Man-Machine Interaction Using Speech

Article *in* Advances in Computers · December 1971

DOI: 10.1016/S0065-2458(08)60632-4 · Source: DBLP

David Hill
The University of Calgary
58 PUBLICATIONS 490 CITATIONS

Man-Machine Interaction Using Speech

DAVID R. HILL[1]

Department of Mathematics, Statistics, and Computing Science
The University of Calgary, Calgary, Alberta, Canada

[1] Present address: Department of Electrical Engineering Science, University of Essex, Colchester, Essex, England.

1. Introduction

1.1 Nature of a Man-Machine Interface Using Speech

What is meant by "man-machine interface using speech"? One could talk instead of a "system-system interface using speech," and thereby avoid the slightly irrelevant question as to which is the man. An interface is a communication boundary between two systems. The essentially *human* character of a speech interface lies in the *nature of speech*. which is also where most of the problems arise. For communication, there must first be transmission of information. Information is characterized in terms of selective power among a set of messages, s the communicants must have message sets in common (which explain why the communication of "new ideas" is difficult). There mus also be some means of ensuring that the information received corre sponds to that which was transmitt 1. *Effective* communication is achieved when the receivers state following the message transmission is that desired by the sender. The receiver's state (of understanding) may have to be determined indirectly by observing the receiver's *behavior*. The sender, therefore, requires (in the absence of guarantees) feedback concerning the pertinent aspects of the receiver's behavior to allow modification of what is transmitted until the informatiom received achieves the effect required by the sender. Some basic step involved in man-man speech communication are indicated in Fig. 1.

It is clear, therefore, that a "system-system interface using speech" implies that *both* systems:

(1) are able to recognize and understand intelligible speech;
(2) are able to generate intelligible speech
(3) have a common corpus of experience—a common language.

One important aspect of speech is that the sole arbiter of what is intelligible speech is the human—preferably, for obvious reasons, in the form of a representative panel of his species. Machine recognition is "successful" if it makes a judgment agreeable to such a panel. Speech units have no absolute existence even within the same language. Speech units are functionally defined entities—two speech phenomena even though physically (acousticallv) dissimilar, are the "same" if they never distinguish two words in the language. A similar philosophy applies at higher levels of language.

1.2 Recognition and Synthesis

Just as the communication situation is symmetric, so there is symmetry in the processes of speech recognition and speech synthesis.

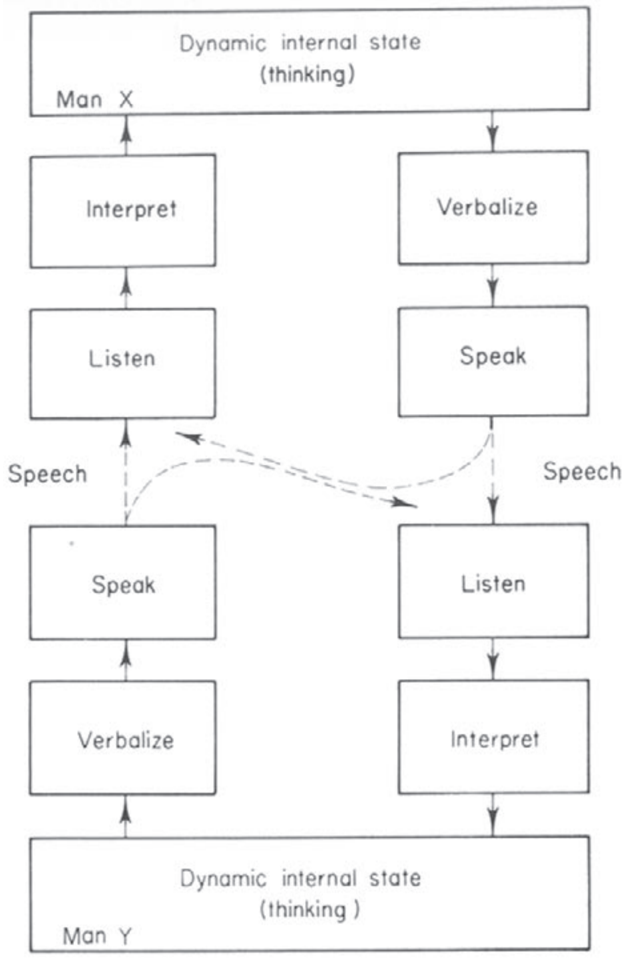

fig. 1. Steps for man-man speech communication

Figure 2 illustrates the parallel levels in both. In either case, it is necessary to *describe* speech adequately. This problem corresponds to the "problem of representation" noted by Feigenbaum [36]. He makes it clear that he considers "representation of data and data structures" to be a different problem to the "problem of representatiom *for problem solving systems*," though he indicates a belief that they have much of importance in common. For those working on speech recognition and synthesis by machine, the "representation of the problem to be solved" and the "data representation" are inextricably interwoven. Much essential research in speech has been concerned with exploration of suitable structures in which to frame worthwhile questions about speech. The problem of representation is the central problem in both speech recognition and speech synthesis in any sense There are two further problem areas in both recognition and synthesis—implementation and application. These are, perhaps, problems of technology and awareness.

