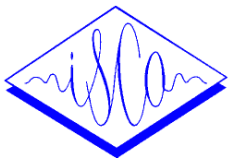

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ON THE PERCEPTUAL DISTANCE BETWEEN SPEECH SEGMENTS

Oded Ghitza and M. Mohan Sondhi

AT&T Bell Laboratories
Murray Hill, New Jersey 07974, USA

ABSTRACT

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. (By speech segments we mean pieces of a speech signal, of duration 50-200 milliseconds. For concreteness we will consider a segment to mean a diphone.) Such a distance metric would be useful for low bit rate speech coders because perturbations introduced by such coders typically last for several tens of milliseconds. It would also be useful for automatic speech recognition on the assumption that mimicking human behavior will improve recognition performance. Yet a third use for such a metric would be to define a just noticeable difference for diphones (a "phonemic" JND). (If a diphone is perturbed, how far from the original must the perturbed diphone be, in order to be perceived as a different diphone?) In this talk we will describe our attempts at defining such a metric.

I. INTRODUCTION

This paper is concerned with psychoacoustical experiments relevant to the perception of speech. In the past, such experiments have been concerned with the perception of what we may call "frame level" properties, that is, properties that can be derived by examining speech through a short (20-30 millisecond) time window. 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 pitch of a tone, JND for formant frequencies, etc.). Speech, however, is a highly nonstationary signal, and it is not at all clear how masking properties and JND's change due to this nonstationarity. Therefore these studies are of limited application to problems such as speech coding at low bit rates and automatic speech recognition. Almost all progress in these areas has come from application of signal processing techniques, with little help from psychophysics.

In this paper we will describe our attempts at improving this situation. We will consider psychophysical experiments involving "segment level" properties, where by segments we mean pieces of speech signals with durations of the order of 50-200 milliseconds. For concreteness we will consider diphones, although longer segments could be studied by similar methods. The main problem we will address here is the derivation of a "perceptual distance" between two such segments of (in general) unequal duration. Useful measures of distance between two speech signals have, of course, been proposed in the past. Our point of view is different, however, in that we would like the distance to have perceptual relevance.

A measure of the perceptual distance would be of interest in its own right. It would also have practical applications. For instance, perturbations introduced by low bit rate speech coders extend 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 a perceptual distance would provide a robust measure for automatic speech recognition.

Our approach to the problem may be described briefly as follows: The 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 based on a parametric distance function. The parameters are optimized so as to mimic the error patterns of the human subjects. In the following sections we will describe these steps in somewhat greater detail.

In Section II we will summarize the DRT paradigm, which is well known as a tool for the evaluation of speech coders. We will also describe the way in which we simulate the paradigm by replacing the human subjects by an array of automatic speech recognizers. In Section III we will describe the interchange of time-frequency tiles alluded to above. This "tiling" experiment has also been described in a recent paper, so we will only summarize it briefly. Finally, in Section IV, we will discuss the optimization procedure and the degree of success achieved by the simulation in mimicking human error patterns.

II. THE DIAGNOSTIC RHYME TEST (DRT)

Psychophysics

For the psychophysical paradigm we have chosen the DRT, which was first suggested by Voiers [8], and which has been in extensive use for evaluating speech coders. We will discuss our reasons for this choice after first describing the test.

In the DRT, Voiers uses 96 pairs of confusable words spoken by several male and female speakers. All the words are of the CVC type, and the words in each pair differ only in the initial conso-

nant. [More recently, Voiers has assembled another database in which the words in each pair differ in the *final* consonant. The corresponding test based on this database is termed the Diagnostic ALliteration Test (DALT). Yet a third database has been recently developed in which the words are of the VCV type, and the words in each pair differ only in the *medial* consonant. The corresponding test is termed the Diagnostic Medial Consonant Test (DMCT). In our experiments we have used the DRT and DALT, but have not yet utilized the DMCT. In this paper, to avoid repetitious statements, we will describe the experiments in terms of the DRT. All statements apply, with obvious modification, to the DALT and DMCT as well.]

The target diphones (initial for DRT, final for DALT and medial for DMCT) 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, Fant and Halle [5]. The dimensions are voicing, nasality, sustension, sibilation, graveness and compactness, 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 for the DRT and DALT may be found in [2].

