Paper presented at the 1981 NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition, Bonas, France.

ACOUSTIC-PHONETIC KNOWLEDGE REPRESENTATION: IMPLICATIONS FROM SPECTROGRAM READING EXPERIMENTS

Victor W. Zue

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

ABSTRACT

This paper presents a summary of several spectrogram reading experiments designed mainly to uncover the amount of phonetic information that is contained in the speech signal. The task involved identifying the phonetic contents of an utterance only from a visual examination of the spectrogram. The results generally support the notion that there is a great deal of phonetic information in the speech signal that can be extracted by the proper application of phonetic rules. From these results, it is argued that phonetic recognition in speech recognition systems can be improved substantially, and that improved phonetic recognition will lead to speech recognition systems of greatly increased complexity and sophistication.

1. INTRODUCTION

Speech recognition is a research topic that has fascinated many speech scientists over the past several decades. Prior to the 1970s, research efforts were primarily directed towards isolated word recognition (Hyde, 1968). Since 1970, we have witnessed a flourish of research activities in continuous speech recognition. Such activities in the U.S. were at least in part due to the 5-year, multi-site effort initiated by the Advanced Research Project Agency of the U. S. Department of Defense. The ARPA project terminated abruptly in 1976, with some of the contractors meeting the original goal. [For a detailed treatment of the initial ARPA goal and a review of the final systems, see Newell et al. (1971) and Klatt (1977), respectively]. Since then, research

in speech recognition has, with few exceptions, taken on a different emphasis. These more recent systems generally operate in a speaker-dependent mode, and recognition is accomplished by recognizing isolated words or phrases using mostly pattern recognition techniques (Lea, 1980). Efforts to generalize such techniques to connected words and phrases have been met with limited success.

Almost all the state-of-the-art recognition systems, whether for isolated-word or continuous speech recognition, depend heavily for their success on the training of the system to a particular speaker's voice. In almost every case, the speaker has to train the system on each word or phoneme. The recognition is performed by matching the input speech to the stored templates and picking the template that is the closest to the input, using some distance measure. While such techniques have been quite successful for a system with a limited vocabulary and for a given speaker, they do not come close to human performance on virtually unlimited speaker and vocabulary.

There are two general types of pattern matching approaches. The first involves storing individual patterns for each word in the dictionary. Such an approach becomes increasingly impractical as the dictionary grows, both in storage and computational requirements. Furthermore, such an approach leads to little flexibility in handling word boundary effects, which can cause major distortion of a given word compared to its form if enunciated in isolation.

The second approach is to store a dictionary of templates deemed to be representative of all of the phonemes of the language. Each word is then entered as a sequence of such templates. This approach is better able to handle word boundary effects, by incorporating appropriate rules into the dictionary structure. However, subtle but specific variations in a given phoneme which may yield important clues to the identity of its neighboring phonemes are generally totally ignored, and in fact would probably serve to lower its score against a "canonic" form stored in the dictionary. As an example, we offer /s/ in an /sk/ environment, which often shows a concentration of energy near 1800 Hz in addition to the usual high frequency energy. This acoustic cue, when present, is extremely reliable in identifying the following velar consonant. In a template matching scheme, such an /s/ would probably still score best against an /s/ template, but there would be no mechanism for utilizing the 1800 Hz peak to aid in the identity of the /k/.

Another problem with phoneme template matching schemes is that they generally weigh all of the information equally regardless of the spoken speech sound. In the phoneme /u/ of American English,

the first formant will always be low in frequency, but the second formant can be almost anywhere in the F2 region. One solution would be to store a whole sequence of dictionary entries for /u/ covering the entire F2 range. It would however be better to focus on F1 for identifying /u/, and use the location of F2 to aid in the identity of adjacent phonemes. Pattern recognition has no place for such decision strategies.

There are at least two other reasons why template matching alone can not be expected to give adequate performance as the systems grows in complexity and usage. First, regardless of the particular type of representation of the speech signal (raw spectrum, LPC coefficients, etc.), there does not exist a set of theoretically motivated distance metrics. As a consequence, acoustic information related to the linguistic (or phonetic) identity of a given sound is confounded with variability related to speaker specific characteristics, and the same distance metric is applied indiscriminately for the different types of speech sounds. Second, the templates tend to be highly speaker-dependent. The adaptation of a system to new speakers is likely to be a time consuming and painful process. The alternative approach, the one that we are proposing, is to incorporate our knowledge of the acoustic characteristics of speech sounds, and how these characteristics are modified by the environments, in order to provide a representation that is more linguistically based.

It is our belief that acoustic-phonetic knowledge representation is a major roadblock in the design of advanced speech recognition systems that are meant to approach human performance. In particular, we believe that the ability of humans to perceive speech in hostile acoustic environments, spoken by familiar as well as unfamiliar speakers, cannot be duplicated easily by machine without adequately incorporating these knowledge sources. Our belief is at least in part due to the results of recent spectrogram reading experiments which have shown that the acoustic signal is rich in phonetic information.

The purpose of this paper is to summarize several spectrogram reading experiments conducted over the past three years, describe the factors contributing to such a positive result, and speculate on the implication of our research on the design of speech recognition systems.

2. SPECTROGRAM-READING EXPERIMENTS

Ever since the invention of the sound spectrograph some thirty years ago (Koenig et al., 1946), the spectrogram has been the single most widely used form of display for speech. The popularity of the spectrogram is at least partly due to the fact that it is

relatively easy to produce, and that it displays the temporal and spectral characteristics that are known to provide the segmental cues to speech sounds. To be sure, a speech spectrogram sometimes introduces distortions to the acoustic structure of speech and often does not provide adequate information on certain linguistically relevant cues, such as stress and intonation. Nevertheless, a speech spectrogram gives a good description of the segmental acoustic cues of speech, and it has been an invaluable tool in the development of our understanding of speech production and perception.

