On the Automatic Recognition of Continuous Speech: Implications from a Spectrogram-Reading Experiment

DENNIS H. KLATT and KENNETH N. STEVENS

Abstract—An experiment was performed in which the authors attempted to recognize a set of unknown sentences by visual examination of spectrograms and machine-aided lexical searching. Ninteen sentences representing data from five talkers were analyzed. An initial partial transcription in terms of phonetic features was performed. The transcription contained many errors and omissions: 10 percent of the segments were omitted, 17 percent were incorrectly transcribed, and an additional 40 percent were transcribed only partially in terms of phonetic features. The transcription was used by the experimenters to initiate computerized scans of a 200-word lexicon. A majority of the search responses did not contain the correct word. However, following extended interactions with the computer, a word-recognition rate of 96 percent was achieved by each investigator for the sentence material. Implications for automatic speech recognition are discussed. In particular, the differences between the phonetic characteristics of isolated words and of the same words when they appear in sentences are emphasized.

The aim of the study reported here is to gain insight into the strategies that might be necessary in a device for the automatic recognition of spoken sentences. An experiment was performed in which the authors attempted to recognize a set of unknown sentences by visual examination of spectrograms.

The use of human interpretation of broad-band spectrograms bypasses the problem of automatically extracting relevant acoustic parameters from the signal such as formant frequencies, fundamental frequency, and rapid spectral changes. Theoretical considerations [1] and experience with spectrographic analyses of speech and with the Haskins Pattern Playback [2] has shown that a broad-band spectrogram preserves most of the perceptually important information in the acoustic waveform. Thus, the results of this study should not be strongly dependent on limitations imposed by the type of front-end preprocessing scheme employed.

Nineteen sentences were selected randomly from a larger corpus of sentences produced at a conversational speaking rate by five different adult male speakers. The corpus consisted of questions that were asked of a computer program whose data base concerned the chemical analyses of moon rocks [3]. Broad-band spectrograms were made of the utterances, and this set of 19 spectrograms was presented to the experimenters (10 sentences to DKH and 9 to KNS). Fig. 1 displays one of the spectrograms used in the study.

Working independently the authors attempted to recognize the sentences. The sentence-recognition task was divided into two parts consisting of 1) an attempt to derive a phonetic transcription of an unknown sentence without identifying any of the words, and 2) use of this transcription and a 200-word computerized lexicon in order to guess at word candidates and finally derive a meaningful sentence that satisfied the syntactic constraints of English. The spectrogram continued to be used during the latter phase to verify working hypotheses and to generate a modified phonetic transcription if a portion of the initial transcription did not lead to any reasonable lexical hypotheses.

Phonetic Transcription Task

The authors first attempted to make broad phonetic transcriptions of the utterances from the spectrographic data. In order to minimize the possibility that words might be identified and used to improve the transcription during this part of the task, a 300-ms window was cut in a piece of cardboard that was overlaid on the spectrogram. The window was moved from left to right across the spectrogram and the transcription was done on a single pass.

The phonetic transcription was made in terms of sequences of specific phonetic segments. If some features of a segment were not visible or identifiable with sufficient confidence, only a partial feature specification was made of a segment.

An example of a phonetic transcription appears below the spectrogram shown in Fig. 1. A partial transcription is indicated in the figure either by identifying the features present or by listing a set of candidate segment types. In some instances, the number of phonetic segments underlying a stretch of the speech signal was not specified exactly. For example, a portion of the signal could be characterized by "one or two sonorants."

Manuscript received April 30, 1972. This research was supported by the Information Processing Techniques Branch of the Advanced Research Projects Agency. This work was performed at Bolt Beranek and Newman Inc., Cambridge, Mass.

The authors are with the Research Laboratory of Electronics and the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

