Distance Measures for Speech Recognition--Psychological and Instrumental*

Paul Mermelstein

ABSTRACT

Perceptual confusion among speech sounds can serve as a guide to the selection of appropriate distance metrics for verification of hypotheses in speech-recognition systems. Known results covering psychological representation of speech sounds are first reviewed. Desirable properties for distance measures for verification are stated, and previously proposed distance metrics for word-recognition are evaluated in this light. This paper reports on one experiment that demonstrates the need for assessing the significance of local differences by any distance metric to be used for verification of syllable-sized hypotheses concerning the speech signal.

INTRODUCTION

Analysis of the continuous speech signal to obtain a phonetic transcription is a significant problem for any speech-understanding system. Speech sounds undergo a complex reorganization of their acoustic properties, from their form when uttered in isolation, to their form in a sentence context. This reorganization is generally accompanied by a loss of information; distinctive differences among sounds become reduced and sometimes disappear altogether.

Analytic segmentation and labeling rules may be constructed to extract the segments of speech that are characterized by unchanging features (Mermelstein, 1975). Due to variations in context and speaker, however, these rules are at best probabilistic in nature, as they only select a highly likely hypothesis concerning the underlying segments. The rules are based on acoustic measurements pertaining only to a short-time interval of the signal in and around the hypothesized segment.

To utilize information from a somewhat larger context, one attempts to verify the analysis-derived hypotheses at the syllable or word level. Word boundaries are not readily apparent in fluent speech; therefore one wants to consider the verification of syllable-sized units. By restricting our analysis to admissible syllables of the language, both those found within words and those

---

*This paper was presented at the Joint Workshop on Pattern Recognition and Artificial Intelligence, Hyannis, Mass., 1-3 June 1976.

Acknowledgment: This work was supported in part by the Advanced Research Projects Agency, Department of Defense.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

spanning word boundaries, we can immediately reject a large number of hypotheses. Additionally, knowing the syllable context, we can utilize predictions concerning the effects of neighboring sounds on each other in order to ascertain whether the data in fact support those hypotheses.

We first review some results concerning human perceptual confusions among speech sounds in order to select an appropriate representation on which to compute distance measures. Next, several desirable properties are cited for a distance metric appropriate for the verification of syllable-length hypotheses. Distance measures previously used for limited word-recognition systems possess these properties to a variable extent. Distance-based recognition is generally inappropriate for selecting one of more than a few hundred distant patterns. For a fixed finite probability of error for any individual membership comparison, the recognition probability tends to zero as the number of patterns is increased. Therefore, we suggest that analysis be used to select only a few reasonable hypotheses concerning the phonetic content of a syllable, and conventional word-recognition techniques be limited to verification of such hypotheses. In order that a metric be appropriate for verification as well as recognition, we require not only that the distance to the correct category be a minimum, but also that such minima lie below a fixed threshold, and distances to incorrect categories lie above that threshold. Finally, we cite a simple experiment whose results emphasize the need for weighting the short-time spectral distances according to the significance of the local differences.

Psychological Distance Representation

Experimental data on confusion among speech sounds by human listeners are available from perception and recall experiments. Miller and Nicely (1955) measured perceptual confusions among single initial consonants under various conditions of noise added to the speech signal. Wickelgren (1966) measured confusion among consonants that were perceived correctly in a serial recall experiment. The confusion patterns were generally similar. Essentially the same feature system could explain the confusions in auditory perception as in short-term memory. Where confusion exists, it can be viewed as the result of selective substitution of features such as voicing, nasality, openness, and place. Similarity among consonants was found to be a monotonic function of the number of features they share. Where confusion among consonant-vowel and vowel-consonant sequences was tested, the order was not significant for vowel errors but was a feature of consonant errors.

Shepard (1972) derived a similarity matrix from the Miller-Nicely confusion data and obtained a spatial representation of the speech sounds. He assumed that similarity is an exponentially decreasing function of interclass distance and minimized the error between the similarity and its distance derived representation,

$$\sum_{i>j} \{S_{ij} - (e^{-bD_{ij}} + c)\}^2$$

$S_{ij} = (p_{ij} + p_{ji})/(p_{ii} + p_{jj})$ is a function of the reported confusion matrix.
D_{ij} is the distance between classes i and j in the spatial representation

recovered, given by $\sqrt{\sum_k (X_{ik} - X_{jk})^2}$, where X_{ik} is the projection of the coordinate of the i^{th} class on the k^{th} orthogonal dimension of the underlying perceptual space. Parameters to be determined are b and c. Over 99 percent of the variance for confusion among 16 consonants was accounted for on the basis of two orthogonal dimensions. These dimensions corresponded roughly to the perceptual features of voicing and a combination of nasality and frication.