A speech spectrogram of the utterance, "The soldiers knew the battle was won," is shown in Figure 1. As Figure 1 reveals, a spectrogram provides a display of the energy in the speech wave in terms of frequency -- along the vertical axis; time -- along the horizontal axis; and intensity -- by the darkness of the markings.

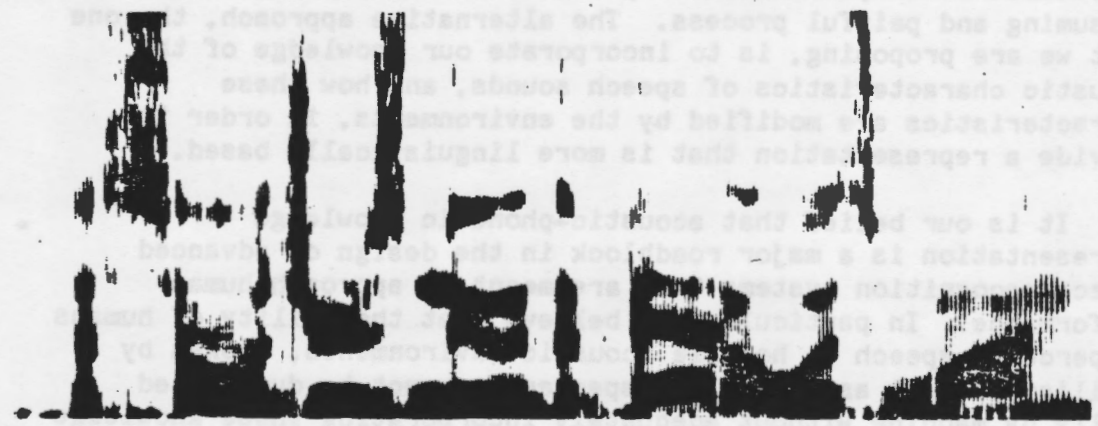

Figure 1 Spectrogram of the utterance "The soldiers knew the battle was won." The sentence was spoken by a male speaker and the spectrogram was made on a Voiceprint Laboratory Sound Spectrograph.

Is it possible to "read" a spectrogram? With sufficient practice, can one examine the spectrogram of an unknown utterance and determine what was said? The common belief is that it is not possible (Fant, 1962; Liberman, 1968; Lindblom and Svensson, 1973) and that no amount of training will allow one to do so. The results of past spectrogram reading experiments by Potter et al. (1947), Klatt and Stevens (1973), and Svensson (1974) are consistent with such a belief. In the classic book **Visible Speech**, Potter and his colleagues reported an experiment in which they attempted to teach normal as well as hearing-impaired subjects to communicate with each other using the Direct Translator, a

real-time spectrographic display. The best subject, a deaf engineer, was able to acquire a vocabulary of 800 words after 200 hours of training. Vocabulary acquisition appeared to be a steady linear function of the amount of training; a word took about 15-20 minutes of practice to acquire. Since the visual patterns were recognized at the word level, it is difficult to assess their results in terms of phonetic recognition.

Klatt and Stevens (1973), serving as their own subjects, attempted to phonetically label a set of 19 spectrograms of unknown English sentences. In order to minimize the possibility of recognizing words in spectrograms, a "mask" was placed over each spectrogram so that only 300 msec of speech was visible at one time. In addition, each spectrogram was read in a single pass from left to right. Klatt and Stevens correctly identified approximately 33% of all segments, and provided a correct partial specification (e.g. "plosive", "nasal") on an additional 40%.

As part of his doctoral dissertation, Svensson (1974; see also Lindblom and Svensson, 1973) asked students who had taken a spectrogram-reading course to phonetically label spectrograms of unknown utterances. The subjects were provided with a flow chart describing a set of binary decisions to use when reading spectrograms. In addition, they were told to provide only one label per segment, and were given an hour to interpret each spectrogram. The results, according to Svensson, were disappointing, with performance ranging from a low of 22% to a high of 51% with an average of 38% of the segments correctly labeled.

In this section we report the results of several spectrogram-reading experiments that we have conducted over the past three years. These experiments collectively address several questions. First, how well can unknown spectrograms be "read". More specifically, how much phonetic information can be extracted from the acoustic signal. Second, how can the incorporation of other knowledge such as syntax, semantics and pragmatics improve the speed and accuracy of spectrogram reading. Third, how (and if) such a skill can be learned by others. Since the speed at which spectrograms can be read is of little concern to speech recognition researchers, the second question will not be addressed in this paper.

2.1 How Accurately Can Spectrograms Be Read?

The first experiment, designed to uncover the amount of phonetic information that can be extracted from a spectrogram, was reported by Cole et al. (1980). The task involves the phonetic transcription of unknown English utterances by visual examination of the spectrogram. In this experiment, spectrograms were made from 23 English utterances spoken by two male speakers. Each

utterance is approximately 2 sec in length and contains an average of 6.4 words. The utterances consisted of normal English sentences, as well as phonetically, syntactically, and semantically anomalous sentences. In addition, spectrograms of 45 words in the carrier phrase, "Say --- again," were also prepared. This second set of spectrograms was intended to uncover to what extent word boundary information may influence performance.

In order to measure the accuracy of the reader's transcriptions, we asked three trained phoneticians to transcribe the utterances by listening to them. Figure 2 compares the transcription produced by the reader with that produced by each transcriber (T) for the sentence shown in Figure 1.

