Interactive Experiments with a Digital Pattern Playback

Patrick W. Nye, Franklin S. Cooper, and Paul Mermelstein

ABSTRACT

Among the most useful tools for speech research have been those that enable spectrograms to be compared with one another, that provide ways of modifying speech data and that permit the user to listen to the modified speech signal. This paper reports an experiment in which such an interactive research tool—a Digital Pattern Playback (DPP)—was used to evaluate a spectrum-matching and dictionary-search technique for speech recognition. The DPP was used to display spectrograms of "unknown" sentences. An analyst divided these sentences into segments of word-length and listed their important acoustic features. Using these features, an interrogation program examined a feature-based spectrographic dictionary and recovered all the words having features that matched each unknown segment. When necessary, additional features were assigned to narrow the search. The reference spectrograms retrieved from the dictionary were compared, one at a time, with the spectrograms of the unknown sentence, and the best match was selected for each unknown segment. In general, the performance of the human analysts was found to be quite low, since only 26 percent of the words contained in the sentences were matched correctly. The paper concludes with a discussion of the factors governing human and machine performance on spectrogram matching.

INTRODUCTION

This paper describes results obtained from a speech analysis experiment that explored methods for organizing the information required for automatic speech recognition. The experiment required that the analysis operations be performed by two human subjects who worked from visual displays. These analysts studied the spectrogram, waveform, and amplitude functions of an unknown sentence and divided the sentence into word-length segments. Having listed the most salient features of each segment, the analysts then sought a set of matching reference words that were retrieved automatically from a feature-labeled dictionary. The identities of the reference words were not known to either of the analysts whose data are reported in this paper. Thus, syntactic and semantic considerations did not play a direct part in the selection of suitable matches.

*This paper was presented in part at the 90th meeting of the Acoustical Society of America, San Francisco, Calif. November 3-7, 1975.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]
Ingemann and Mermelstein (1975) have reported the results of some similar experiments that were carried out with conventional paper spectrograms. Their experience showed that the clerical problems became serious when subjects were required to work with reference libraries as large as 100 words. The present work represented a continuation of those experiments but avoided the inconvenience of handling volumes of paper by using a computer-based display system.

THE DISPLAY SYSTEM

The speech signals were displayed by an interactive research tool—called the Digital Pattern Playback (DPP)—which has been built around a PDP 11/45 and GT40 computer system (Nye, Reiss, Cooper, McGuire, Mermelstein, and Montlick, 1975). The system organization is sketched in Figure 1. The PDP 11/45 runs a general-purpose operating system allowing multiprogram access from several terminals. The GT40 supports the display functions. The analyst, seated at the keyboard, can selectively access the PDP 11/45 or the GT40. Using this facility, he may display two spectrograms lying one above the other on the same screen—each representing 1.6 secs of speech (see Figure 2). The lower spectrogram display field is usually occupied by a reference item that has been selected from the dictionary and installed there for direct comparison with the unknown. A cursor, controlled by a knob, can be moved to any point along the time axis of the upper, unknown spectrogram and the cross-section at that point can be displayed. A similar cross-section facility is also available for the lower spectrogram. In addition, the user has the freedom to examine waveform plots for the unknown at points indicated by the cursor, and to examine the intensity and fundamental frequency functions of selected segments of speech data. Other facilities include provisions for manipulating speech spectra and hearing the results through a channel vocoder. The system forms a general speech analysis-synthesis facility, only a few of whose capabilities were employed in the experiments described here.

ORGANIZATION OF THE RETRIEVAL PROGRAM

Each of the reference spectrograms consisted of a candidate word presented in the sentence frame "Please say again." These spectrograms made up a lexicon of 100 reference items of which 20 had both stressed and unstressed forms represented, giving a grand total of 120 entries. The items were stored on a disk in such a way that they could be selectively retrieved by means of a specially designed program that also collected data on each analyst's decisions and analysis procedures. A general model of this process is given in Figure 3.

Before commencing the experiment, the two analysts were each asked to select a personal set of up to 16 descriptive features that were considered to be useful in correctly selecting matching words from the lexicon. Each analyst then used his chosen features to label each member of the reference list. Any one of three discrete values could be assigned to each feature; either present, absent or unspecified.

The retrieval program listed the features that an analyst found in a word-segment of the unknown sentence and used this list (or feature vector)
Figure 1: An overall view of the DPP showing its principal components. The physical components actually used by the analysts were the HP display and cursor, the GT40 display and the keyboard.
Figure 2: A copy of the computer-displayed spectrogram of a portion of the unknown sentence (top) and the hypothesis obtained from the first analyst after pass 1 (bottom).
Figure 3: Model of the information pathways available to the analysts who formulated an hypothesized spectrogram from a sequence of reference-word units.
to extract a subset of the reference lexicon that shared the same features. An additional property of the program was that "unspecified" feature values, in reference words, matched both "present" and "absent" assignments of those features in the unknown segment.