This spatial representation is shown in Figure 1. A hierarchical clustering procedure which sequentially clusters sound pairs in the order of their

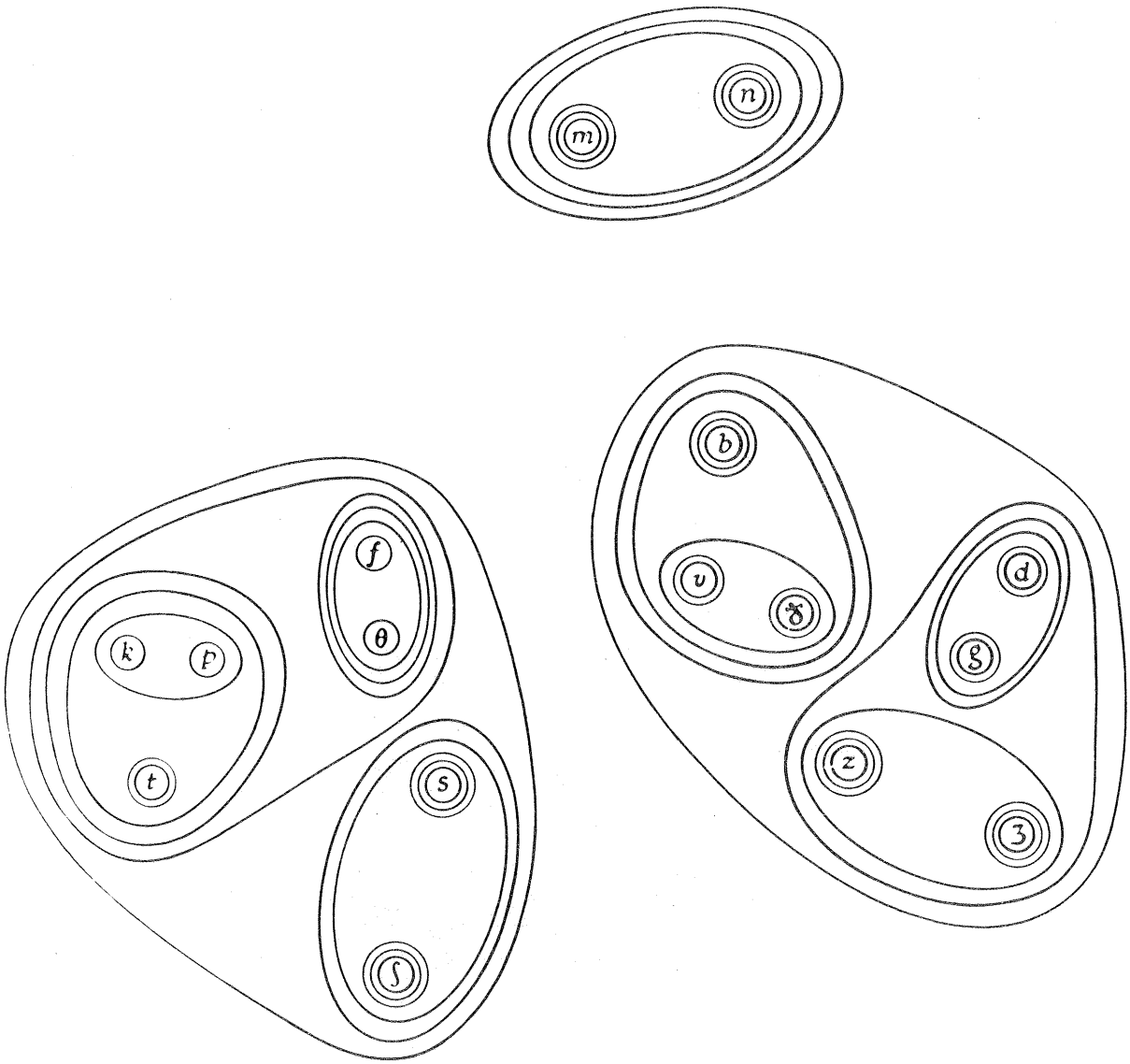

Figure 1: Spatial and hierarchical representation of the perceptual similarity between consonants. From Shepard (1972), copyright 1972, McGraw-Hill, Inc. Used with permission of McGraw-Hill, Inc.

similarity yields the clusters indicated. These clusters roughly correspond to those one derives on the basis of confusions at decreasing levels of signal to noise ratio. There appears to be a good correlation between the similarity values under different noise conditions--decreasing signal to noise increases the confusion among similar sounds.

It is significant to note that the sound space is not uniformly populated. A distance sufficiently large to cross the boundary between /p/ and /k/ is probably not significant for variation among different tokens of /s/. The technique relies on confusion data; therefore, the distance between distinct tokens of members of the same phonemic category is assumed to be zero. Since any continuous instrumental measure must be sensitive to both intercategory and intracategory variation, these results can only be used as a guide to the construction of an appropriate distance metric.