In order to evaluate segmentation accuracy, we arbitrarily adopted the criterion that a segment is assumed to exist when two

---

VZ: ð ə s o^{w} ǰ ɚ z n ü ð ə b ɛ ɾ ə w ʌ z w ɔ n
č ʒ yu v v æ ! l ə

T1: ð ə s o l ǰ ɚ z n u ð ə b æ ɾ ! w ɨ z w ʌ n

T2: ð ʌ s o l ǰ ɚ z n u ð ə b æ ɾ ! w ʌ z w ɔ n

T3: ð ə s o l ǰ ɚ z n u ð ə b æ ɾ ! w ə z w ʌ n

Figure 2 Transcriptions produced by the reader while reading the spectrogram, and by the three Ts who listened to the speech. The utterance is the one shown in Figure 1.

---

or more Ts produced a segment label. According to this criterion, there were 499 segments in the 23 utterances, of which the reader identified the existence of 485, or slightly more than 97%. It should be noted that segmentation is not an unambiguous process. In fact, 20 alternative segment paths were introduced, of which 13 were correct. It is interesting to note, however, that no extra segments were introduced by the reader.

The reader used a single label to identify a segment 52% of the time, two labels 35% of the time, and three labels 6% of the time. On the remaining 33 segments, the reader produced 17 partial transcriptions, 13 optional segments, and provided no label in three cases. If we exclude optional labels from consideration, and count each partial transcription as three labels, then the reader produced an average of 1.53 labels to each segment. When more than one label was given, they were almost always rank ordered so that it was possible to score the results for first, second, and third choices. We arbitrarily decided to score all partial transcriptions as equivalent to a third choice.

Table 1 summarizes the results on labeling. It can be seen that segment labels produced by the reader agreed with at least one T on 425 of 499 segments, or 85%. Approximately two out of three

TABLE 1
Agreement on Segment Labels between the Reader and Any T

	Proportion	%
First Choice	335/499	67
Second Choice	64/499	13
Third Choice	26/499	5
TOTAL	425/499	85

times the agreement is in the first choice. Closer examination indicates that there is a tendency for better agreement in consonants than vowels. This measure of labeling accuracy might seem, at first glance, quite generous, since the reader's labels were compared individually with that by each transcriber. However, it must be kept in mind that phonetic transcription is an interpretation imposed upon a continuously varying signal, and that transcribers rarely agree completely on the transcription of a given utterance. In this experiment, the agreement among all three transcribers is approximately 85%.

A detailed analysis of the reader's performance revealed a number of insights into his methods. Perhaps most interesting, higher-level knowledge about the syntactic and semantic structure of English was rarely used to interpret spectrograms. Performance was significantly better on sequences of words and nonsense words (92%) than on English sentences (83%), suggesting the use of phonological compensation by the transcribers based on contextual information.

The results on words in a known carrier phrase represents substantial improvement over that of normal and anomalous sentences. The existence of all 201 segments was correctly identified by the reader. The labels produced by him agreed with at least one of the Ts 93% of the time.

2.3 Can Spectrogram Reading Be Taught?

The results cited in the previous section are encouraging in that they illustrate the richness of phonetic information in the acoustic signal. These results in themselves are of little scientific interest, however, if it can not be demonstrated that such skill can be represented by explicit knowledge and thus can be

taught to others. With respect to the design of advanced speech recognition systems, the ability to transfer such knowledge to others must first be demonstrated before any claims can be made regarding improved machine performance.

Several attempts have been made to teach the spectrogram reading skill to others. One such attempt was reported by Cole and Zue (1980) in which five students who enrolled in a graduate course in speech production and perception were asked to label phonetically five unknown spectrograms. The students' performance was compared to the transcriptions produced by two experienced phoneticians who listened to the sentences. Working as a group, the students identified the existence of approximately 94% of the segments. Averaging over all classes of sounds, the students agreed with either of the transcribers on the first, second, and third choices, respectively, 51, 24, and 8% of the time.

In a separate experiment, Seneff (1979) reported on the results of a spectrogram reading experiment that involved phonetic labeling and word hypothesizing for a set of 49 sentences spoken by one male speaker. Prior to the reported experiment, Seneff estimated herself to have studied about 30 to 40 spectrograms of continuous speech. During the experiment, she was provided with feedback immediately following the completion of each spectrogram. The feedback included not only the correct answer, but also instructions regarding acoustic cues or phonological processes. Phoneme identification scores improved steadily from 59% for the first ten sentences to 69% for the last ten sentences for an average of 64%. It should be noted that Seneff allowed herself only one label per segment.

The results reported by Seneff are more significant than those reported by Cole and Zue (1980), in this author's opinion, for several reasons. First, the corpus of the Seneff study (49 sentences and 1153 segments) is considerably larger than that reported by Cole and Zue (5 sentences and 104 segments). Second, it is very difficult to infer from Cole and Zue the performance of individual students, since the students worked as a group throughout ther experiment, whereas in the second experiment Seneff was the only subject. Finally, the experiment by Seneff was more controlled both in style and content, and a careful log of the learning experience was kept, thus allowing a more direct assessment of the transferability of this skill.

In addition to the two experiments described above, the spectrogram reading course has been given by this author on at least three other occasions to groups of students ranging in number from 8 to 13. The students varied significantly in their background, ranging from seasoned speech researchers to naive subjects with no exposure to acoustic phonetics. The course

usually started with a brief description of the acoustic theory of speech production and the acoustic characteristics of consonants and vowels in stressed consonant-vowel (CV) environment. This is then followed by a description of the coarticulatory and phonological processes that govern the concatenation of segments into larger units. By the end of the course, which is approximately 40 hours in duration, the students usually have been exposed to spectrograms of approximately 50 sentences. Although a formal test of the subjects was not performed, it is estimated conservatively that the phonetic recognition accuracy on sentences is about 50%.