While the matching program was under way, an analyst could specify additional feature information about the unknown by changing values or specifying previously unspecified values. Alternatively, he could relax feature assignments by increasing the number of unspecified features and thereby increase the size of the matching word list. The number of reference spectrograms that matched any specified feature-vector could be rapidly determined. In the event that too many reference items matched the specified features, the analyst was allowed to revalue features in the reference list to achieve greater precision. When the number of retrieved items fell to a sufficiently low level, the analyst could scan through them one by one, each time displaying the potential match above the unknown. In order to make a unique selection, he could then invoke additional information not included in the previous feature assignment; for example, expected formant shifts from the reference form to fit the apparent context of the unknown. If none of the retrieved items matched sufficiently well, the feature assignment was then modified to select a new list of matching words.

The analyst could also display a series of potential word matches in an appropriate order, side by side, and judge whether coarticulation effects could account for the remaining discrepancies between the reference words and the unknown. After the analyst had arrived at a hypothesized reference-word sequence that satisfied his criteria, the sequence of items was given to the original speaker to be spoken in the same tone of voice and with the same intonation pattern used in the original unknown sentence. This production form of the matched sequence was then added to the data base for the analysts' examination. At this point, new reference words could be substituted where the analyst noted that a mismatch with the unknown sentence had occurred.

The record-keeping section of the retrieval program noted the number of searches of the reference library that were made by both the analysts and all of the reference words that they examined. This record allowed the authors to trace the significant information feedback paths in the system--those that resulted in new searches of the reference library with differing feature-vectors. These feedback paths are noted in Figure 3. The extent to which lexical information can modify an analyst's segmentation and feature assignment was not surprising. In fact, through this attempt to model explicitly the information flow among the various subtasks of the analysis process, we have uncovered a structure similar to the model for speech recognition proposed by Fant (1970) nearly 7 years ago.

**EXPERIMENTAL OBSERVATIONS**

Both analysts found the 16 assignable features to be insufficient and would have used a larger number, had there been provision to do so. However, even the assignment of sixteen features to 120 reference items was very time-consuming. In order not to impose any prior feature organization on our analysts, all features were considered equally important in establishing a
match. The analysts were frustrated by the necessity to explicitly mark the absence of many features—a requirement imposed by the single-level feature organization. Use of a multilevel or hierarchic feature organization necessitating the selection of secondary features only if they were appropriate in the light of specific assignments of the primary features, would have overcome this difficulty.

Both analysts found little difficulty establishing reference-word matches to the prominent words of the unknown sequence. In fact, they were surprised to discover how little information (possibly only 3 or 4 features) sufficed for the retrieval of no more than 6 matching items. More severe difficulties were encountered in attempting to find the matches for the less prominent words or syllables. Here the analysts did not trust their feature assignments—an indication of the difficulty that they encountered in making those assignments in the first place. One analyst resorted to an exhaustive scanning of the list of unstressed reference items. The other compared pairs of stressed and unstressed reference items to infer which features could be expected to be harder to detect under reduced stress. He then relaxed the feature assignment for the corresponding unstressed items.

The second analyst attempted to overcome the word segmentation problem by selecting prominent syllables around which to organize a retrieval attempt. The ability to look at variations in the spectrum envelope as the cursor swept through successive time intervals of the spectrograms proved to be quite helpful in selecting the most prominent syllable of a sequence. Organizing the retrieval strategy around prominent syllables permitted the rapid examination of alternative hypotheses. For example, the first hypothesis might be a monosyllabic stressed word, the second a bisyllabic word with an additional unstressed syllable. Information about additional consonantal segments could be added to the feature vector used for retrieval until the number of retrieved items was small enough to be individually scanned. Even though only a few salient features located near the prominent vowel were assigned, the retrieval process frequently resulted in an obvious match to a much longer segment of the unknown.

The features describing vowel color were not found very useful by either analyst. There are two reasons that may account for this finding. First, contextual influences on the vowel formant-frequencies of both the reference word and the unknown word-segment made reliable feature assignment difficult. Second, very few of the reference items differed by vowel color alone. Thus, the specification of vowel color features did not significantly reduce the number of retrieved matches in contrast to leaving them unspecified.