A similar spatial distribution can be achieved for vowel sounds and is given in Figure 2. Although the data are shown in three dimensions, which

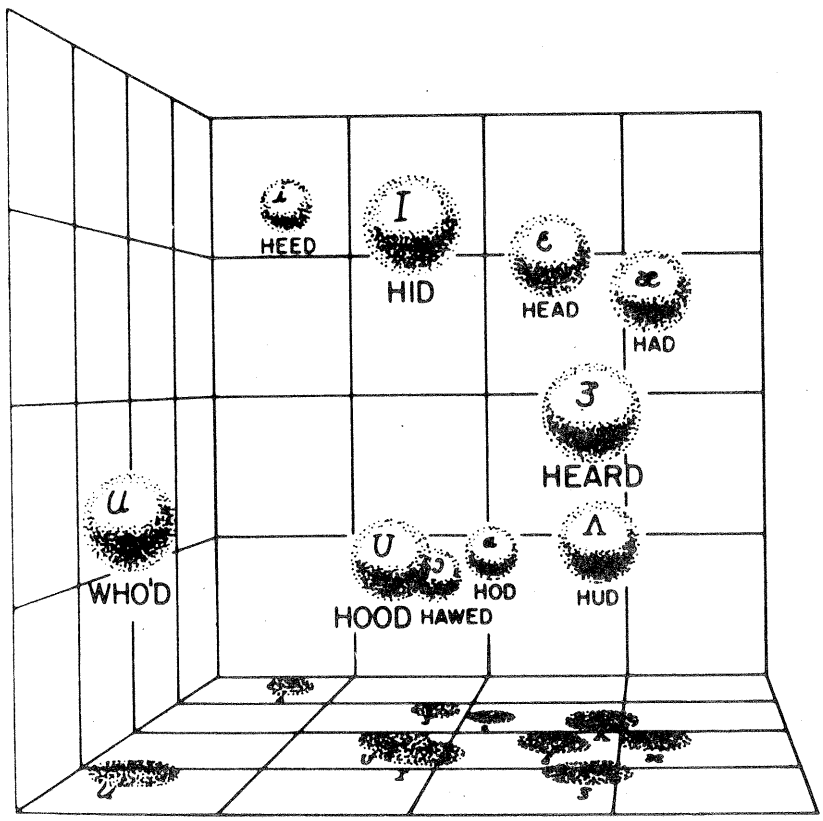

Figure 2: Three dimensional spatial representation for 10 vowel phonemes. From Shepard (1972), copyright 1972, McGraw-Hill, Inc. Used with permission of McGraw-Hill, Inc.

account for 99 percent of the variance, the first two dimensions account for 97 percent. While the principal dimensions correspond roughly to the first two formant frequencies of the vowels, the second dimension appears to be compressed roughly logarithmically with frequency. These results correlate well with known data concerning the spacing of critical bands in the human auditory system--the band within which noise effectively masks a signal of fixed frequency. These critical bands are about equally spaced with frequency below 1000 Hz, increasing logarithmically thereafter. The mel-frequency scale reflects that spacing.

Confusion between vowels and consonants seems quite rare, but no data are available. It is unfortunate that the semivowels and glides were not included in the Miller-Nicely confusion experiments since these would have yielded the most interesting consonant-vowel confusion data.

Compound consonants present additional problems. Despite the close fusion in articulation between the component consonants of a compound, the confusions of the compounds can be explained in terms of the confusion of the components (Pickett, 1958). This result may be due to phonological constraints among the compounds. Since stops and fricatives are relatively rarely confused, the classes of compounds in which they participate will also be rarely confused. Confusion predominates among the stop-liquid compounds in initial and the nasal-stop group in final position.

According to Wickelgren (1966) consonant similarity and vowel similarity can be considered as independent dimensions in syllable recall. However, co-articulation effects modify the acoustic cues for consonants, depending on the syllabic vowel. Therefore the possibility of perceptual interactions between consonant and vowel must be recognized.

Desirable Distance Measure Properties

In view of the above results, a distance measure that models human performance should ideally recognize the phonemes, and construct the distance measure from phoneme confusability data. Failing such recognition, we can at best approximate the peripheral, precategorical aspects of human speech perception behavior.

Let us postulate a set of desirable properties for a distance measure for the verification of syllable-sized segments.

1. The measure should operate on time-aligned versions of the tokens to ensure consonant-to-consonant and vowel-to-vowel comparison. Since syllables have but one prominent vowel, the best aligned tokens can be viewed as those that will minimize vowel-vowel differences as well as differences in the prevocalic and postvocalic position.

2. If the final distance measure is a time integral of some distributed distance function, an appropriate weighting function that assesses the significance of the contributions from the individual short-time segments must be used.