3. PHONETIC KNOWLEDGE REPRESENTATION

The conclusion to be drawn from the experiments described in the previous section is that the acoustic signal is rich in phonetic information, that such information can often be extracted from a spectrographic representation of the signal, and that the process of extracting this information often involves the application of explicit rules that can be taught to others. The results are all the more interesting considering that past attempts at spectrogram reading have met with much less encouraging results. These results are also significantly better than those reported in the literature on phonetic recognition in speech recognition systems (Klatt, 1979). What, one must logically ask, is the cause of such improved performance?

One of the most important factors that contributed to the improved performance in spectrogram reading is our improved understanding of the acoustic characteristics of fluent speech. To be sure, there has been ongoing research on the acoustic properties of speech sounds over the past few decades, and a great deal of knowledge has been acquired. However, with few exceptions, these research efforts have been focused on the acoustic properties of consonants and vowels in stressed consonant-vowel syllables. It was not until the past decade that researchers began to focus on the acoustic characteristics of speech sounds in continuous speech. We now have a much better understanding of the properties of speech sounds in different phonetic environments [see, for example, Umeda (1974), Kameny (1975), Klatt (1975), Zue (1976), Umeda (1977), and other citations in the following sections]. Furthermore, we are beginning to develop a quantitative understanding of the phonological processes governing the concatenation of words [see, for example, Oshika et al. (1975), Cohen and Mercer (1975)]. In fact, a few of the effects have been studied in some detail (Zue and Laferriere, 1979; Zue and Shattuck-Hufnagel, 1980). Therefore, our improved ability to interpret spectrograms is primarily due to the collected research efforts of many researchers. In addition, as a consequence of studies on the properties of speech sounds and

of the auditory responses to speech-like sounds (see, for example, Kiang, 1980), we are gaining better insights into how the speech signal is processed in the auditory system, what portions of the signal are carrying the principal information concerning the distinctive phonetic dimensions, and what portions show more variability with respect to these dimensions. For example, the role of the burst spectra and amplitudes and of rapid onsets and offsets in identifying place of articulation and other features for consonants has been documented (Zue, 1976; Blumstein and Stevens, 1979).

To summarize, it is our feeling that our ability to extract a great deal of phonetic information from the acoustic signal is primarily a reflection of our improved understanding of the factors that contribute to the phonetic identities of speech sounds and their acoustic correlates. Spectrogram reading is nothing more than a demonstration of the proper utilization of our gained knowledge of how the acoustic cues for phonetic contrasts are encoded in the speech signal. Native speakers of a language demonstrate this ability whenever they communicate with each other by voice.

In the next section, we will attempt to identify the various ways phonetic knowledge is encoded in the speech signal and show how these sources of knowledge are utilized in spectrogram reading. We will argue that a phonetically-based speech recognition system must incorporate these knowledge sources.

3.1 Acoustic Phonetic Knowledge Sources

If the ultimate goal of a speech recognition system is to reach the performance of human speech perception, i.e., extremely high recognition on nearly all speakers, it would be important to know how acoustic phonetic knowledge is being utilized by humans. After all, the task of phonetic recognition faced by a computer is not unlike that faced by the native speaker of a language when trying to decode the utterance spoken by another native speaker. Aside from knowing the inventory of the speech sounds used in a language, i.e., the phonetic segments and their acoustic correlates, there are several levels at which acoustic phonetic knowledge is utilized heavily. We shall address each one of these knowledge sources in turn.

The first of these sources involves the constraints imposed by the language on the allowable sequences of speech sounds. In American English, there are only some forty possible consonant clusters that can start a word. Thus, for example, a native speaker knows that "vnuk" is not an English word. As another example, if we specify that the third phoneme of a three-element, word-initial cluster is /l/, then the remaining two phonemes (/s/

and /p/, as in "splash") are uniquely determined **without** the need of further acoustic evidence. The utilization of sequential constraints is not restricted to the word-initial position. In American English, for example, nasal-stop clusters almost always follow the homorganic rule requiring that the nasal and the stop have the same place of articulation (as in "camp", "tent"). The proper identification of one element of the cluster again determines the other element without the need of further acoustic evidence. The knowledge of constraints on allowable phoneme sequences enables the listener to fill in phonetic details that might be distorted or even absent in the acoustic signal. A speech recognition system must incorporate this knowledge such that it can also infer phonetic identity of speech sounds in the absence of clear acoustic cues.

When speech sounds are connected to form larger linguistic units, the canonic acoustic characteristics of a given speech sound will change as a function of its immediate phonetic environment. As an illustrative example, consider the utterance, "Tom Burton

Figure 3 Spectrogram of the sentences, "Tom Burton tried to steal a butter plate," spoken by a male speaker. The spectrogram illustrates the various acoustic realizations of the phoneme /t/.

tried to steal a butter plate," shown in Figure 3. Every word, except "a", in this sentence contains a single occurrence of the phoneme /t/. However, depending upon the immediate phonetic environment and stress pattern, the underlying /t/'s are realized alternatively as an aspirated /t/ ("Tom"), an unaspirated /t/ ("steal"), a retroflex /t/ with extended aspiration ("tried"), an unreleased /t/ ("plate"), a flap ("butter"), or a glottal stop ("Burton"). The acoustic characteristics of these realizations are seen to be drastically different.