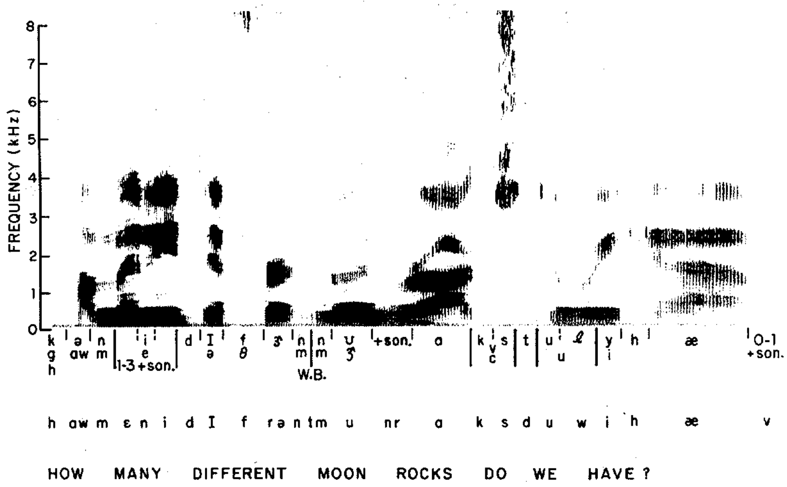

Fig. 1. **Broad-band spectrogram of one of the unknown sentences, "How many different moon rocks do we have?", is shown. Phonetic transcription of DHK is indicated immediately under the spectrogram. This partial transcription can be compared with the correct transcription that appears below it.**

A so-called "correct" phonetic transcription appears below the transcription of the experimenter in Fig. 1. An arbitrary decision was made to define the correct transcription in terms of the phonetic segment representation for each word as stored in the computerized lexicon. There are a number of phonological rules of English that modify the pronunciation of words when these words appear in different sentence environments. A more realistic correct transcription would take into account the effects of these phonological rules. The result might be a slightly closer correspondence between our transcriptions and the correct ones.

A summary of the overall phonetic transcription results is presented in Table I. Over 650 phonetic segments were transcribed. The phonetic transcription contains many errors and omissions: 10 percent of the segments were omitted from the transcriptions, 17 percent were incorrectly transcribed, and an additional 40 percent were transcribed only partially in terms of phonetic features. A segment was improperly inserted in the transcription when in fact none was present less than one percent of the time.

Stated in positive terms, 33 percent of the segments were transcribed correctly. If one adds those segments correctly transcribed in terms of partial phonetic feature information (40 percent), a total of 73 percent of the transcription contained no errors or omissions.

Comparison of the results for the two experimenters shows that DHK tended to make more complete transcriptions than KNS. That is, if he had doubt concerning the identity of a feature, KNS would leave the feature unspecified, whereas DHK would more often try to identify all features of the segment. Some feeling for the relative merits of these two strategies can be obtained by computing the conditional probability of an error given that a partial or a complete transcription was attempted. These probabilities are presented in Table II for the two experimenters. It is seen that KNS is able to make fewer mistakes by avoiding complete transcriptions when he is not sure. The cost attached to this strategy is that he is presenting substantially less information to the next stage of the recognition process.

TABLE I
Transcription Error Rates (in Percent) for All Segments

	DHK	KNS	Average
Total number of segments in corpus	359	299	
Correctly transcribed	41	24	33
Partially transcribed with no errors	31	50	40
Segments with errors	19	15	17
Omitted	9	11	10

TABLE II
Conditional Probability of a Transcription Error

	DHK	KNS
Complete transcription attempted	0.22	0.13
Partial transcription attempted	0.19	0.19

Vowel Errors

The transcription results for vowels and consonants were analyzed separately. Some characteristics of the transcription results for vowels are presented in Table III. The distribution of vowel errors is similar to the overall data shown in Table I. However, only three percent of the vowels went undetected. Thus, the syllabic structure of the utterance was usually identified from the spectrographic representation. Omitted

TABLE III
Transcription Error Rates for Vowels

	Number	Percent Correct Complete	Percent Correct Partial	Percent Error	Percent Omission	Percent Error + Omission
All vowels	256	33	44	20	3	23
+ Stressed	140	25	50	22	3	25
— Stressed + Reduced	82	43	40	10	7	17
— Stressed + Reduced	34	38	24	38	0	38

TABLE IV
Transcription Error Rates for Consonants

	Number	Percent Correct Complete	Percent Correct Partial	Percent Error	Percent Omission	Percent Error + Omission
All consonants	402	32	39	6	13	29
Prestressed	142	47	38	11	4	15
Nonprestressed	168	29	40	20	11	31
In a prestressed cluster	23	23	32	36	9	45
In a nonprestressed cluster	69	26	37	9	28	37

vowels tended to be preceded by a stressed vowel or sonorant.