3. The distance measure between tokens should be symmetric, $D(X,Y) = D(Y,X)$.

4. It should be possible to utilize the distance measure to determine phonetic equivalence. If X and Y are phonetically equivalent, but X and Z are not, the $D(X,Y) < D(X,Z)$.

5. Let A,B be parametric representations of two tokens, then $M(A,B) = M(B,A) = (A+B)/2$ is a template for the class (A,B) such that $D(A,M) \leq D(A,B)$ and $D(B,M) \leq D(A,B)$.

Templates are used as compact descriptors for equivalence classes. Consider the class of metrics defined as the weighted sum of elemental metric components for short-time segments. Let P be some space of time-warping transformations such as shown in Figure 3:

$$D(X,Y) = \min_{p(\tau)\varepsilon P} \sum_{\tau} w(\tau) \; d[x(\tau),y(\tau)]$$

where $d(\tau) = d[x(\tau), y(\tau)]$ is an elemental metric component over a short-time segment of the path $p(\tau)$ that maps $1 \leq t_x \leq T_x$, and $1 \leq t_y \leq T_y$ onto τ and $w(\tau)$ is some positive semidefinite weighting function that assesses the significance of the contribution from each element of the path.

Among requirements that we may want to impose on the elemental distance metric between any two short-time segments are

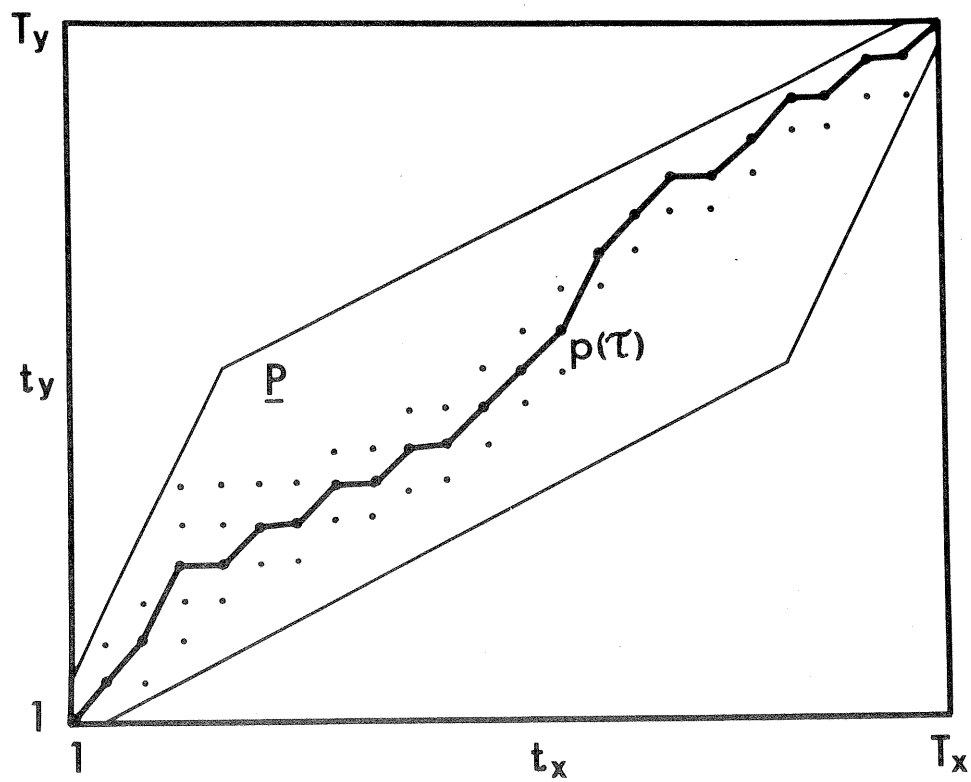

Figure 3: Typical path in the space of time-alignment transformations between two speech segments.