Synthesis is a valuable aid to improving our representation of speech, since we can observe the inadequacy of the synthesized speech, and infer the inadequacy of our representation (see also Section 4.2)

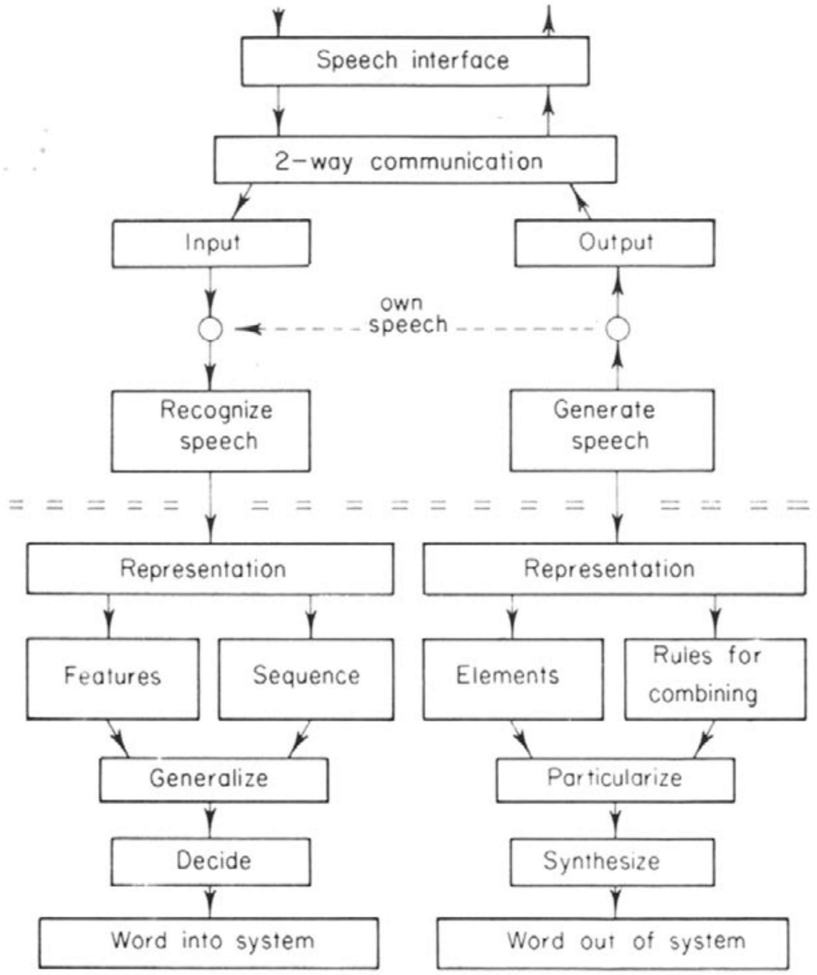

Fig. 2. Symmetry of the speech interface

In this way synthesis is closely related to recognition, though caution is necessary, for synthesis is *particular*, whereas an adequate representation of speech must be *general*. This is one reason for the advanced state of speech synthesis compared to the rudimentary state of speech recognition; a second reason lies in the efficiency of the human recognition process, which is able to compensate for some inadequacy in the machine generator. A good example of the "particular" nature of synthesis is the "identifiability" of both human and mechanical talkers. Iles at Edinburgh University [171] has adapted Holmes' [64] rules to suit PAT, the synthesizer invented by Lawrence [2]. The resulting synthesis is identifiable with Holmes' rules. and characteristic of PAT are clearly noticeable. Recognition of individual machine generated speech would seem to offer little more than a technological problem, but an individual speaker varies more than an individual machine. The principle factor that has underlain continued failure to "solve" the recognition problem has been an inability to give a generalized representation of speech. and the recognition process, which embraces such variation. The representation adopted even restricts what questions may be asked concerning speech.

2. Reasons for Requiring a Man-Machine Interface Using Speech

2.1 Introduction

The reader is entitled to ask why a man-machine interface using speech is so attractive. It cannot be denied that some interface is essential, so it is permissible to concentrate on the notion of using speech.

At the 1967 IEEE International Convention, during a panel discussion, J. C. R. Licklider is quoted by Lindgren [90] as saying:

> I am concerned that the assumption is made that people are going to continue to talk mainly with people. Whereas it seems to me, we will have in due course a dynamic, organized body of knowledge, not in anyone's brain, but in some vast machine or chemical thing and, for scientific and technical and intellectual purposes, we'll want to talk mainly with it. Talking among people might be mainly for positive reinforcement of one another, although I suspect that even that could he handled better through some organized reinforcement arrangement.

