Speech recognition through spectrogram matching

Frances Ingemann* and Paul Mermelstein

Haskins Laboratories, New Haven, Connecticut 06510
(Received 19 July 1974; revised 7 October 1974)

In order to assess human analysis of acoustic data before attempting such analysis by machine, a series of experiments was conducted in which subjects were asked to match spectrograms of continuous speech to reference spectrograms of the same words. Although error rates varied with sentence difficulty and size of vocabulary, comparison of the matches shows greater agreement in phoneme segments than other experimenters have obtained in phonetic transcriptions of unknown utterances without semantic of syntactic processing. Accuracy in matching can be further improved by feedback to the human analyst in the form of spectrographic representation of a sequence of tentative matches spoken as if they made up the unknown utterance. The results suggest that automatic matching of word- or syllable-sized acoustic patterns may provide a more accurate phonemic input to the syntactic–semantic component of a speech recognition system than direct transcription of individual phonetic segments.

Subject Classification: 70.65, 70.60, 70.30.

The limited performance of speech recognition systems to date indicates to us that improved acoustic analysis as well as good syntactic–semantic analysis are prerequisites to better system performance. Human analysis of acoustic data without the use of nonacoustic information can be expected to assist the design of improved acoustic analysis systems.

The difficulty which people have in accurately identifying the phonetic content of spectrograms of unknown utterances has long been recognized by researchers in the field.[1] Until recently, little experimentation had been undertaken since the early pioneering work at Bell Laboratories.[2] Within the past few years interest in spectrogram reading has been renewed, at least partially in response to attempts at automatic speech recognition in the expectation that cues available to human spectrogram readers could be programmed into an automatic speech recognition system.

Studies by Klatt and Stevens[3] and Lindblom and Svensson[4] have shown that subjects who are experienced in examining spectrograms can label phonetic segments correctly less than half the time when they are presented with spectrograms about which they have no additional information. These experiments have also shown that the addition of syntactic, semantic, and prosodic information can improve performance significantly.

Our interest lay in finding out whether speech recognition could be improved without recourse to nonacoustic information. The technique we chose was the matching of spectrograms of unknown utterances with reference spectrograms identified only by number so that success of the task depended almost entirely on the ability to match patterns visually. Klatt and Stevens[3] also used spectrographic matching but because the reference words were known, syntactic and semantic considerations entered into the selection of suitable matches. Our experiments were undertaken to evaluate human spectrogram matching performance before attempting spectrogram matching by machine.[5]

I. EXPERIMENT I

The first experiment was in the nature of a limited pilot study to determine whether subjects could match spectrograms at all. In this experiment, as in all the experiments described in this paper, spectrograms were based on the speech of a single female speaker. Wide-band spectrograms were produced on a Voiceprint spectrograph using a frequency scale of 0–4800 Hz.

Subjects were asked to locate, within spectrograms of five test sentences, 10 words given in reference spectrograms. The reference words were content words consisting of one, two, or four syllables spoken in the context "Say ——again." Each reference word occurred once in the sentences, except that two monosyllabic words occurred in a suffixed form as well as in the uninflected form given as the reference word. Subjects were not told the meaning of either the reference words or the test sentences, but they were given the meaning of the sentence frame in which the reference word appeared.

Three subjects were used: one who had extensive experience examining spectrograms, one who had moderate experience, and one who had no experience. All three subjects performed the task with few errors (75%–83% correct).

II. EXPERIMENT II

Since Experiment I had shown that spectrograms could be matched, a second experiment was devised to include all words in a randomly selected text to determine whether the task could be done as successfully with a larger set of reference words, some of which were unstressed. A passage, four sentences long, was chosen at random from a publication. These sentences contained a total of 70 words, of which 51 were different.

Reference spectrograms were made of the 51 words in the context "Say——again," in which *again* was given major sentence stress to prevent both contrastive stress on the reference word and a possible phrase boundary juncture between the reference word and *again*. In addition, a second version of some monosyllabic function

words in an unstressed context was provided. The second context was not identified for the subjects, who were told only that the word was in unstressed position in the second version.