The modification of the acoustic properties of speech sounds as a function of the phonetic environment is not a phenomena that is restricted to within a word. When words are concatenated to form phrases and sentences, significant acoustic changes can result, as evidenced in the following example. Figure 4 shows a spectrogram of the seven words "did", "you", "meet", "her", "on", "this", and "ship", spoken in isolation as well as in a sentence,

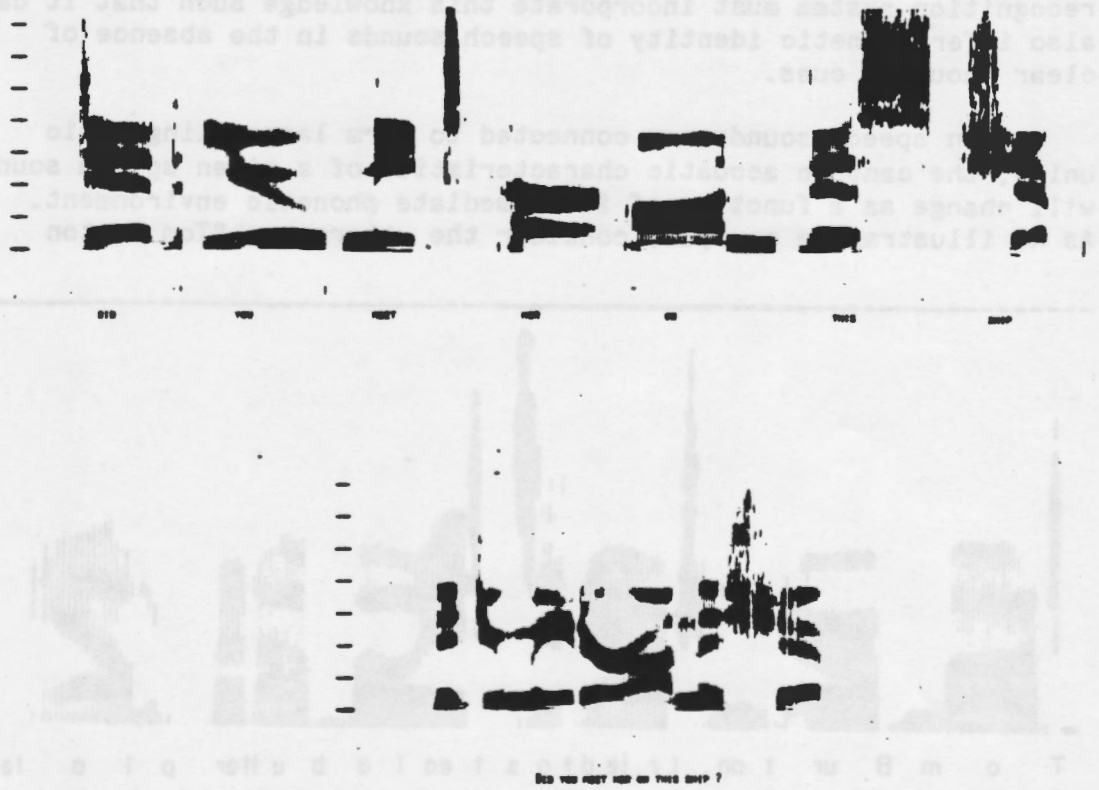

Figure 4 Spectrogram of the words "did", "you", "meet", "her", "on", "this", "ship", spoken in isolation and in a sentence by a male speaker. The spectrogram illustrates the phonological processes such as palatalization and flapping that can operate at word boundaries.

"Did you meet her on this ship?" We can see, for example, that the word-final /d/ and the word-initial /y/ in the word pair "did you" are realized acoustically as a single /ǰ/; the word-final /t/ and the word-initial /h/ are realized as a single flap; and the word-final /s/ and the word-initial /š/ in the word pair "this ship" are realized as a single, long /š/. These kinds of phonetic changes at word boundaries, particularly when there are adjacent word-final and word-initial consonants, are extremely common in

normal American English. In order to properly perform lexical access, the nature of these phonological rules must be understood.

Whereas it is true that the acoustic realizations of phonetic segments are highly context-sensitive, most of the variations, such as the ones illustrated in Figures 3 and 4, are, in fact, systematic and can be captured by explicit rules (e.g. /t/ becomes a glottal stop [?] when preceded by a stressed vowel and followed by a syllabic nasal [n], as in "Burton"). Over the past decade, research in fluent speech has enabled us to gain a good understanding of the nature of these rules and how they interact. Although our present knowledge of the inventory of such rules is still incomplete, such knowledge, however fragmented, must be incorporated into a speech recognition system such that words can be recognized from the seemingly ambiguous acoustic signals.

Another factor affecting phonetic knowledge representation is the fact that a given phonetic contrast often has multiple acoustic cues. For example, the voicing characteristic of English stop consonants can manifest itself in many ways (Klatt, 1975; Zue, 1976). Some of the more salient acoustic cues include: a) the prevoicing during closure, b) the duration of the voice-onset time (VOT), c) the extent of the first formant transition, d) the amplitude of the burst, e) the fundamental frequency contour, and f) the duration of the preceding vowel. While it is often true that these multiple acoustic cues provide redundant phonetic information, there are also situations where only selected cues are available to aid recognition. For example, if a stop consonant is not released (as the /t/ in "basketball"), then those cues related to the release will not be present, thus rendering a decision algorithm based on VOT or the burst spectra useless.

We have demonstrated the fact that phonetic contrasts often have multiple acoustic cues. On the other hand, a given acoustic cue could also have multiple causes. For example, the duration of phonetic segments is influenced by a multitude of factors ranging from phonetic and phonological (such as inherent duration, stress, segmental interaction) to syntactic, semantic, and discourse knowledge (Klatt, 1976). Therefore, the application of acoustic phonetic rules must not be binding so that the effects of other factors can be taken into account when additional knowledge about the utterance becomes available.