In a Chomsky-Halle generative phonology of English [4] vowels are subdivided into stressed and nonstressed categories. Unstressed vowels are usually shorter in duration. The unstressed vowels are further subdivided into reduced vowels (e.g., "*a*bout") and unreduced vowels (e.g., "bab*y*", "*a*nd"). Transcription errors have been broken down into these three vowel categories in the next three lines of Table III. The data indicate that the reduced vowels produced somewhat fewer transcription errors and the most completely correct responses. Unstressed unreduced vowels were most difficult to recognize.

When transcription errors are made for vowels, it is of interest to study the particular confusions involved. Since incorrect responses are usually similar to correct responses, an approximate technique for studying segmental confusions is to identify the phonetic feature values that are in error [5]. Analyses of the data from one experimenter (DHK) suggest that errors were approximately randomly distributed among the chosen features, although there was a surprisingly high number of front-back confusions. The division between front and back vowels is quite sharp in the Peterson and Barney [6] formant data, but this apparent natural distinction is not as clearcut in coarticulated normal utterances.

Consonant Errors

The transcription error rates for consonants are shown in Table IV. Single consonants have been divided into 1) prestressed and 2) nonprestressed categories in the table. The figures show that about twice as many errors were made in transcribing nonprestressed consonants.

Consonant clusters are known to involve special coarticulation effects in English. Clusters have been analyzed separately in Table IV. A cluster was defined in a somewhat unorthodox manner: a cluster consisted of a sequence of two or more consonants not separated by a word boundary[1]. Prestressed clusters were rare in the data sample. The nonprestressed clusters contained a large number of segment omission errors. Most of these errors involved omitting either a plosive or a nasal consonant. The majority of the clusters producing omissions were (homorganic) nasal-plosive sequences. These particular transcription omissions may not be errors but rather the result of a phonological rule of English pronunciation.

Errors seem to be more or less randomly distributed among the manner and place-of-articulation features for consonants. A number of voicing errors for stops were due to a rule of English phonology that shortens the aspiration duration in a voiceless stop in preunstressed and word-final positions.

The data from individual vowels and consonants were scanned to determine whether certain segments were easier to identify than others. The stressed vowels [i, a, u], the schwa vowel, single prestressed voiceless consonants, and single nasal consonants (prestressed or not) had particularly low error rates. Some caution should be used in interpreting this observation because the sample size may have been too small to

[1] The cluster definition was motivated by the observation that a word boundary blocks certain kinds of cluster rules. For example, in an [st] sequence, the [t] is normally unaspirated, but the [t] is fully aspirated if preceded by a word boundary.

determine significant differences in error rates for individual segment types.

Word Boundaries

Cues to the location of word boundaries in a sentence are present to some extent in the acoustical signal, but these cues are not sufficient to transcribe the probable location of each word boundary. A word boundary precedes the glottal stop in English, and this cue was used by DKH to predict ten word boundaries in the corpus without error.

Word-boundary hypotheses must be made in order to specify lexical scans efficiently. The authors attempted to use durational cues, consonant-cluster sequence limitations, and certain kinds of allophonic rules as well as semantic information to generate probable word-boundary locations.

Sentence-Identification Results

In the second part of the experiment, the spectrographic data and the phonetic transcription were used in conjunction with a computerized lexicon developed by Makhoul [7] in order to identify the words of the sentences. The overt strategy of the experimenters was to begin at the left of the spectrogram and select several segments from the phonetic transcription. The segment specification was then typed into a computer console by the investigator, and all of the lexical items that matched the specification were printed out automatically in response. An example of a question is "Give me all single words that begin with a [p] or a [t], followed by a nonlow front vowel, followed by zero or one sonorants, followed by an [s]." It was possible also to ask for single words or two-word sequences that contained a segment specification when the segments did not begin the word or word pair.

The word candidates were then observed and either accepted or rejected on the basis of a second detailed look at the acoustic data and the experimenter's knowledge of the syntactic and semantic constraints of English. If none of the word candidates proposed by the computer were acceptable or if no candidates were proposed by the computer, the experimenter attempted to rephrase his question or change some feature value in the transcription on the basis of reexamination of the acoustic data. At other times, a word was identified directly from the spectrogram without the aid of the computerized lexicon. If the experimenter was unable to recognize a given section of the spectrogram, he would jump ahead and return to that section later when he knew more about the rest of the sentence.