Licklider speaks with authority, and is among those concerned with making the facts of the future. Even his last remark is fair as may be judged by the fact that some intelligent people could not be convinced that Weizenbaum's [146] primitive automatic psychiatrist—ELIZA—was only a suitably programmed computer. Thus ELIZA passes Turing's test [152] in this restricted context.

Thus the first reason for requiring a speech interface, as opposed to any other, is based in the psychology of the human. It is natural, convenient, and indeed, almost expected of any system worth communicating with.

2.2 Advantages

There are, however, a number of very practical advantages. Speaking for the nontechnical man John Lotz (quoted by Lindgren from the same discussion) notes the following advantages of speech communication:

(1) No tool is required;
(2) It can be used in the dark as well as it can in the light;
(3) It does not require a free line, of sight;
(4) It can he varied from a (confidential) whisper to a loud (long distance) shout;
(5) It requires very little energy;
(6) It is omni-directional; and
(7) It leaves the body free for other activities.

The last two are, perhaps, the most important advantages from the purely practical point of view. Lea [82] working for NASA translates these advantages into concrete terms for us. After noting, with Licklider, that the use of speech is rooted deep in man's psychological makeup, he notes some general advantages of speech communication and also some advantages of special significance to astronauts. A speech interface (where numbers following cross-reference prior advantages):

(8) Offers more natural communication;
(9) Allows physical mobility during operation (6, 7);
(10) Increases the communication capacity, by exploiting multi-modal communication [he notes that speaking is, in any case, much faster than typing];
(11) Gives additional reliability-since channel failure is less a problem when there are many channels;
(12) Is very suitable for "alert" or "break-in" messages (6);
(13) Allows communication without specialized training (1, 8);
(14) Is essential, when hands and eyes are both occupied, as they often are in busy space missions (7);
(15) May allow physiological/psychological monitoring of the state of the operator, at least as a back up to electrical sensing;
(16) May be less affected by acceleration than other channels [there have been no reports of hearing loss during acceleration].

He notes further considerations of special relevance to astronauts:

(17) In space, the environmental noise is under perfect control [though he admits it may still be noisy] because no noise can come in from "outside."
(18) In a spacecraft. the speaker population is very restricted.
(19) Voice control allows machinery inside the spacecraft to be operated from outside without the need to carry around complex communicators

We may add to these as follows. A speech interface is extremely compact, the speech interface need take up little or no panel space. Shortage of panel space has been a problem in some situations, and especially so in small space capsules, with so many demands for what little space is available. With voice control of machine functions it would be easy to monitor control sequences. Only a radio transmitter and receiver would be required. When speech is a normal means of communication (as in Air Traffic Control) the communication load on the operator can be reduced if the machine is able to accept input in the form in which it is normally generated. In currently planned ATC schemes the controllers must talk to the aircraft and store the same information in a local computer. Even when ATC is largely compute controlled, this will remain true in a significant sense. Speech also offers special advantages for disabled persons (other than those with speech or hearing difficulties), especially for the blind. The transmission of information by speech is compatible with the cheap and widely available telephone system.

The possibility of "voice dialing" is an additional attraction. Finally, using voice-printing, or similar techniques, identity checks may be performed on the speaker—a potentially valuable security measure. These advantages may be summarized. A speech interface:

(20) Takes up little or no panel space;
(21) Allows simple monitoring of command sequences by remote personnel ;
(22) Can allow a single channel to serve more than one purpose (e.g., communication with aircraft at the same time as data entry to a computer);
(23) Allows blind or disabled persons to operate machines easily ;
(24) Is compatible with the widely available and inexpensive telephone system;
(25) Allows security checking.

To realize all these advantages fully requires better performance than is offered by any equipment now in sight.

2.3 Disadvantages

Lea notes two disadvantages. First, a message from the computer leaves no permanent trace in voice form—but (with a proper interface) presumably the astronaut could ask for a repeat of the message if necessary; from point of view of "the record" all conversation is recorded anyway. Second, if machines are used which are excessively limited (e.g., in terms of vocabulary, or acceptable "accent"), they may he so aggravating to converse with that they will be unusable. We may add to these as follows. A speech input may be subject to competing inputs—a special case of the noise problem. The recipient of spoken machine output is compelled to progress at the rate dictated by the machine; he may not scan a page, taking note only of the odd sentence, nor reread parts easily. Special care must be taken to verify key messages; a mechanical device, unless it fails, will produce a definite signal a spoken word, even for the most perfect machine and under completely "noise-free" conditions, may be ambiguous. The problem of ambiguity can be alleviated by careful choice of message structure and vocabulary, and such precautions are normally taken, by the military, even for man-man communication using speech. Thus we may summarize the disadvantages of a speech interface:

(1) It is rather susceptible to various sorts of environmental noise:
(2) It could prove aggravating, and therefore unusable, if too far short o f perfection;
(3) It is transitory, arid requires additional complexity of equipment to allow recording or repetition;
(4) Verification of some messages, prior to execution, might be essential;
(5) A speech output may not be scanned easily.