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 of the individual speakers. 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 and may be analyzed in various ways, as discussed in Section IV.

The conditions of the paradigm are such that the subject is given as much of the "cognitive" information about the stimuli as possible, so that the errors may be attributed entirely to the peripheral processing in the auditory pathway. This is the aspect of perception that we are interested in. And the fact that the DRT allows us to focus on it, is the main reason for our choice of this paradigm.

Simulation

As mentioned above, the subject is given as much of the cognitive information about the stimuli as possible. We make the assumption that the subject is able to utilize this information. Thus, when presented with an utterance to be identified, the subject is able to process it through two models (one for each word in the pair displayed visually) and choose the one judged closest. To simulate the DRT, therefore, we implement an array of automatic speech recognizers, one for each pair of words in the database. The unknown utterance is examined by the appropriate recognizer and scored by the models for each of its two words. The utterance is classified as the word whose model gives the best score.

This method of simulating the DRT has been described in [3], and the reader is referred to that article for details. The particular type of speech recognizer that we use in the simulation is also described in a recent article [4], so we will not describe it in detail here. Suffice it to mention that the recognizer comprises Hidden Markov Models with nonstationary states, where each state is a template of a diphone. When used in the DRT, the recognizer is essentially a recognizer of the initial diphone, since the second diphone of the CVC is identical for the two words in each pair. Thus correct recognition occurs if and only if the initial diphone of the utterance is closer to the model for the initial diphone of the correct word than to the model of the other word of the pair.

The errors made in this simulation are entirely governed by the definition of distance between the test diphone and the model diphone. This distance may be defined in parametric form in a variety of ways. Our definition is as follows: Define a diphone as a sequence of feature vectors — one for each frame. Our choice of feature vector is a 30-dimensional EIH [1]. Let $\mathbf{O} \equiv [\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_N]$ and $\hat{\mathbf{O}} \equiv [\hat{\mathbf{o}}_1, \hat{\mathbf{o}}_2, \cdots, \hat{\mathbf{o}}_M]$ be the sequences of feature vectors for the test utterance and the template, respectively. For the template we define two matrices: M_c for the consonant, and M_v for the vowel. In terms of M_c and M_v we define the distance between two vectors $\mathbf{o}$ and $\hat{\mathbf{o}}$ as

$$d(\mathbf{o}, \hat{\mathbf{o}}) = (\mathbf{o}-\hat{\mathbf{o}})' M' M(\mathbf{o}-\hat{\mathbf{o}})$$

where M is replaced by M_c if $\hat{\mathbf{o}}$ is in the consonant portion of the template, and by M_v if it is in the vowel portion. With this definition of distance between vectors, the distance between the sequences $\mathbf{O}$ and $\hat{\mathbf{O}}$ is defined as

$$D(\mathbf{O},\hat{\mathbf{O}}) = \min_{n(m)} \sum_{1}^{M} d(\mathbf{o}_{n(m)}, \hat{\mathbf{o}}_m).$$

This is the usual Dynamic Time Warp (DTW) distance, except for the introduction of the matrices M.

The matrices M_c and M_v may, in general, be different for every template. This, however, is not feasible. Note that the total number of diphones is on the order of 2000 in english. Even the number of diphones in the DRT and DALT is about 400. We therefore restrict the number of matrices by using the same sets for diphones with "similar" properties. At present we group together consonants into six categories (bilabial, labio-dental, dental, alveolar, palatal and velar) and the vowels into four categories (low back, high back, low front amd high front). This gives us 24 classes of diphones, and we assign a set of matrices to each such class.

We also have the freedom of choosing the structure of the matrices M_c and M_v. Again, it is not feasible to use full 30×30 matrices. We have tried diagonal and tridiagonal structures for them.

When all the parameters in all the matrices have been specified, the definition of D gives us a parametrized distance which depends on the template (or state). The choice of parameters is optimized so as to match the error patterns of the simulated DRT to the error patterns of the human subjects, in the experiments to be defined in the next section.