The 19 sentences consisted of 158 words. Each investigator misidentified three words in his corpus, but in the few sentences common to both the corpuses, no word was misidentified by both investigators. The following analysis describes statistical characteristics of the computer responses to word-search questions by DHK. These results help to reveal the recognition strategy that was employed.

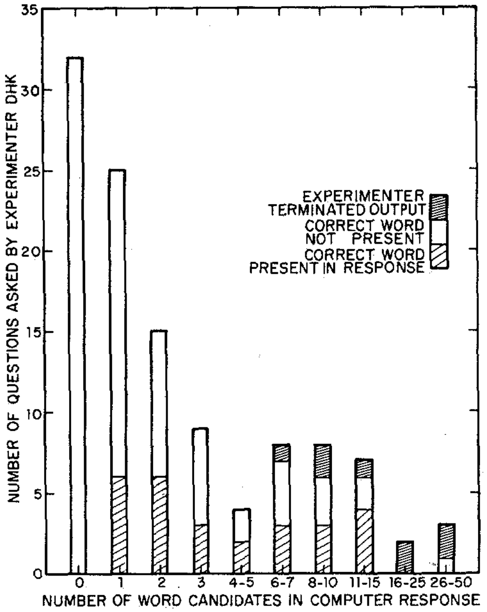

Fig. 2. Distribution of the number of word candidates in the computer responses to 114 lexical search questions posed by DHK.

About 114 word-search questions were asked of the computer program during the course of the recognition efforts. All were requests to find a single word having a feature matrix consistent with a given partial feature specification. The computer responded with the number of words that matched the segmental specification. If this number was greater than five, the experimenter had the option of terminating the output before the words were printed and could rephrase his question if he so desired. A total of eight questions were terminated in this way.

For the remaining 106 questions, zero or more words were printed on the teletype and the experimenter attempted to judge whether any of the printed words matched the spectrographic data sufficiently to be accepted. The correct word appeared in the output only 25 percent of the time.

The number of words in the response to a question varied greatly. A distribution of the number of words printed for each question is shown in Fig. 2. If the correct word was present, the median number of words in the response was 4.5. Thus, the investigator usually had to select the word from several alternatives, and the correct word was among the printed alternatives only one quarter of the time. Although this presented many difficulties for the investigators, it was observed that reference back to the spectrogram permitted fairly reliable acceptance and rejection of word candidate hypotheses. Subsequent analysis revealed that when the correct word was present in the computer response, the word was recognized and accepted in every case.

The words from the ten sentences were sorted into eight syntactic categories. Nouns formed the dominant class, with verbs, adjectives, pronouns, prepositions, determiners, verbal auxiliaries, and conjunctions represented about equally. The 27 words that were sought and found with the aid of the computerized lexicon did not have the same distribution. Most of the function words and familiar single-syllable nouns such as the digits were recognized without the aid of the lexicon.

Discussion

Previous attempts to recognize words from spectrographic examination have indicated that this is a very difficult endeavor. Potter, Kopp, and Green [8] found that after an extended period of training, deaf subjects were able to learn to recognize no more than about four words per hour of training with spectrographic displays. Word acquisition did not accelerate with time. These data suggest that subjects were treating a word pattern as a whole and not attempting to break it up into phonetic units. Our own informal experience and that of our students with spectrogram-reading tasks has shown that, given sufficient time, it is possible to identify, essentially without error, isolated words drawn from a closed vocabulary of 50-100 words spoken by a variety of talkers. In the isolated word case, it is generally agreed that the difficulties that prevent subjects from learning large vocabularies are largely due to the context-dependent encoding of phonetic information in the acoustical display [9].

The final identification score for the words in the sentences of this study was 96 percent. This score is comparable to that reported by Lindblom and Svensson [10] for a task involving nine sentences, although their score of 97 percent was based on only a small number of samples (only one error in 34 lexical items for one subject) in a task that differed from ours in that certain prosodic information was provided to the observer in their experiment.

Isolated Words Versus Sentences

Sentence recognition differs from isolated word identification in several important ways that make the sentence task considerably more difficult. Our experience with spectrograms of sentences indicates that: 1) word boundaries are not clearly marked and significant coarticulation occurs between adjacent words; 2) segmental durations are, on the average, about half of durations observed in spoken isolated words so that a greater degree of coarticulation and articulatory target undershoot is observed; 3) linguistic stress and syntactic bracketing interact through a system of phonological rules to modify segment durations, superimpose an intonation contour, and modify segmental features, generally by reducing vowels and obscuring the acoustic distinctiveness of unstressed consonants and consonant clusters; and 4) certain segments are deleted or radically changed at word boundaries according to other phonological rules that operate at the level of the phrase.