3. An Outline of Speech Production, Perception, and Some Related Topics

3.1 Introduction

Speech forms the visible tip of the vastly complex iceberg that is language. Although speech is physically an acoustic phenomenon it is a gross simplification to consider speech production and perception solely in terms of the speech (acoustic) waveform. Both speech production and speech perception depend heavily on all manner of constraints that form part of the total language process. It would he quite outside the scope of this article. however, to do more than outline the bare essentials of speech production and perception, as far as possible at the acoustic level. It is true that it is not possible to avoid, entirely, some of the higher level processes. We may restrict discussion of production to purely physical terms; perception, on the other hand, is essentially subjective and depends on too many nonphysical circumstances. It is generally accepted that the acoustic waveform does not always contain enough information to he unambiguously identifiable on acoustic criteria alone. A good illustration of the kind of difficulty is the fact that a person listening to a language he or she does not understand is appalled not because the words cannot be understood, but because very little word structure is even apparent. Though the process of perception depends at least in part on the acoustic waveform, either "now" or as experienced in the past, it depends on other factors within the province of the psycliolinguist. the neurophysiologist, the physiologist, the anatomist. and the psychologist. We know less about such factors than we should like to. Even the manner in which we use our past experience of the acoustic waveform is subject to uncertainty of a similar kind, and there is but a tenuous link between the physical properties of the acoustic waveform, and the perception of a sound.

The problems of syntax and semantics, and the problem of how these mediate the production and perception of speech involve formidable new kinds of problems. for which it would seem we require a new insight before we even start to tackle them. Sager [26] has briefly reviewed sonic approaches to the analysis of syntactic structure, and has presented details of one particular approach. Further significant progress in the way of a practical procedure for syntactic analysis has been made by Thorne and his colleagues [148]. Chomsky and Halle have recently published an "interim" report [17] on their work, aimed at the unification of grammar and phonology, on the basis of a transformational model (originated by Chomsky [16]). Automatic translation has proceeded, not too successfully, at the syntactic level, because our ability to handle the problems at the semantic level is so restricted. Work at the semantic level has tended to be restricted to word association networks ([44, 75, 76], for example), and information retrieval—especially question-answering programs [16. 170] and "Advice Taker" type programs ([91]. for example). Borko [9] and also Salton [129, 130] provide more recent relevant material. Miller [103], Cherry [14] and Broadbent [10] have provided valuable overviews of the communication process. Meetham and Hudson's Encyclopaedia [102] provides useful snapshot information on topics in speech and linguistics. The MITRE' syntax analyzer [163, 164] is important and there is a forthcoming book by Foster [41] on automatic syntactic analysis. Simmons *et al.* [138] have considered the problems of modeling human understanding and provide further references. Finally references [4, 38, 55, 98] cover aspects of language analysis and semantics. All of this must be declared to be beyond the scope of this article.

Even while confining our attention, in this section, to the acoustic level, much that could be said must be left unsaid. Flanagan's book [37] provides excellent, detailed coverage of the hard-core of acoustically oriented speech research in the three areas of analysis, synthesis, and perception. Fant [33] is concerned with the acoustics of the vocal tract, and his book is a classic, though unfortunately (at time of writing) out of print. Both these books provide excellent bibliographies. The literature relevant to speech research is widely scattered among electrical engineering, psychology, linguistics, speech and hearing research, physiology, neurophysiologv, physics, acoustics, audio-engineering, and a variety of specialist books and journals. Good bibliographies are difficult to compile, and rarely complete. A broad spectrum of papers numbering around 500, and covering up to 1965/1966, appear in abstracted form [57] but the selection is somewhat inconsistent, and is heavily biased towards automatic speech recognition. Stevens' *Handbook of Experimental Psychology* [142] is a valuable reference for many of those areas which are of indirect but vital concern to speech research and, in particular, includes papers on "Speech and Language" by Miller, "Basic Correlates of the Auditory Stimulus" by Licklider, "The Perception of Speech" by Licklider and Miller, "The Mechanical Properties of the Ear" by Bekesy and Rosenblith, and the "Psycho-Physiology of Hearing and Deafness" by Hallowell Davis. A large body of spectrographic data on speech sounds appears in Potter et al. [118], recently reprinted from a 1948 edition.

Speech communication is surely one of the most interdisciplinary subjects imaginable. This fact creates one of the more formidable problems for those wishing to work in this area, for progress in understanding the process of speech communication will depend on the combined use of skills from many disciplines.