One other factor that is important in the representation of acoustic phonetic knowledge is the fact that acoustic phonetic information is encoded differentially in the speech signal. Speech perception experiments (Cutler and Foss, 1977), spectrogram-reading experiments (Klatt and Stevens, 1973; Cole et al., 1980), and analysis of early speech recognition systems (Lea, 1980) all seem to indicate that humans and machines can perform better recognition

for vowels in stressed syllables and consonants preceding stressed syllables, suggesting that the acoustic information is more robust around these syllables.

3.2 Example of Acoustic-Phonetic Knowledge Utilization in Spectrogram Reading

Having discussed the multitude of factors that influences the representation of acoustic phonetic knowledge, we will now turn to an example of how these knowledge sources interact during spectrogram reading. We want to stress the fact that the spectrogram-reading aspect of our example is only a method to demonstrate the knowledge interaction. We do not feel strongly that phonetic recognition should necessarily be performed on such a representation of the speech signal. In fact, as we shall see, some acoustic information is poorly represented in a spectrogram.

Figure 5 shows the spectrogram of an unknown utterance spoken by a male speaker. The spectrogram was generated digitally, and the display included the speech waveform and several parameters that, we believe, complement a conventional spectrogram. The quality of the digital spectrogram is, in our opinion, comparable to its analog counterpart. We will attempt to walk through this utterance from left to right, illustrating the process of applying acoustic-phonetic knowledge.

There is a well-enunciated glottal stop at the beginning of the sentence (t=0.17 sec) suggesting that the sentence began with a vowel, with no preceding consonant. The glottal stop appears as an irregularity in the fundamental frequency, which is visible both in the time waveform and in the spectrogram. (Another cue for the absence of a preceding consonant is the absence of noticeable rising F1 transition, typical from a consonant to a vowel. Since F1 and F2 are both high (around 700 Hz and 1800 Hz, respectively) in the middle of the vowel, the most likely choice for the vowel is /æ/. (/ɛ/ may be a second choice, but the long duration of the vowel makes it a distant second choice.)

Following the vowel is a region of low acoustic output except in the low-frequency portion. Since the energy dropped abruptly from the preceding vowel, the phoneme is probably a voiced stop. The identity of the stop can be sought by a process of elimination. If the stop were a labial stop (/b/), one would expect all formants to fall sharply at the end of the vowel. If it were a velar stop (/g/), then one would expect F2 and F3 to move towards each other. Since there is little motion in F2 or F3 at the end of the vowel, the most likely candidate is a /d/.

It might be supposed that the burst of energy following the stop gap (t=0.38 sec) is the burst of the /d/. However, it is

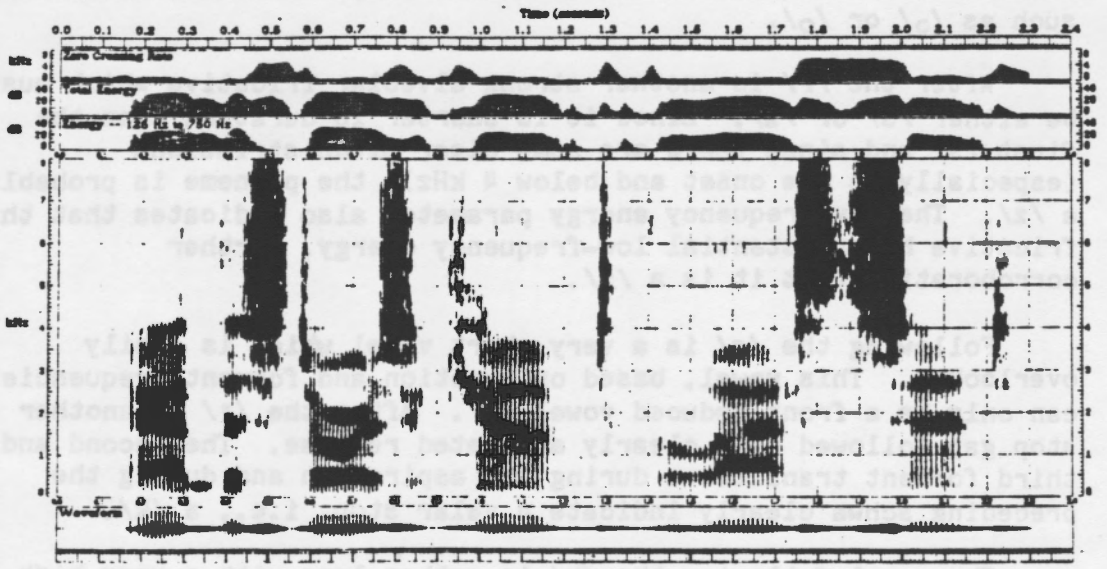

Figure 5 Digital spectrogram of an unknown utterance which is used to illustrate the application of acoustic-phonetic knowledge. In addition to the spectrogram, the waveform and several parameters are also displayed to complement the information available on a spectrographic display.

instead associated with the voiced fricative /ð/, which is often realized in a stop-like manner when preceded by a consonant. The stop is therefore unreleased. A true /d/ burst would be stronger in amplitude and lower in frequency than the frication noise generated for /ð/. (See, for example, the stop burst at t=0.57 sec.)

Following the phoneme /ð/ is a short, moderately front, vowel, possibly an unstressed /ɪ/. After the short vowel is a strong fricative (a definite /s/) followed by a short gap and a burst of energy (t=0.57 sec) which exhibits frequency characteristics typical of a dental stop (/d,t/). Since there is little aspiration following the burst, a /d/ or an unaspirated /t/ are the likely choices. The stop could therefore either be a /d/ with a boundary preceding it (since /sd/ is not an allowable English cluster), or a /t/ in a cluster with the /s/ (/t/ is not aspirated when it appears in a cluster with /s/).