Examples of these four effects are illustrated in Figs. 1 and 3. The latter figure contains a spectrogram of the sentence "If the cube is not blue, pick it up." Also shown are spectrograms of the same words spoken in isolation in random order by the same talker. Comparisons between the words in isolation and in the sentence reveal the extent to which a recoding of the spoken message has taken place.

1) The words in the sentence shown in Fig. 3 contain only one syllable so that it is not obvious from this example that word boundaries and syllable boundaries are largely indistinguishable. However, an indication of this fact can be seen by comparing the first syllable boundary in the word "different" (Fig. 1) with the word boundary in the sequence "if the" (Fig. 3). In addition, it is frequently unclear whether to assign a given consonant to the previous syllable or the following syllable. Thus, in Fig. 3, there is no acoustic cue to suggest that [b] of "blue" is syllable initial, but [b] of "cube" is syllable final.

2) Segmental durations are shorter in the sentence context than in isolated words in Fig. 3. Stressed vowels are not shortened as much as other segments and a sentence-final lengthening rule of English prolongs the duration of the word "up." The word "the," being unstressed, is shortened to such an extent that the vowel formants are considerably displaced from their positions when spoken in isolation. Coarticulation with adjacent segments causes the second formant to be in motion throughout the vowel and to be about 400 Hz higher.

3) The orthographic comma at the phrase boundary in the sentence of Fig. 3 is realized acoustically by a slight increase in the silent closure interval following the word "blue" and by a prolongation of the vowel in the word "blue" over the duration it would assume had no syntactic break of this type occurred [11]. The fundamental-frequency contour (not shown) also displays certain characteristics that help to signal phrase boundaries [12] and words carrying primary stress. An optional phonological rule involving spread of nasalization to segments adjacent to a nasal consonant has apparently caused the vowel of "not" to be partially nasalized, whereas it was not nasalized when spoken in isolation.

4) Several examples of phonological rules that induce changes in segments at word boundaries can be seen in Figs. 1 and 3. The release of a word-final [k] into a following vowel is not aspirated, as shown in the word sequence "pick it" (Fig. 3). The [t] in "not blue" (Fig. 3) is not released at all due to the presence