III. THE TILING EXPERIMENT

The psychophysical experiment used in our search for the perceptual distance is what we call the "tiling" experiment. Details of this experiment may be found in [2]. Briefly, we divide the time-frequency plane into non-overlapping regions called "tiles" that cover the target diphone in each pair of words in the DRT (or DALT). 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 that we could use six tiles. The six regions were chosen as illustrated in Fig. 1. The selection was made on the basis of the following rough reasoning: On the time axis a break at the boundary between the C and V is an obvious choice. The break at 1 kHz is suggested by the known change in the properties of nerve firings at approximately this frequency. The break at 2.5 kHz corresponds roughly to the upper limit of the second formant frequency [7].

We interchange a tile (or some combination of tiles) between the target diphones of each pair in the database, as illustrated in Fig. 2. A total of 14 different distorted versions of the database are created in this way. Each of these versions is used in both the psychophysical and the simulated DRT, as described in Section II. The error pattern induced by each of these distortions is recorded. Some examples of the patterns of errors along the Jakobson, Fant, Halle dimensions are shown in Fig. 3.

IV. THE OPTIMIZATION

Let us denote by **M** the set of all the parameters entering the 24 sets of matrices M_c and M_v defined in Section II. Starting with a trial set of parameters, the error patterns for the simulated DRT are computed for each of the distorted databases of the tiling experiment described in the last section. The parameters are optimized so as to minimize the difference between human and machine performance. This difference is measured by a cost function , C, defined as the squared difference between the human and machine errors, accumulated over all 14 DRTs with the 14 tiled versions. Thus

$$C = \sum_{tilings} \sum_{ti} (h_{ti} - m_{ti})^2,$$

where h_{ti} and m_{ti} are the errors made by the human and machine respectively, for the i-th dimension and the t-th tiling. In order to make C a continuous function of the parameters **M**, a "soft" definition of error is used. Thus if a test diphone is at a distance D_1 from the correct model, and at D_2 from the incorrect model, then the soft error e_s is defined as

$$e_s = \frac{1 + arctan\,[\,k\,(D_1 - D_2)]}{2}$$

As k becomes large e_s approaches 0 if the test is closer to the correct diphone, and 1 if it is closer to the incorrect diphone.

Since it is not possible to analytically compute the gradient of the cost function C, we use a gradient-less optimization procedure. The one we have chosen is the simplex method [6].

V. REFERENCES

[1] Ghitza, O. (1994). "Auditory models and human performance in tasks related to speech recognition and speech coding", *IEEE trans. on Speech and Audio, SAP-2(1)*. Special issue on Neural networks for Speech Processing, 115-132.

[2] Ghitza, O. (1993). "Processing of spoken CVCs in the auditory periphery: I. Psychophysics", *Journal of the Acoustical Society of America, 94(5)*, 2507-2516.

[3] Ghitza, O. (1993). "Adequacy of auditory models to predict internal human representation of speech sounds", *Journal of the Acoustical Society of America, 93(4)*, 2160-2171.

[4] Ghitza, O. and Sondhi, M. M. (1993). "Hidden Markov Models with Templates as Nonstationary States: An Application to Speech Recognition", *Computer Speech and Language, 7(2)*, 101-119.

[5] 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, Mass.

[6] Nelder, J. A. and Mead, R. (1965). "A Simplex Method for Function Minimization", *Computer Journal, 7*, 308-313.

[7] Peterson, G. E. and Barney, H. L. (1952). "Control methods used in a study of the vowels", *Journal of the Acoustical Society of America, 24*, 175-184.

[8] Voiers, W. D. (1983). "Evaluating processed speech using the Diagnostic Rhyme Test", *Speech Technology, 1(4)*, 30-39.

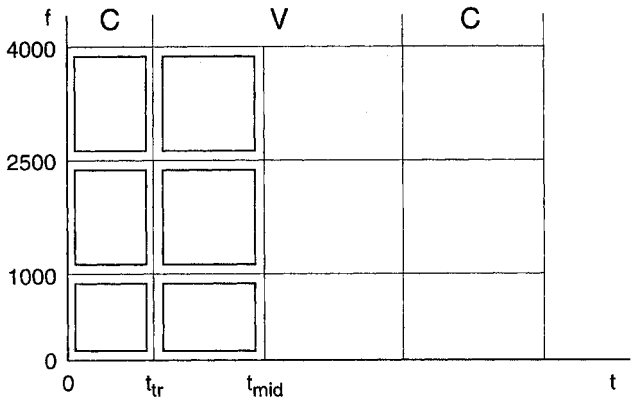

Figure 1. A diagram of the time-frequency region occupied by a spoken CVC word. The time-frequency region of the initial diphone is sub divided into 6 "tiles". The frequency boundaries (from the bottom up) are 0 Hz, 1000 Hz, 2500 Hz and the highest frequency in the band, say 4000 Hz. The time landmarks are (from left to right) the beginning of the word (t = 0), the transition from the initial consonant to the vowel ($t = t_{tr}$) and the mid-point of the vowel ($t = t_{mid}$). For stop consonants, t_{tr} is the transition from the stop release to the vowel.