After the stop is a vowel with sizeable second formant motion. Since F3 is quite low, especially for the second-half of the vocalic segment, a postvocalic /r/ is proposed. The vowel counterpart /ɝ/ is ruled out because F2 is much too low. After taking into account the formant transitions due to the preceding

alveolar stop and the following /r/, we propose a low back vowel such as /ɔ/ or /o/.

After the /r/ is another strong alveolar fricative which must be either /s/ or /z/. Since it is shorter in duration than the first /s/ and since there are some clear pitch striations (especially at the onset and below 4 kHz), the phoneme is probably a /z/. The low-frequency energy parameter also indicates that this fricative has substantial low-frequency energy, further corroborating that it is a /z/.

Following the /z/ is a very short vowel which is easily overlooked. This vowel, based on duration and formant frequencies, can only be a front reduced vowel /ɪ/. After the /ɪ/ is another stop gap followed by a clearly aspirated release. The second and third formant transitions during the aspiration and during the preceding schwa clearly indicate a velar stop, i.e., a /k/.

The vowel following the /k/ is rather long with a very high F1. The high F1 suggests /a/ or /æ/. However, /æ/ is ruled out because F2 is too low (compare with the word "add"). Notice that there is also quite a bit of motion in F2, suggesting the possibility of a diphthong. However, F1 seems to be quite steady, although somewhat diffuse and weak in energy. An experienced reader will recognize the diphthong /aw/ in a nasalized environment. The additional nasal pole and zero in the low-frequency portion of the vowel spectra cause the appearance of an unusually high F1. Both the nasal formant and the vowel formant are visible in this digital spectrogram, since the analyzing filter is narrow enough to resolve them.

When there is a nasalized vowel there is usually a nasal. However, in this case the nasal is very hard to find. In fact, a phonological rule involving homorganic nasal-stop clusters must be invoked. The rule states that the duration of a homorganic nasal is greatly reduced when it is followed by a voiceless stop (Zue and Laferriere, 1979). Since F2 is rising slightly at the end of the vowel, /nt/ is a likely choice.

The stop gap ends in a strong burst with spectral characteristics indicating an alveolar place of articulation. It is tempting to associate the burst with the /t/ in the /nt/ cluster. However, the stop gap is quite long in duration, particularly unusual for a stop in a cluster. Thus a more appropriate hypothesis would be that there is another /t/ following the /nt/ cluster and that the first /t/ is, in fact, unreleased.

The release of the /t/ (t=1.28 sec) is quite long, to be followed by yet another consonant which is low in amplitude. Phonotactic considerations rule out the possibility of this /t/

forming a cluster with another consonant. Thus, the aspiration of the /t/ probably included a devoiced schwa, a common acoustic realization of an unstressed "to" in continuous speech.

The weak consonant following the aspirated /t/ is voiced, as evidenced by the periodicities in the waveform. The only candidates are the two voiced fricatives /v,ð/ or the voiced stop /b/. (/d/ and /g/ are not probable due to the relatively weak release.) Given our previous discussion on the realization of /ð/ following another consonant, /ð/ appears to be the most logical choice.

The consonant is followed by a schwa, which is followed by an intervocalic sonorant. The slow formant transitions suggest that the sonorant is either a /w/ or an /l/. /l/ is the more likely candidate due to the distinct separation of F1 and F2 and the rather marked discontinuity leading into the following vowel.

Following the /l/ is a relatively long, fairly steady-state vowel with a reasonably high F1 and mid range F2. The vowel /æ/ is the only vowel to fit this category. Note that the second formant for this /æ/ is somewhat lower than that of the first /æ/ we encountered. This is due to the coarticulatory effect of the /l/, which tends to lower the second formant of front and central vowels.

After the /æ/ is another /s/, extremely long in duration, with a fading out of energy mid way through. The length alone suggests that there must be more than one segment. In fact, this pattern is characteristic of the /sts/ cluster, with the /t/ closure never completely formed.

The final vowel is clearly nasalized, and, as in the previous nasalized vowel, no distinct nasal murmur is observed. The spectral characteristics of the release enable us to propose another /nt/ cluster. The first formant for the vowel is very difficult to locate, a problem often encountered with nasalized vowels. In such situations one learns to rely mainly on F2. The only vowels in English that can have such a second formant value are /ɛ/ and /æ/. Since the vowel is not very long, especially for the last segment of an utterance, /ɛ/ is probably a better choice than /æ/.

The string of phonemes that we have proposed thus far is shown in Figure 6. Even with an open lexicon, the number of words that can be proposed is quite small. After ruling out several potential word strings on syntactic or semantic grounds, the reader will probably accept "Add the store's account to the last cent" as the most likely candidate for a complete sentence.

ʔ æ dˀ ð ɪ s t ɔ r z ɨ k aw n tˀ tʰ ð ə l æ s t s ɛ n t
ɛ z d o æ

Figure 6 Proposed phoneme string for the utterance shown in Figure 5.

4. IMPLICATIONS TO SPEECH RECOGNITION

The design philosophy of most speech recognition (or understanding) systems in the past decade has been based on the belief that the acoustic signal does not provide sufficient information to identify the linguistic content of an utterance. In fact, the change in terminology from "speech recognition" to "speech understanding" reflects a departure from the view that speech can actually be "recognized" by machine, and the acceptance of the view that syntactic, semantic, and pragmatic knowledge must be used heavily to recognize an utterance (Reddy, 1976). While we are in no way disputing the importance of higher-level knowledge sources in the process of speech recognition by humans and machine, the results from spectrogram reading experiments clearly show that there is a great deal more phonetic information in the speech signal than was previously believed, and that such information is often explicit and can be captured by rules.