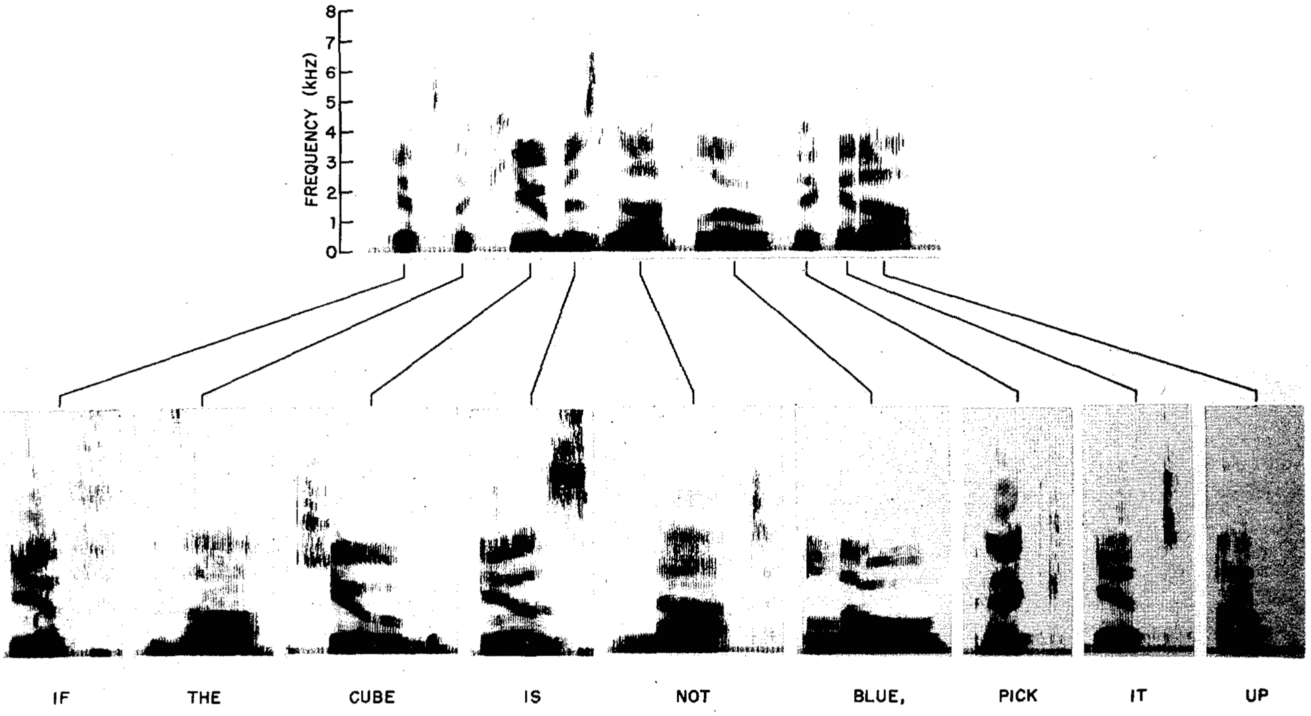

Fig. 3. Broad-band spectrogram of the sentence "If the cube is not blue, pick it up" is shown in the top half of the figure. Spectrograms of the same words spoken in isolation appear below it.

of a following stop. The dental stop in "it up" (Fig. 3) is realized as a very brief tongue flap in this unstressed environment. The consonant cluster [ntm] of "different moon" (Fig. 1) is produced by substituting a glottal stop for the [t] and by realizing the glottal stop as a brief precipitous fall in fundamental frequency in the middle of the nasal sequence.

All of these factors mitigate against the use of a set of stored acoustic patterns for words in an automatic sentence-recognition device. Stored acoustic patterns cannot be expected to match the input sentence (except perhaps for the case of multisyllabic stressed words wherein at least the central portion of the word is somewhat immune to the effects noted above). In general, it is clear from these results that successful automatic recognition or understanding of spoken sentences will necessarily involve a stage at which phonetic segments or syllables are at least partially recognized, and phonological decoding rules are applied to infer the possible missing and modified segments as well as to guess at word boundaries and even map out tentative syntactic boundaries.

Overcoming Transcription Errors

Results for a phonetic transcription task comparable to ours, based on visual observation of spectrograms, have been reported by Lindblom and Svensson [10]. In this experiment, subjects were presented with nine spectrograms of "pseudosentences" that had the prosodic patterns and phonological patterns of normal meaningful (Swedish) sentences, but in fact were meaningless utterances. Thus, the observers were unable to make use of lexical information or of syntactic or semantic context in making their transcriptions. (These effects were eliminated in our experiments by using a 300-ms observation window). Apparently, the subjects were not permitted to give partial specification of phonetic segments in the Lindblom-Svensson experiments. The results showed that 48 percent of the segments were correctly transcribed, a value between our 33 percent correct and 73 percent correct plus partially correct.

The phonetic transcription performance data that we have described could be used to simulate the front end of a sentence recognition scheme if one wished to concentrate on optimizing the lexical search strategy and the application of syntactic and semantic constraints without building a total system for dealing with the raw acoustic data. However, we would caution against overoptimism using this approach because the transcription error rate is very high, and we found it necessary to return to the spectrographic data to verify lexical hypotheses.

Since the average number of phonetic segments in a word is between four and five, and each segment has a probability of 0.73 of being transcribed without an error, then the probability of obtaining the correct word in the list of words that is printed out in response to a lexical search question is about 0.25, assuming the word boundaries are known. If word boundaries are not known, the error rate is likely to be higher. A 25 percent word-hit rate is consistent with the experience of DHK, as noted previously.

A further datum of interest deducible from Fig. 2 is that for only about five percent of the word-search questions that were asked of the computer program by DHK, did the answer consist of one and only one word that was the correct word. That is, in only about five percent of the cases was the correct word obtained from the initial phonetic data alone, without the necessity of further reference to the acoustic data and without reference to syntactic and semantic constraints on the sentences. If the lexicon size were to be increased, this percentage would probably go down, and the average number of words printed out in response to a lexical search question would go up in proportion to the relative increase in lexical items.