3.2 Speech Production

The human vocal apparatus is depicted, diagrammatically, in Fig. 3. The main constituents are the larynx—in which are the vocal folds (also, mistakenly, called “cords”); the tube connecting the larynx to the mouth—comprising the pharynx and the oral cavity; the nasal cavity; the velum—a valve which may open to connect the nesal cavity in parallel with the oral cavity; the teeth; the tongue; and the lips. In the most relaxed articulatory state, and with the velum closed, the larynx, together with the tube to the mouth orifice, may act somewhat like an organ pipe. Forcing air through the glottis (the elongated “slit” between the two vocal folds within the larynx) causes forced vibration of the vocal folds, and provides a harmonic.-rich quasi-periodic source of excitation which sets up standing wave: within the vocal tract, as illustrated in Fig. 4. The frequency of thcs standing waves, or resonances, under these circumstances, depend:

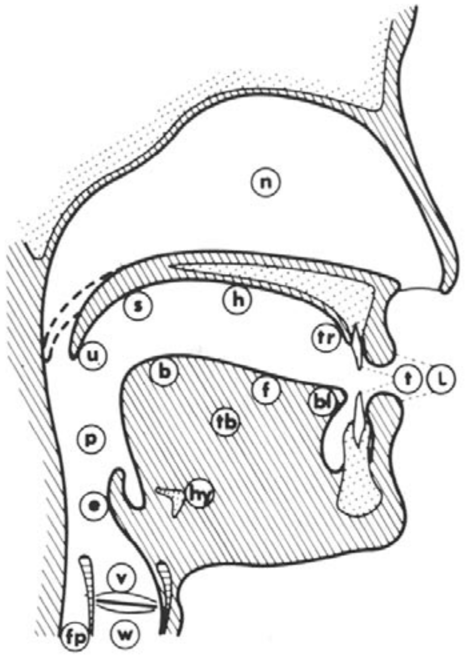

Fig. 3. Schematic diagram of the human vocal apparatus: b, back of tongue; bl, blade of tongue; e, epiglottis; f, front of tongue; fp, food passage; h, hard palate; hy, hyoid bone; l, lips; n, nasal cavity; p, pharynx; s, soft palate or velum; t, teeth; tb, tongue body; tr, teeth ridge or alveolar ridge; u, uvula; v, vocal folds (and glottis); v, windpipe (trachea).

principally on the length of the tube from the larynx to the lips. The length is about 17 cm for the adult male, and leads to resonances of about 500 Hz, 1500 Hz, 2500 Hz, etc. The energy supplied by the *source* (carried as a modulation of the airflow through the glottal orifice) is concentrated into particular frequency regions by the *filter effect* of the vocal tract, and it is worth considering speech production in term of the so-called *source-filter* model [33]. The Fourier spectrum of typical glottal volume-velocity waveform (which is, roughly, a train of triangular pulses) falls off with increasing frequency at about 12 dB/octave [37]. The behavior of the vocal tract may be regarded as analogous to the behavior of an electrical network in which current represents volume flow, and voltage represents pressure. There is, in fact, a quantitative correspondence: the resonant behavior of the appropriate electrical network, which is based on the assumption that the irregularly shaped vocal tract may be approximated by the electrical equivalent of many short cylindrical sections of varying diameters, parallels that of the vocal tract. To the extent that the model is valid (and there is considerable agreement between model and reality [33]) one can investigate the transmission properties of the tract analytically. It is possible to treat the source and filter in terms of network theory to derive expressions representing the output from the vocal tract in terms of the excitation function and the transfer function of the tract [33, 37].

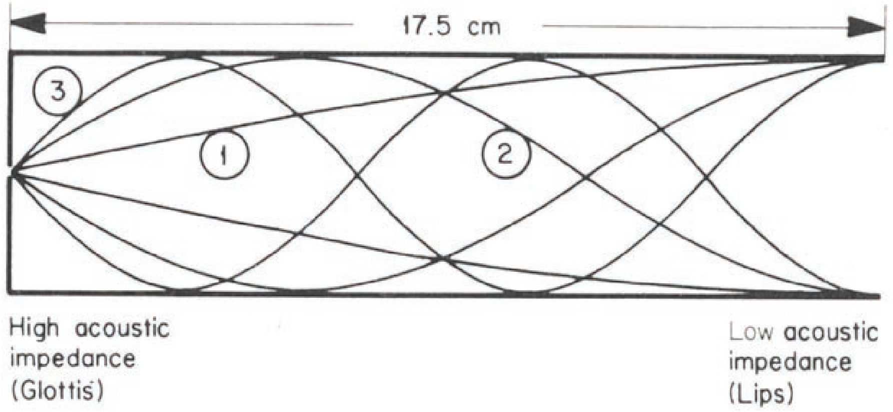
17.5 cm
3
1
2
High acoustic impedance (Glottis)
Low acoustic impedance (Lips)

Fig. 4. Standing waves in a uniform tube closed at one end

The spectral peaks (in a frequency analysis) of the sound radiated from the vocal system are termed *formants*. For most practical purposes formant frequencies may be considered the same as the resonant frequencies of the vocal tract [33]. Schroeder [134] expresses the distinction between the representation of speech in terms of vocal tract transfer function and excitation function on the one hand, and the representation in terms of spectral envelope and spectral fine structure (of a frequency analysis of the radiated sound) on the other, very clearly.