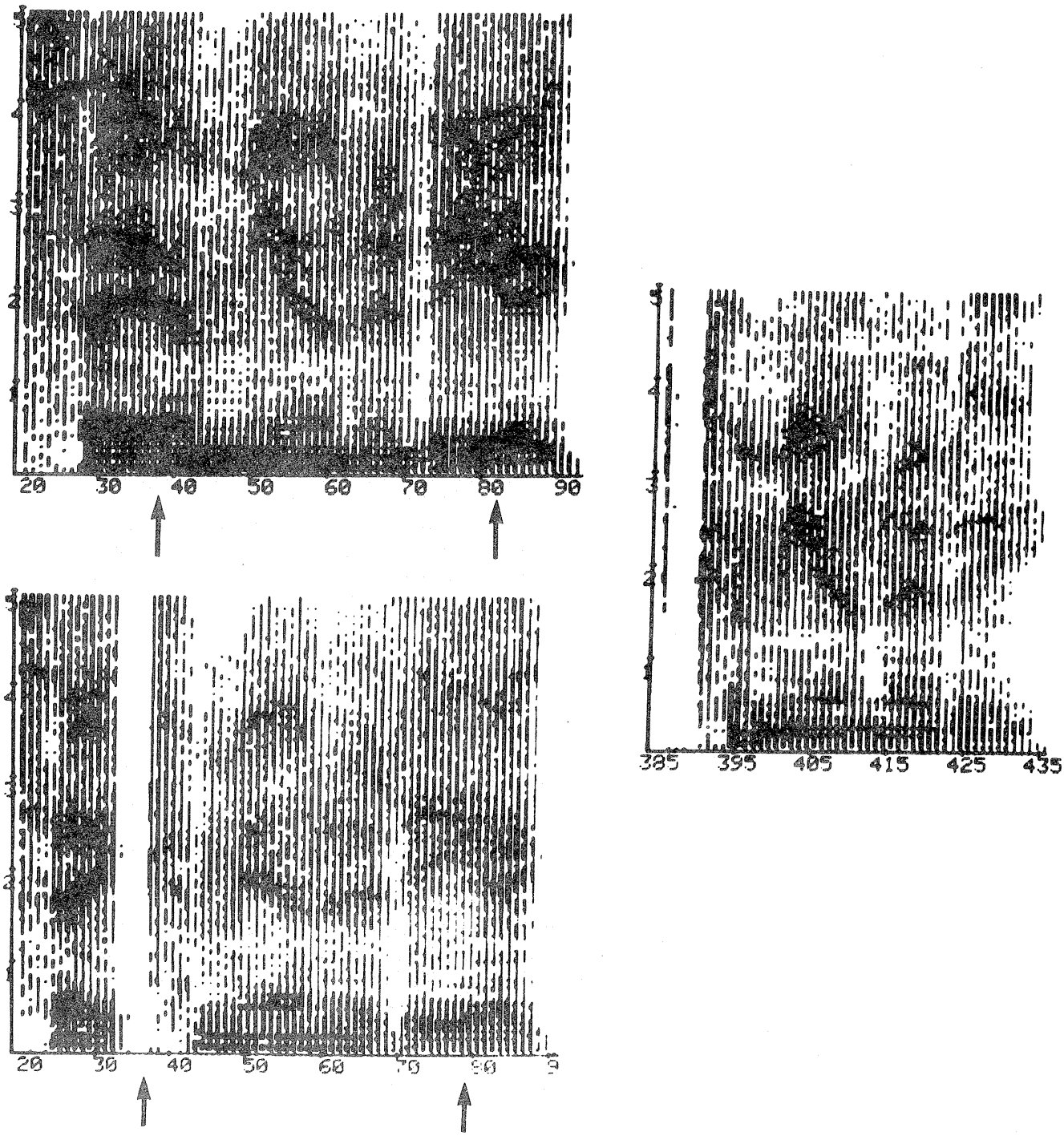

Figure 4: Spectrograms for the reference words "immunity" (top), "community" (bottom), and unknown (right). Frequency in kHz units, time in increments of 12.8 msec.

1. positive semidefinite, $d(x,y) \geq 0$ (this ensures that the global metric is also positive semidefinite),

2. symmetric, $d(x,y) = d(y,x)$,

3. that it satisfies the triangle inequality $d(x,y) + d(y,z) \geq d(x,z)$.

4. that it satisfies a perceptual weighting of the frequency components of the power spectra of the signals. If variation in $s(\omega_1)$, the energy at frequency ω_1, is perceptually more significant than that in $s(\omega_2)$, then $d[x, x + \Delta s\ (\omega_1)] > d[x, x + \Delta s\ (\omega_2)]$.

The need for careful assessment of the significance of spectral variations was realized when we carried out the following experiment (Nye, Copper, and Mermelstein, 1975). Human spectrographic pattern recognizers were asked to match the words of an unknown sentence, presented in spectrographic form, with the same words from a reference library of spectrographic patterns. The reference library was generated by the same speaker, stored in computer retrievable form, and displayed through specification of a list of required features. Since the phonetic transcription of the reference words was not made available, the subjects wide discouraged from using syntax and semantics to assist the pattern matching operation. While the subjects had no problem in rejecting the phonetically dissimilar words, they encountered frequent confusions between similar words. Figure 4 shows the two reference words "community" and "immunity" at left, and the unknown word at the right. In the presence of some uncertainty concerning the word boundary, the disagreement in the unstressed syllable at the top just to the right of the first arrow was accepted by two observers in view of the wide agreement over the rest of the word. The region of significant spectral disagreement between the two extends for no more than 100 msec. Clearly we need a rather sophisticated metric to resolve such distinctions.