Six subjects took part in the experiment, all of whom had experience examining spectrograms. Each subject was given only one or two test sentences so that the reference set contained approximately twice as many words as a subject would find in his sentence.

The overall score of correct identifications was 67%. Most errors were made on monosyllabic words, particularly function words.

III. EXPERIMENT III

Because monosyllabic words seemed to be more difficult to match than polysyllabic words, a third experiment consisting only of monosyllabic words was designed to examine this area more carefully. The difficulty of the task was increased by adding to the reference words other words which were phonetically similar.

A test sentence consisting of 10 monosyllables was constructed and the reference set consisted of 40 words: the 10 words in the test sentence and 30 words similar to phonetic strings in the test sentence. Once again, for some of the monosyllabic function words a second unstressed variant in a context not known to the subject was provided.

Four subjects took part in the experiment, all of whom had experience examining spectrograms. Words were identified correctly only 48% of the time. In contrast to the previous experiment, content words were not more easily matched than function words. Content words were matched correctly only 45% of the time, while function words were matched 50% of the time.

This experiment also pointed up the difficulty of locating word boundaries when the reference set includes words which can be confused. For example, *ask* was identified twice as *last* and twice as *lass* because the *l* of the preceding word *will* was assumed to be part of this word; *to* was once identified as *took* when it preceded *pay*.

A comparison of the string of phonemes in the sentence with the string of phonemes in the matched reference words show that the percent of phonemes correctly matched is considerably higher than the percent of words. Seventy-two percent of the phonemes in the sentence were found to be correctly matched and 35% errors made. The total of these two percentages exceeds 100 because two phonemes in the reference words were sometimes matched to a single phoneme in the sentence.

When considered from the point of view of word recognition relative to phoneme recognition, the results correspond rather closely to the relationship found by Fletcher[6] between syllable recognition and "letter" recognition in testing noisy speech transmission systems. Fletcher's curves would predict 77% "letter" [phoneme] recognition to accompany 48% syllable recognition. The predicted sentence intelligibility for human listeners under these conditions is 94%. These facts suggest that an automatic speech recognition system whose performance on acoustic analysis is comparable to the visual performance of our human subjects and whose performance on the syntactic–semantic level is comparable to that of human listeners could be expected to "understand" 94% of simple questions or instructions such as given in Fletcher's Intelligibility List.

IV. EXPERIMENT IV

The number of errors in preceding experiments led us to try spectrographic feedback as a means of improving performance. Experiment IV began, as did the previous two experiments, by asking the subject to match words in a sentence to reference words. Two subjects were used, one who had participated in all three of the previous experiments and one who had participated in none.

After the subject had tentatively matched the sentence, the reference words he selected were read as a sentence, with stress and intonation as close as possible to the original utterance. Spectrograms of this sequence of tentative matches were given to the subject to compare with the original sentence. He was then allowed to revise his list of matches and once again he was given spectrograms of the sequence of matches. This process was repeated until the subject indicated that he no longer wished to continue. Both subjects stopped with their third attempt. The subject was then asked to give a confidence rating for each of his matches.

The subjects differed greatly in their matching ability, although spectrographic feedback improved both of their performances. Whereas one subject on the third try correctly matched all the words, the other only attained 38% correct. Furthermore, only 50% of the matches in which the second subject expressed high confidence were in fact correct. However, the ratings did have some validity in that none of the low-confidence matches were correct.

There are a number of possible explanations for the difference between the performances of the two subjects, since test conditions differed slightly. The more successful subject had fewer reference words (78 vs 120) and spent twice as much time making matches. He also asked for and received spectrograms of a second reading of the original sentence.

Although the second subject's word recognition score was only 38% on this experiment, a comparison of phonemes between the words he matched and the words in the original sentence gives a score of 82% correct and 24% error. When we compare this result with Experiment III, we see that although the second subject made more word matching errors than the average for Experiment III, he matched more phonemes.