The probability of obtaining a correct word in the lexical search can be enhanced by two strategies: 1) by obtaining greater accuracy in the initial feature extraction procedure, possibly by making use of input data that have a better dynamic range and intensity resolution than the spectrographic data used in this experiment, by making more precise measurements on the data, or by going through some sort of speaker normalization procedure such as to have available a spectrogram of a standard known utterance spoken by the same talker whose sentence is currently being recognized, and 2) by leaving more features unspecified, particularly those features in which there is least confidence. KNS was able to improve his conditional probability of a correct complete transcription by 0.06 over that of DHK by avoiding more of the difficult decisions. As an alternative strategy, the lexical search could identify not only words that provided an exact fit to the phonetic transcription but also words that differed in one or more features from the specifications derived from the spectrogram.

There are indications that our transcription might improve in time due to learning effects. Prior to this experiment, we had dealt with spectrograms of isolated words and nonsense forms for the most part. Sentence production involves another set of acoustic-phonetic transcription rules, most of which must be discovered through study of spectrographic material, because rules involving sentence materials are not to be found in the literature.

Developing Strategies for Sentence Recognition

While improved knowledge of acoustic phonetics and a more precise input representation of the acoustic signal might provide some increase in the initial transcription score, it is probable that the improvement will not be substantial. Any strategy that relies on direct analysis of the input signal and utilizes a system of rules that transforms the results of this analysis into a pattern of segments and features, will provide error-free specification of only a small fraction of the phonetic features in the kind of sentence material used in this study. It will be possible to identify only a few of the words based on this kind of information alone. Almost all of the words were eventually identified in this study only after some hypotheses were made concerning the possible word sequences. These hypotheses were based in part on the original phonetic transcription, in part on further examination of the spectrogram, and in part on the requirement that the sentence be meaningful and syntactically correct. Reference to the spectrogram was then made to confirm or to deny the hypotheses.

Therefore, in terms of an analysis-by-synthesis paradigm [13] the observers in this spectrogram-reading experiment derived relatively little information from the initial preliminary analysis, although, of course, this information was necessary to initiate the hypothesis generation when no other information about the sentence was available.

Lindblom and Svensson [10] have shown how the spectrogram-reading task can be simplified when the subjects are provided with a formalized strategy for requesting information on prosodic features of the stimulus (which help to provide inferences about grammatical aspects of the utterance) and when the grammar of the sentence is limited and is known to the observers. Utilization of these kinds of constraints would require a lexical search through a smaller number of word candidates at each point in the recognition process. It is recognized that the strategies used by the investigators in determining features from the acoustic data did not take full advantage of acoustic attributes relating to linguistic structure (such as relative segment durations and the fundamental frequency contour). One deterrent to the use of this kind of acoustic data is that the rules relating these attributes to the segmental and syntactic aspects of an utterance are not as yet well quantified in English.

Our findings and those of Lindblom and Svensson emphasize the crucial importance of developing systematic procedures for specifying syntactic categories and semantic features of words based both on acoustic data and on sentence context, and for utilizing these data, in conjunction with the partial phonetic transcription, to carry out a lexical search. Also needed are effective techniques for verifying hypotheses about possible words or word sequences by referring to the acoustic data. These two tasks, the inclusion of syntactic and semantic constraints and the verification of hypothesized word sequences against the spectrographic representation, were performed in an *ad hoc* fashion by the experimenters in the present spectrogram-reading exercise. In developing a model for the understanding of sentences, procedures for implementing these tasks must, of course, be carefully specified.

In conclusion, it is suggested that any serious worker in the field of automatic speech recognition should undertake to read spectrograms in an organized way similar to the experiments that we have described. It is an excellent way of learning a great deal about speech, and it is the only sure way to convince oneself of the complexities involved and of the necessity for approaching the problem with more sophisticated forms of analysis.

Acknowledgment

The authors would like to thank J. Makhoul for obtaining the sentence materials upon which this study was based and for writing the computer program that performed the lexical searches.

References

[1] C. G. M. Fant, *The Acoustic Theory of Speech Production.* D's-Gravenhage: Mouton, 1960.

[2] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psych. Rev.*, vol. 74, pp. 431-461, 1967.

[3] W. Woods, "The lunar science natural language information system," Bolt Beranek and Newman Inc., Cambridge, Mass., Int. Rep., Contract NAS9-1115, 1971.