Acoustics Based Distance Measures

Let us now examine some distance measures proposed previously in the light of these requirements. Sokoe and Chiba (1971) constructed an Euclidean distance metric on short-time spectral samples obtained from a bank of band-pass filters. When the words were aligned in time through use of a dynamic programming algorithm to minimize the total word-to-word distance, they achieved 99 percent recognition of the 100 two-digit Japanese numbers of five speakers. Klatt (1976) has proposed weighting the spatial distance metric with a function that reflects the increased perceptual importance of differences near the spectral peaks, and reduced perceptual importance of the differences near spectral minima. Itakura (1975) suggested use of the minimum prediction residual as a distance measure for isolated word recognition. This measure computes the ability of the linear predictor that is optimum for the reference-word segment to predict the signal waveform of the target-word segment,

$$d(X/a) = \log\ (\underline{a}\ \underline{V}\ \underline{a}' / \underline{\hat{a}}\ \underline{V}\ \underline{a}')$$

That is, the distance between the target segment characterized by process X and the reference segment, having the optimum linear-prediction vector $\underline{a}$, is given by the log-likelihood ratio where $\underline{\hat{a}}$ is the optimum linear predictor of X, and $\underline{V}$ is

the vector of autocorrelation coefficients of X. While this measure can be computed rather quickly from the signal waveform, it is not symmetric between reference and target. To overcome this, Gray and Markel (1975) have suggested a symmetric modification of the linear-prediction residual, namely

$$d_s(X/a) = d(X/a) + d(a/X) .$$

The linear-prediction residual is a measure of the unpredicted signal energy. There is no attempt to assess the significance of the suboptimum prediction of the signal waveform. For some signals even a rough spectrum approximation appears adequate, for others a finer representation is required.

White and Neely (1975) performed a comparative evaluation of the Euclidean spectral distance measure and the one based on the linear-prediction residual. He found them roughly equivalent in terms of performance for recognition of a 36-word and a 91-word vocabulary of one speaker. They concluded that the major improvement over previous results arose from the use of the various dynamic programming algorithms for word alignment. Use of the dynamic programming technique for word recognition was first proposed by Velichko and Zagoruyko (1970).

Atal (1974) has used a non-Euclidean distance measure for speaker recognition, namely

$$d(\underline{\mu}_1, \underline{\mu}_2) = (\underline{\mu}_1 - \underline{\mu}_2)W^{-1} (\underline{\mu}_1 - \underline{\mu}_2)'$$

where the $\underline{\mu}_j$ are parameter vectors to be selected and W is the covariance matrix of $\underline{\mu}$. He explored representations in terms of linear-prediction coefficients, impulse response coefficients, autocorrelation function samples, predictor derived area functions, and cepstral parameters. The cepstral coefficients c_k are related to the linear-prediction parameters by

$$\sum_{k=-n}^{k=n} c_k e^{-jk\theta} = \ln[\sigma/|A(e^{j\theta})|]^2$$

where $1/A(e^{j\theta})$ is the linearly predicted signal spectrum and σ is the rms energy. Among the different parametric representations, the cepstral coefficients gave the highest speaker identification accuracy. Representation in terms of cepstral coefficients has the advantage that a set of coefficients of the same order can be averaged, and the result equals the cepstral representation of the average of the log power spectra (after normalization to unity gain). Use of the covariance matrix normalizes the contributions of the components of the parameter vectors independently of any linear transformations they may undergo.

Bridle and Brown (1974) used a set of 19 weighted spectrum-shape coefficients given by the cosine transform of the outputs of a set of nonuniformly spaced bandpass filters. The filter spacing is chosen to be logarithmic above 1 kHz and the filter bandwidths are increased there as well. We will, therefore, call these the mel-based cepstral parameters. Pols (1971) showed good word recognition results using only the three shape variation components maximally contributing total spectral shape variation. These components resemble the mel-based cepstral parameters rather closely in terms of their frequency variation. The mel-based cepstral parameters have the advantage that generally fewer parameters suffice for an adequate representation of the power spectrum than the

linear-prediction coefficient series. A truncated cepstral representation corresponds to a frequency-smoothed power spectrum, one from which evidence concerning the individual harmonics of the speech signal is missing. To the extent that the spectrum of the excitation signal is invariant between successive voiced segments of the speech signal, the mel-based cepstral measure corresponds to a mel-weighted summation of the difference between the two smoothed vocal tract transfer functions.