V. CONCLUSIONS

Human subjects can match spectrograms of unknown utterances with reference-word spectrograms better than they can directly transcribe the spectrograms in terms of a sequence of phonetic elements. Our results indicate that the subjects do not make full use of the

acoustic information present in the spectrograms unless they are given a means to assess the significant differences between the spectrographic manifestations of different words. One such means is comparison of spectrograms.

Even when the number of words correctly matched is low, the number of phonemes matched is higher than the number of correct identifications reported for other acoustic phoneme-recognition schemes. This suggests that the output of this analysis-by-matching technique could yield a more accurate input to the syntactic–semantic component of a speech recognition system than is now available.

A process that generates the sequence of matched words in a manner resembling as closely as possible that of the unknown utterance serves as a useful source of feedback to the subjects. However, since only two subjects took part in the experiment with feedback, the degree of improvement which might generally be obtained cannot be predicted.

Subjects' error rates in matching vary significantly with sentence difficulty, size of vocabulary, and general ability to predict the changes a word pattern may undergo when placed in an unknown context and spoken with a different prosody. At least for limited vocabularies, subjects are able to determine whether two spectrographic patterns do or do not correspond to different productions of the same word more reliably than they are able to assign phonetic labels to the speech stream as seen in the spectrogram. However, as the size of the vocabulary increases, the likelihood of selecting the correct word decreases. Since the analysis time and paper-handling difficulties also increase with vocabulary, manual execution of such tasks can rapidly become impractical. Subjects' ability to generalize the acoustic cues they observe so that they need not be presented with all spectrographic forms but only a limited subset remains to be investigated.

Automation of this matching process would entail storage of a complete vocabulary in spectrographic form, an exorbitant requirement. Various parametric representations suggest themselves but the corresponding storage savings will have to be weighed against deterioration of performance as compared with full spectrogram matching. Although generalization of the appropriate acoustic information into a sufficient set of analysis rules remains the ultimate goal, studies of word matching provide useful comparisons for the performance of any other method.

We believe the results of our experiments warrant continuation of our studies with computer-assisted word retrieval as a means of developing automatic pattern matching techniques that make best use of those cues found by humans to be useful in establishing reliable word matches.

ACKNOWLEDGMENTS

We wish to acknowledge the contribution of F. S. Cooper and P. W. Nye to the initiation of these experiments. We are also grateful to the following people who not only served as subjects but also provided useful insights into the spectrogram matching process and suggestions for further experimentation: L. Barton, R. Collier, F. S. Cooper, J. Gaitenby, L. Lisker, R. McGuire, P. W. Nye, and G. Sholes. This research was supported in part by the Advanced Research Projects Agency of the Department of Defense under Contract No. N0014-67-A-029-002 monitored by the Office of Naval Research. The views presented here do not necessarily represent the views of the Department of Defense.

*Present address: Linguistic Department, University of Kansas, Lawrence, Kansas 66045.

[1]See, for example, A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Why are Speech Spectrograms Hard to Read?" Am. Ann. Deaf 113, 127–133 (1968). See also C. G. M. Fant, "Descriptive Analysis of the Acoustic Aspects of Speech," Logos 5.4 fn (1962).

[2]R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech* (Van Nostrand, New York, 1947).

[3]D. H. Klatt and K. N. Stevens, "On the Automatic Recognition of Continuous Speech: Implications from a Spectrogram-Reading Experiment," IEEE Trans. Audio Electroacoust. AU-21, 210–217 (1973).

[4]B. E. F. Lindblom and S-G. Svensson, "Interaction Between Segmental and Nonsegmental Factors in Speech Recognition," IEEE Trans. Audio Electroacoust. AU-21, 536–545 (1973); S-G. Svensson, "Prosody and Grammer in Speech Perception," Monograph 2, Institute of Linguistics, University of Stockholm (1974).

[5]A more detailed description of these experiments will appear in Haskins Laboratories Status Report on Speech Research 39/40 (1974).

[6]H. Fletcher, *Speech and Hearing* (Van Nostrand, New York, 1929).