In the neutral articulation considered above, therefore, the lowest formant, called F1, had a frequency of 500 Hz. The *frequency* of the *n*th formant is conventionally represented as F_n, thus, for the neutral unstressed vowel under consideration, F_2 = 1500 Hz and F_3 = 2500 Hz. Such a neutral unstressed vowel occurs in English as the first vowel in "above." The formants higher than F_3 are less important for intelligibility of speech, though they affect speech quality, and give clues as to speaker identity. Even F_3 is important for only a few sounds, such as the "r" in "three" in General American.

If the vocal tract is constricted at some point, then the resonant properties are modified. and the formant frequencies change, leading to perception, by a listener, of vowels other than the neutral vowel. A very simple view, considering a given formant, is that a constriction at a velocity maximum for that formant, within the tract, decreases tlu formant frequency, while a constriction at a pressure maximum increases the formant frequency [33] The constriction may be produced by a tongue "hump" raised at some point anywhere from the front to the back of the oral cavity, the height of rise being anywhere from high (maximum constriction consistent with production of a vowel sound) to low. Vowels are frequently classified in terms of height and place of tongue hump constriction. Figure 5 shows four relatively extreme vowel articulations.

Besides the vowels, produced as a continuous sound with a quasi periodic source (voicing or phonation) and a relatively unobstructed vocal tract, other sounds may he produced. Some depend on excitation by incoherent noise, produced at a severe constriction in the tract with or without accompanying voicing. These are termed *voiced or unvoiced fricatives* and, for English, include, for example, /v,z,f,s/. Such sounds are characterized by the random source, and the amplitude and frequency of the main spectral peak(s). Strevens [144] has published spectral data on English fricatives. In the case where noise is produced at the partly closed glottal orifice (instead of forcing the folds into vibration) it is termed *aspiration*. Aspiration is used as the source, in place of voicing, during whispered speech. The output is characterized by the random source, and formant patterning. In "breathy voice" both voicing and aspiration may be present .

Another class of sounds, the *stop consonants*, result from complete closure of the vocal tract, Those sounds may or may not be accompanied by various phenomena. For example, phonation during closure (in voiced stops); the release of overpressure (resulting in transient excitation of the vocal tract); aspiration (voiceless stops in particular); or frication (production of incoherent noise) at the point of closure, as the closure is released. As there is relative silence during stop sounds they are characterized almost entirely by the