Experiments With a Mel-Based Cepstral Distance Measure

I have been concerned with the adequacy of a mel-based cepstral distance measure to discriminate phonetically similar words and syllables. To evaluate the contribution of time-dependent significance functions to an integrated distance measure, I conducted the following experiment: four speakers, two male, two female, recorded one production of each of the twelve phonetically similar words, "stick," "sick," "skit," "spit," "sit," "slit," "strip," "scrip," "skip," "skid," "spick," and "slid" in a reference context "say ___ again." The words were excised from the carrier by listening to a specifiable delimited segment of the signal. Spectra were computed for all the words and reduced to a two-dimensional cepstral respesentation. The respective interword distances were determined for all possible pairs of words by time alignment with Itakura's dynamic algorithm. The unweighted metric used was

$$d(a,b) = \frac{1}{N} \sum_{\tau=1}^{N} \sum_{k=1,2} [C_k^a(\tau) - C_k^b(\tau)]^2$$

$C_k^x(\tau)$, $\tau = 1, \ldots, N$; $x = a,b$; $k = 1,2$ are the time-aligned, two-dimensional, mel-based cepstral coefficient vectors for the two words. Figure 5 shows histograms of the interword distances for the same word spoken by two different speakers, as well as for all other pairs comparing different words spoken by the same or different speakers. The complete overlap between the two comparison categories is surprising. Although the unweighted distance measure is useful to differentiate phonetically distant words, it is clearly not applicable to the discrimination of phonetically similar words.

I next generated templates for each of the words by time warping the words of each speaker onto the one with longest duration using the same dynamic programming algorithm. The mean and variance of the first two cepstral parameters were next computed for the time-aligned versions and used as templates representative of the respective words. Next the weighted distance between each token x and template A was determined using the inverse of the variance for weighting each cepstral coefficient difference, for example,

$$d_w(x,A) = \frac{1}{N_A} \sum_{\tau=1}^{N_A} \sum_{k=1,2} [(C_k^x(\tau) - C_k^A(\tau))/\sigma_k^A(\tau)]^2$$

The time-alignment path, $\tau = 1, \ldots, N_A$ is now a function of the local cepstral variance, $[\sigma_k^A(t)]^2$.

A fixed distance threshold allowed the correct assignment of all but 2 of the 48 tokens to the appropriate word class. The two confusions arose through incorrect assignment of one token of "slit" to "sit" and one token of "spit" to

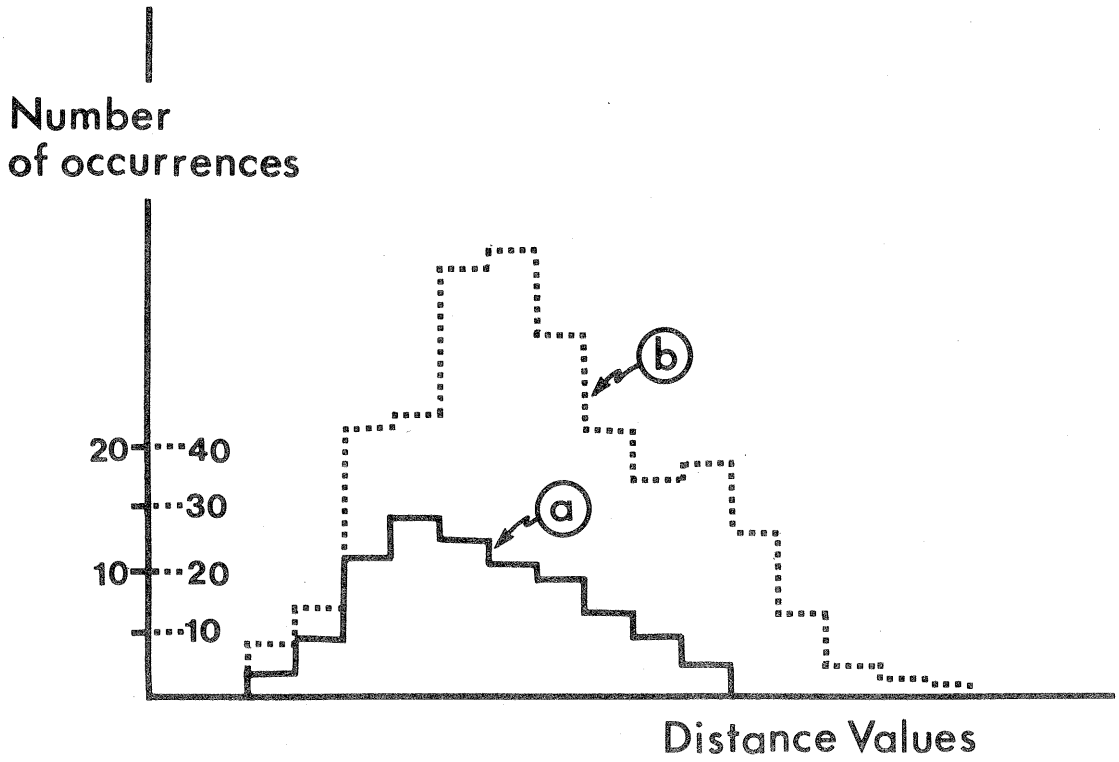

Figure 5: Histograms of computed interword distance values for (a) words from the same category (different speakers), (b) words from different categories (same or different speakers).

