken, H. J. M. (**1985**). ''A review of the MTF concept in room acoustics and its use for estimating speech intelligibility,'' J. Acoust. Soc. Am. **77**, 1069–1077.

Jakobson, R., Fant, C. G. M., and Halle, M. (**1952**). ''Preliminaries to speech analysis: the distinctive features and their correlates,'' Technical Report No. 13, Acoustic Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

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

Peterson, G. E., and Barney, H. L. (**1952**). ''Control methods used in a study of the vowels,'' J. Acoust. Soc. Am. **24**, 175–184.

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