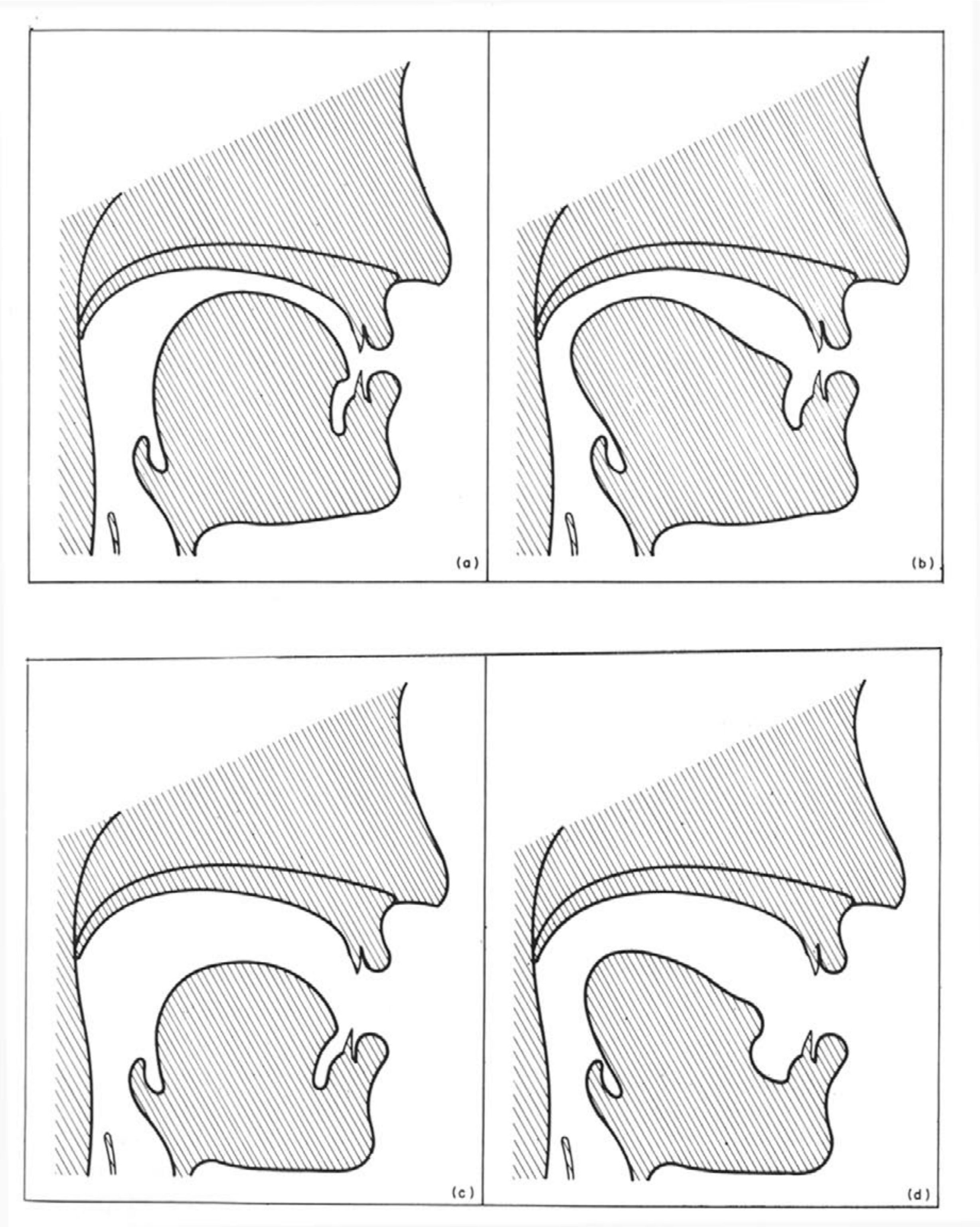

Fig. 5. Four relatively extreme vowel articulations: (a) high front vowel (similar to that in "heed"); (b) high back vowel (similar to that in "who'd"): (c) low front vowel (similar to that in "had"); (d) low back vowel (similar to that in "hard").

transient phenomena that are associated with them, including the rapid changes of formant frequencies preceding and/or following the silent interval. These formant transitions are important acoustic phenomena associated with many consonants, and are more rapid than the changes of formant frequency associated with, say, diphthongs, where a glide from one vowel configuration to another takes place. Voiced stops are distinguished from unvoiced stops, in the main, by the relative timing of phonation and closure. For example, an initial unvoiced stop such as /p/ in "pat" is distinguished from an initial voiced stop such as /b/ in "bat" mainly by the fact that in the latter case voicing starts before, or virtually coincident with the release of closure, whereas in the former case the start of voicing occurs some time after release of the closure, and the first part of the associated formant transitions are superimposed on aspiration noise.

Nasals constitute a further class of speech sounds. These depend on coupling the nasal cavity to the vocal tract by opening the velum. Additional formants appear in the radiated sound, and there is a reduction in the amplitude of the higher frequency components. If the oral cavity is relatively unobstructed, nasalized vowels are produced. If the oral cavity is closed, nasal consonants are produced; for example, in English, /m n/.

Both stops and nasal consonants are commonly classified according to the *place* of closure of the vocal tract—in English, velar, alveolar, and bilabial, for example. Other sounds are (similarly) classified according to the *place* of maximum constriction of the tract. An orthogonal classification is according to *manner* of articulation—"how" as opposed to "where" the sound is articulated—e.g., voiced stop, and nasal.

Affricate sounds are somewhat like the stops, but involve a slower separation of the articulators during release, with a greatly extended period of frication. They are commonly represented by the symbol for the stop sound followed by the symbol for the appropriate fricative. The initial sound in English "chip" is an unvoiced palato-alveolar affricate, represented /tʃ/.

Finally, there are the four *glides*, the sounds /w/ as in "we," /j/ as the first sound in "you," /r/ as in "real," and /1/ as in "let." The sounds are vowel-like, in that they are characterized by formant values and are continuant. However, they are characterized also by the change of formant values, not unlike the diphthongs, together with a degree or type of constriction that places them away from vowels.

3.3 Phonemes

The number of distinguishably different sounds that may be produced by the human vocal apparatus, though not unlimited (due to thc built-in constraints of the vocal apparatus) is very large. Nevertheless these sounds may be grouped, for a given language, into a comparatively small number of categories—*phoneme* categories. Two sounds are placed within the same phoneme category if they never form the basis of a distinction between two words in the language. In representing the spoken form of a language, various marks and letters may be used to represent the various sounds. Each symbol may represent a phoneme, in which case the transcription is "broad" (*phonemic*) and the symbols are usually enclosed within virgules (/ ... /); or the precision of the representation may be greater, so that a "narrow" (*phonetic*) transcription results, and the symbols are usually enclosed within square brackets. The greater precision of the narrow transcription allows different sounds from within the same phoneme category to he represented differently, and such a transcription is non-phonemic. The different sounds falling in one phoneme category are termed allophones of the phoneme. It is important to realize that the acoustic variation between different allophones of a given phoneme may be very great indeed. For doubters who need to be convinced, the section describing the detailed realization of the /r/ phoneme in R.P. English in Daniel Jones' classic [73] is worth reading. The whole book is an excellent, detailed reference , which will reveal some of the gross simplification involved in the outline of some speech sounds of English given in Section 3.2.