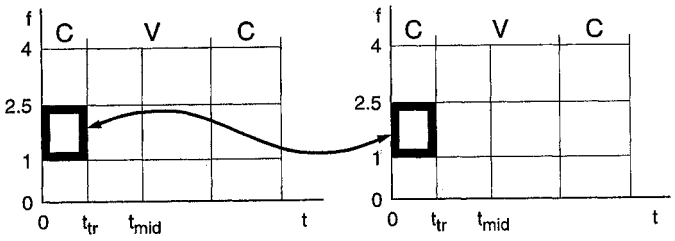

Figure 2. A diagram of the time-frequency region occupied by a prototype DRT word-pair, where the regions corresponding to the initial diphones are divided into 6 tiles each. The interchange of one of the tiles is illustrated.

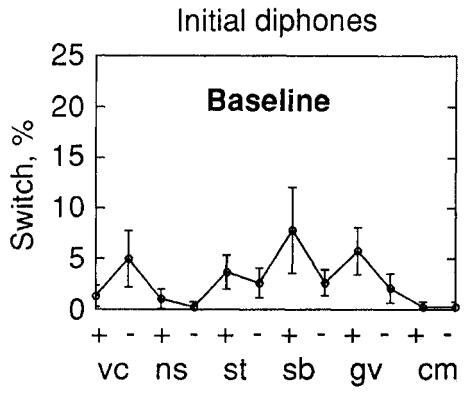

Figure 3(a). The average and the 95% confidence interval for the DRT without any interchange of tiles. The abscissa of every plot indicates the 12 phonemic categories: "vc" is for Voicing, "ns" for Nasality, "st" for Sustention, "sb" for Sibilation, "gv" for Graveness and "cm" for Compactness". The "+" sign stands for attribute present and the "-" sign for attribute absent. The ordinate is termed "switch", and it 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 of 16 (the total number of words per phonemic category).

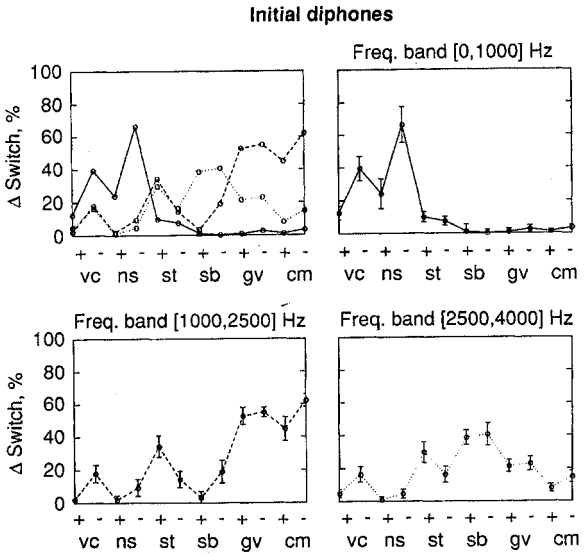

Figure 3(b). Human performance on the DRT database, under interchange of each frequency band over the entire diphone. The upper left plot is a summary of the other 3 plots, with the confidence-interval bars omitted. The abscissa is as in Fig. 3(a). The ordinate is termed "Δ*switch*", since it represents the *additional* number of switches, relative to the baseline version, that occured due to the particular interchange operation. 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 interchange condition. The upper right plot shows the amount of Δ*switch*, in percent, under interchange of the 0-1kHz band. The lower left plot is for the 1 kHz - 2.5 kHz band, and the lower right plot is for the 2.5 kHz - 4 kHz band. Notice that *Voicing* and *Nasality* are strongly correlated with the first frequency band of the diphone, *Graveness* and *Compactness* with the second frequency band of the diphone, and *Sibilation* with the third frequency band of the diphone.