[4] N. Chomsky and M. Halle, *Sound Pattern of English.* New York: Harper and Row, 1968.

[5] G. A. Miller and P. E. Nicely, "Analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.*, vol. 27, pp. 338-353, 1955.

[6] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175-184, 1952.

[7] J. Makhoul, "Computer-assisted reading of spectrograms," presented at the 82nd Meeting of the Acoustic Society of America, Denver, Colo., Oct. 1971.

[8] R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech.* Princeton, N.J.: Van Nostrand, 1947, and New York: Dover, 1966.

[9] A. M. Liberman, F. S. Cooper, D. O. Shankweiler, and M. Studdert-Kennedy, "Why are speech spectrograms hard to read?" *Amer. Ann. Deaf,* vol. 113, pp. 127-133, 1968.

[10] B. Lindblom and S. G. Svensson, "Interaction between segmental and nonsegmental factors in speech recognition," presented at 1972 Conf. on Speech Communication and Processing, Boston, Mass., Apr. 24-26, 1972.

[11] D. H. Klatt, "A generative theory of segmental duration in English," *J. Acoust. Soc. Amer.*, vol. 51, p. 101 (A), 1972.

[12] W. A. Lea, "An approach to syntactic recognition without phonemics," in *Proc. 1972 Conf. Speech Communication and Processing,* Boston, Mass., Cat. 72 CHO 596-7 AE, 1972, pp. 198-201.

[13] M. Halle and K. N. Stevens, "Speech recognition: A model and a program for research," *IRE Trans. Inform. Theory,* vol. IT-8, pp. 155-159, Feb. 1962.

Recovering Parentheses From Spoken Algebraic Expressions

MICHAEL H. O'MALLEY, DEAN R. KLOKER, and BENAY DARA-ABRAMS

Abstract—A study of the relationship between the syntactic and prosodic organization of spoken algebraic expressions is reported. It was found that subjects were very consistent in their placement of junctures when reading algebraic expressions slowly. Furthermore, there was an almost perfect correlation between measured silence and perceived juncture. Rules were developed for inserting parentheses based on the location and measured duration of silence intervals in an utterance. Listeners were asked to insert parentheses, given the spoken form, and the consistency of their answers was measured by a chi-square test. For those cases where there was listener agreement on a single answer, the rules were tested and found to agree with the listeners from 91 to 95 percent of the time. Mathematically experienced and mathematically naive listeners displayed similar performance that suggests that the acoustic cues used by the speaker to indicate syntactic structure in this restricted domain of discourse may have a more general applicability.

Manuscript received November 1, 1972. This research was supported by ONR Contract N00014-67-A01-81-0039 and ARPA Contract DAHC04-72-C-0021.

M. H. O'Malley and D. R. Kloker are with the Phonetics Laboratory, University of Michigan, Ann Arbor, Mich. 48104.

B. Dara-Abrams was with the Phonetics Laboratory, University of Michigan, Ann Arbor, Mich. 48104. She is now with the Department of Computer Science, University of California, Santa Cruz, Calif.

I. Introduction

The precise relationship of syntactic and semantic structure to prosodic features (fundamental frequency and amplitude contours, durations, pauses, and rhythm) is an important theoretical problem in psycholinguistics and an important practical problem in automatic speech recognition. It seems clear, for example, that sentence boundaries and certain major divisions within sentences are often marked by some kind of phonological "juncture."[1] If juncture is a function of syntactic grouping, and if it can be recognized in the acoustic signal, then it could play an important role in human or machine perception of sentences.

However, the number of junctures in an utterance is a function of speech rate and the location of these junctures does not always match syntactic boundaries [1]. Furthermore, hesitation pauses, which resemble junctures in some ways, may occur at almost any point in an utterance. In addition, the acoustic signals of linguistic junctures are neither simple nor easy to measure. Some investigators [2] seem to argue that the syntax is recognized first and that perceived junctures are then derived from syntax.

Spoken algebraic expressions form an interesting special case for the study of prosodies. In contrast to most English sentences their syntactic and semantic structure is well understood. They have a context-

[1] A juncture is an abstract linguistic unit that is postulated to account for the ability of a native listener to locate certain kinds of boundaries in a spoken utterance on the basis of direct acoustic cues and/or his knowledge and expectations about the lexical, syntactic, and semantic constraints of English.