The one analyst who attempted to make use of segment duration in his feature assignment found it to be useful only in extreme cases. For the most part, the segmental durations of unknown words varied considerably as a function of stress, syntactic role and position in the sentence, making small durational differences ineffective for discrimination purposes.
Results

The average proportion of words that the two analysts succeeded in correctly matching was only 26 percent, and this figure did not increase after one cycle of feedback. Although one error was corrected, an additional error was introduced in the words hypothesized on the second attempt. The overall word-matching performance was thus significantly lower for the machine-assisted word-matching experiment than for the similar experiment conducted with conventional spectrograms by Ingemann and Mermelstein (1975). There are several possible reasons for this deterioration in matching performance. The relative unfamiliarity of the display—in particular the way acoustic features seen on the DPP are affected by the limited time resolution of the display—may have been one factor. More importantly, perhaps, the sentence in the current experiment was longer (21 words vs. 16 words) and somewhat more complex. The lexicon used in the DPP experiment intentionally included more words that had close phonetic similarities to the unknown words of the sentence.

The word-identification scores are broken down by analyst, stress, and number of syllables in Table 1. While 52 percent of the words that contained at least one stressed syllable were correctly identified overall, practically all of the matches with unstressed words were incorrect. Overall performance on multisyllabic words was somewhat higher than on monosyllabic words. Here the relative performance of the subjects differed significantly. The analyst who used the strategy that focused on prominent syllables did better on monosyllabic words but worse on multisyllabic words. The strategy led to frequent errors on the unstressed syllable of a multisyllabic word—particularly when phonetically similar words were included in the lexicon. Substitutions in the unstressed syllables of those words were quite frequent. Examples of such substitutions are "immunity" for "community", "human" for "humor", "arrive" for "derived", and "salt" for "assault."

<table>
<thead>
<tr>
<th>TABLE 1: Percent correctly identified words.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass 1</td>
</tr>
<tr>
<td>Tokens</td>
</tr>
<tr>
<td>Monosyllabic words</td>
</tr>
<tr>
<td>Multisyllabic words</td>
</tr>
<tr>
<td>Unstressed words</td>
</tr>
<tr>
<td>Stressed words</td>
</tr>
<tr>
<td>All words</td>
</tr>
</tbody>
</table>

Percent correctly identified syllables

| All syllables | 30 | 43 | 47 | 43 | 47 |

94
CONCLUSIONS

The single most important observation to emerge from the results is the poor performance of the analysts on unstressed words. A reference token, whether spoken in a stressed form or in a different unstressed environment, does not provide sufficient information to enable the analyst to effect a match. Perhaps a larger number of reference tokens taken from a variety of contexts in which the word may occur might be useful, since it is evident that analysts are usually unable to predict the transformations that the acoustic features of words can undergo if they are uttered in phonetically different contexts. Analysts generally judge similarity in terms of common features between the unknown and reference tokens. They do not pay particular attention to the variability of those features and thus do not differentiate among the features according to their reliability in establishing matches. It seems likely that intensive learning sessions on the variability of acoustic features are required before improved word matching results can be obtained.

The lack of any significant improvement following feedback of the hypothesized words spoken as a sentence is probably due to the fact that the overall performance was initially too low (that is, the initial hypothesis was offered with such a low level of confidence that it contributed as much to the analyst's uncertainty as it did to his knowledge). It appears to be that a higher minimum performance must be reached before the information supplied by feedback can be usefully absorbed. If an unknown word is embedded in the correct context, its appearance is likely to be quite similar to its form in the unknown sentence. However, if the context is incorrect as well, a new production of the reference form is obtained that may not be any more similar to the unknown than it was to the original.

Let us now consider the prospects for implementing an entire feature assignment and word-matching procedure in algorithmic form for execution by a machine. The selection of matching words on the basis of assigned feature values is clearly the easiest procedure to implement, and, in fact, this has already been successfully carried out. Heuristics are available for the assignment of values to most acoustic features and, therefore, we can expect that this analysis procedure can be implemented at a cost that increases roughly linearly with the number of features used. We anticipate more difficulty, however, with the process labeled "similarity". We are not, as yet, able to quantify a general similarity metric that assigns perceptually appropriate weights to specific differences. Events of short duration, such as bursts, may contribute a great deal to measures of similarity, whereas differences in events of longer duration, such as shifts in formant frequencies in vocalic intervals, may be of less significance.

It is possible that the comparison of word-sequences might be implemented with the aid of a speech synthesis program; however, it appears that finding an appropriate metric of similarity is the most difficult problem. Given any general difference measure, we do not yet know how to separate differences between speakers from differences between words, and until we can learn what the important distinctions are that we must look for, word identification through spectrum matching by a human analyst, or by a machine, will not be a practical art.
REFERENCES