As we have stated earlier, it is our belief that acoustic phonetic recognition is the major road-block in the design and implementation of advanced speech recognition systems with capabilities approaching that of humans. We believe that developing phonetic recognizers with significantly better performance than those reported in the literature should be possible, and such improved phonetic recognition will enable us to tackle problems with relaxed constraints imposed by syntax, semantics, and pragmatics, thus enabling us to approach problems of increased complexity and sophistication.

Just exactly how the phonetic knowledge should be captured and represented is a matter that must be tackled jointly by researchers in diverse disciplines. The speech scientists and acoustic phoneticians must specify the phonetic rules of a given language in quantitative detail. Parameters capturing the relevant acoustic characteristics must be determined and extracted from the speech signal. Researchers in artificial intelligence must specify how such knowledge should be represented in the computer such that information can be accessed and digested in a cohesive manner. We are far from solving the problem, but the spectrogram reading experiments do provide us with reenforced conviction that the problem may be solved in the not so distant future.

REFERENCES

Blumstein, S.E. and Stevens, K.N. (1979) "Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characteristics of Stop Consonants," J. Acoust. Soc. Am., Vol. 66, No. 4, 1001-1017.

Cohen, P.S. and Mercer, R.L. (1975) "The Phonological Component of an Automatic Speech Recognition System," in **Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium**, ed. D.R. Reddy, 275-320, (Academic Press, New York).

Cole, R.A. and Zue, V.W. (1980) "Speech as Eyes See It," Chapter 24 in **Attention and Performance VIII**, ed. R.S. Nickerson, 475-494 (Lawrence Erlbaum Asso., Hillsdale, New Jersey).

Cole, R.A., Rudnicky, A.I., Zue, V.W., and Reddy, D.R. (1980) "Speech as Patterns on Paper," Chapter 1 in **Perception and Production of Fluent Speech**, ed. R.A. Cole, 3-50 (Lawrence Erlbaum Asso., Hillsdale, New Jersey).

Cutler, A. and Foss, D.J. (1977) "On the Role of Sentence Stress in Sentence Processing," Language and Speech, Vol. 20, 1-10.

Fant, G. (1962) "Descriptive Analysis of the Acoustic Aspects of Speech," Logos, Vol. 5, 3-17.

Hyde, S.R. (1972) "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature," in **Human Communication: A Unified View**, edited by E.E. David and P.B. Denes (McGraw-Hill, New York).

Kameny, I. (1975) "Comparison of Formant Spaces of Retroflexed and Nonretroflexed Vowels," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, 38-49.

Kiang, N. S.-Y. (1980) "Processing of Speech by the Auditory Nervous System," J. Acoust. Soc. Am., Vol. 68, 830-835.

Klatt, D.H. (1975) "Voice Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters," J. Speech and Hearing Research, Vol. 18, 686-706.

Klatt, D.H. (1976) "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," J. Acoust. Soc. Am., Vol. 59, No. 5, 1208-1221.

Klatt, D.H. (1977) "Review of the ARPA Speech Understanding Project," J. Acoust. Soc. Am., Vol. 62, No. 6, 1345-1366.

Klatt, D.H. and Stevens, K.N. (1973) "On the Automatic Recognition of Continuous Speech: Implications from a Spectrogram-Reading Experiment," IEEE Transactions on Audio and Electroacoustics, AU-21, 210-217.

Koenig, W., Dunn, H.K., and Lacey, L.Y. (1946) "The Sound Spectrograph," J. Acoust. Soc. Am., Vol.18, 19-49.

Lea, W.A. (1980) **Trends in Speech Recognition**, (Prentice-Hall, Englewood Cliffs, New Jersey).

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1968) "Why Are Speech Spectrograms Hard to Read?" American Annals for the Deaf, 1968, Vol. 113, 127-133.

Lindblom, B.E.F. and Svensson, S.G. (1973) "Interaction between Segmental and Nonsegmental Factors in Speech Recognition," IEEE Transactions on Audio and Electroacoustics, AU-21, 536-545.

Newell, A., Barnett, J., Forgie, J.W., Green, C.C., Klatt, D.H., Licklider, J.C.R., Munson, J., Reddy, D.R., and Woods, W.A. (1973) **Speech Understanding Systems: Final Report of a Study Group** (North-Holland/American Elsevier, Amsterdam).

Oshika, B.T., Zue, V.W., Weeks, R.V., Nue, H., and Aurbach, J. (1975) "The Role of Phonological Rules in Speech Understanding Research," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, 104-112.

Potter, R., Kopp, G., and Green, H. (1947) **Visible Speech**, (van Nostrand, New York).

Seneff, S. (1979) "A Spectrogram Reading Experiment," Term paper submitted for a Graduate Course on Sound, Speech, and Hearing, Massachusetts Institute of Technology.

Svensson, S.G. (1974) **Prosody and Grammar in Speech Perception**, Monographs from the Institute of Linguistics, University of Stockholm, (MILOS), Vol. 2.

Umeda, N. (1975) "Vowel Duration in American English," J. Acoust. Soc. Am., Vol. 58, 434-445.

Umeda, N. (1977) "Consonant Duration in American English," J. Acoust. Soc. Am., Vol. 61, 846-858.

Zue, V.W. (1976) "Acoustic Characteristics of Stop Consonants: A Controlled Study," Sc.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology; Also published by the University of Indiana Linguistic Club.

Zue, V.W. and Laferriere, M. (1979) "Acoustic Study of Medial /t,d/ in American English," J. Acoust. Soc. Am., Vol. 66, No. 4, 1039-1050.

Zue, V.W. and Shattuck-Hufnagel S. (1980) "Palatalization of /s/ in American English: When is a /š/ not a /š/?" J. Acoust. Soc. Am., Vol. 67, S27.