"spick." The same tokens were used to generate the template and to test them; therefore, this represents a biased test of discriminability. When I attempted to generate templates from fewer tokens, editing problems near the word edges, such as whether the release of the final stop was properly included, the result was significantly poorer discrimination. Nevertheless, the dramatic difference, as compared to the use of unnormalized distances, underlines the necessity of including appropriate modeling of the significance of the encountered variation of the parameters.

One result of using the inverse of the parameter variance for the weighting function, is to assign more significance to silent segments where the variance was actually zero (assigned a finite nominal value), than to the segments having finite energy. Since all our tokens began with the phoneme /s/, we could not explore the question of the relative weights to be assigned to fricatives and voiced sounds. Presumably, the relative cepstral distance among the class of unvoiced fricatives is larger than that among the vowels. Therefore one would want to tolerate larger differences in fricative regions than in vowellike regions before rejecting a given hypothesis.

A further desirable property of a time dependent weighting function appears to be the assignment of larger weights to regions of high spectral variation than to stationary regions. Otherwise, for steady-state segments the contributions to overall distance are proportional to the durations of the segments. Under those

conditions vowel differences would be overemphasized. No experimental results are as yet available on this point.

Discussion and Conclusions

Synthesis represents an alternative technique for generating the reference templates. Klatt (1975) and Cook (1976) have proposed a word verification procedure based on synthesis of the hypothesized word. Its use offers large potential savings in storage requirements at the costs of a small increment in processing requirements.

The prime motivation of using templates derived from actual productions at this point is the need to establish quantitatively the amount of speaker and context dependent variation for which verification techniques must provide. While synthesis procedures generally give us a perceptually acceptable representative of the class to which the token may be assigned, they provide no information concerning the admissible variation in the individual parameters. As we gain more insight into the relative significance of short-time variations in speech spectra and achieve an ability to model the process adequately, synthesis will undoubtedly become a more cost-effective procedure for the generation of templates. Until that time, however, one must resort to the generation of templates from actual productions in the exploration of hypothesis verification techniques.

Our attempt to utilize insights from speech perception processes as an aid to improved speech verification techniques suffers from an inability to separate the peripheral and central processes in human speech perception. There remains a large gap in our knowledge concerning the transformations that the signal undergoes before the segmental information is extracted. We do not yet have an adequate model of the extent of acceptable variation among tokens that belong to a segmental equivalence class. Nevertheless, known properties of perception may be used to guide us toward perceptually relevant representations of the speech signal. We have some evidence that improved verification results are obtainable by focusing on those representations of the speech signal which have proven to be of interest for human speech perception.

REFERENCES

Atal, B. S. (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55, 1304-1312.

Bridle, J. S. and M. D. Brown. (1974) An experimental automatic word recognition system. JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.

Cook, C. (1976) Word verification in a speech understanding system. Conference Record, IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, 553-556. (Available from IEEE, 345 East 47th Street, New York, N. Y. 10017.)

Gray, A. H., Jr. and J. D. Markel. (1975) COSH measure for speech processing. J. Acoust. Soc. Am., Suppl. 58, S97(A).

Itakura, F. (1975) Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 67-72.

Klatt, D. (1975) Word verification in a speech understanding system. In Speech Recognition, ed. by D. R. Reddy (New York: Academic Press), pp. 321-341.

Klatt, D. (1976) A digital filter bank for spectral matching. Conference Record, IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, 573-575. (Available from IEEE, 345 East 47th Street, New York, N. Y. 10017.)

Mermelstein, P. (1975) A phonetic-context controlled strategy for segmentation and phonetic labeling of speech. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 79-82.

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

Nye, P. W., F. S. Cooper, and P. Mermelstein. (1975) Interactive experiments with a Digital Pattern Playback. J. Acoust. Soc. Am., Suppl. 58, S105(A).

Pickett, J. M. (1958) Perception of compound consonants. Lang. Speech 1, 288-304.

Pols, L. C. W. (1971) Real-time recognition of spoken words. IEEE Trans. Computers 20, 972-978.

Sakoe, H. and S. Chiba. (1971) A dynamic-programming approach to continuous speech recognition. Reports of the 7th International Congress on Acoustics, Budapest, 20-C-13, 65-68.

Shepard, R. N. (1972) Psychological representation of speech sounds. In Human Communication, a Unified View, ed. by E. E. David and P. B. Denes (New York: McGraw-Hill), pp. 67-113.

Velichko, V. M. and N. G. Zagaruyko. (1970) Automatic recognition of 200 words. Intl. J. Man-Machine Studies 2, 223-234.

White, G. M. and R. B. Neely. (1975) Speech recognition experiments with linear prediction, bandpass filtering and dynamic programming. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-24, 173-188.

Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Am. 39, 388-398.