For a known language, since the different conditions under which different allophonic variations of a particular phoneme occur are known, the broad, or phonemic transcription usually suffices to transcribe the spoken utterance into a written equivalent, and native speakers are usually unaware of the extent of the variation.

Phonemes may be regarded as speech segments or units only in a descriptive sense, and considerable training is required to learn to "hear" the sounds of speech. In addition to segmental descriptors of speech such as phonetic value and stress, there are supra-segmental characteristics—those that extend over more than one segment—and these may affect the meaning of an utterance radically. Thus the intonation pattern, or rise and fall of pitch (voicing frequency) must be controlled as a vital part of the speech production process.

3.4 Speech Perception

The natural medium of speech is the acoustic waveform. Speech is first an acoustic phenomenon, acting on the auditory system comprising the pinna (the externally visible part of the ear); the external canal or meatus; the ear drum which vibrates three small bones or ossicles in the middle ear; the cochlea—a coiled chamber divided for almost its entire length into halves by an elastic partition, each half being filled with fluid; and a system of nerves, leading ultimately to the cortex, in particular the auditory cortex.

A schematic diagram of the uncoiled cochlea is shown in Fig. 6. The elastic partition is shown as a single wall, though it actually contains a duct comprising additional components-including the basilar membrane and the important "organ of Corti" in which are found hair cells, the ultimate end organs of hearing. These hair cells generate nerve pulses in response to disturbance of the basilar membrane.

Acoustic vibration is transformed into movement of the perilymph in the cochlea by the ossicles acting upon the oval window. At very low frequencies (below 20 Hz) this movement is a to and fro movement of fluid between the oval and round windows. Higher frequencies act to produce displacement of the basilar membrane so that the point of maximum displacement varies with frequency. The highest frequencies are quickly dispersed, and produce maximum displacement towards the oval window end of the membrane. The lowest frequencies produce the greatest effect towards the helicotrema. Thus the response of the basilar membrane and organ of Corti to incoming sound, in terms of frequency and phase, is rather like a continuum of overlapping bandpass filters of comparatively broad band-width. Frequency resolution is best at the low-frequency end, the "Q" being roughly constant [37]. Signals related to the displacement of the basilar membrane over its entire length are transmitted to the brain, which may be regarded as the seat of perception.

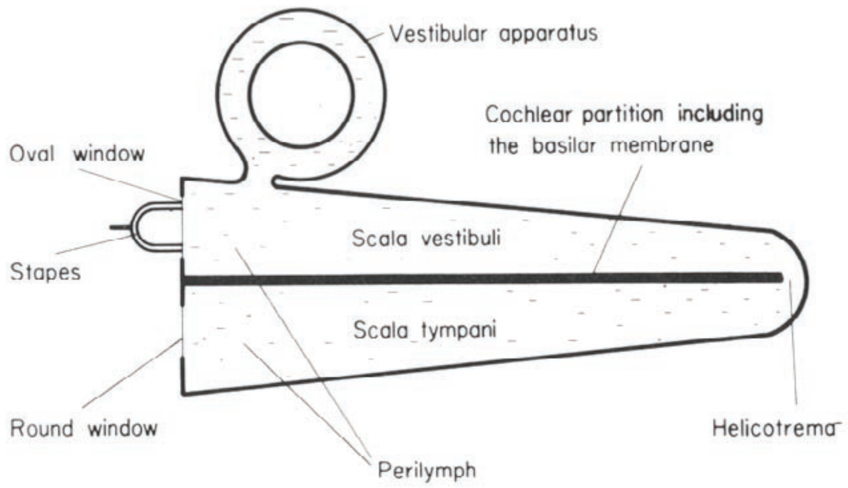

FIG. 6. Schematic diagram of thc uncoiled cochlea.

Interest in this system—from the point of view of perception—may take two forms. The classical approach of psychophysics may be used to determine the accuracy or resolution of the system in the discrimination of elementary sounds. On the other hand, the perception of speech, per se, may be investigated, for there is considerable evidence to support the view that the

perception of multidimensional, temporally complex speech sounds is of a different character to the naive perception of simple sounds. Speech perception involves learning, and judgments are categorical. A tone frequency may be perceived a lying "somewhere" in between two reference tone frequencies. In the "speech mode" however, a sound will be heard as either this referenc sound, or that reference sound. Stevens [140] has tentatively concluded that even vowels, in context, are subject to similar categorical judgments. This is at variance with earlier belief based on experiment with isolated vowels, which had suggested vowels were different from other speech sounds in this respect. There is a large body of knowledg derived from the psychophysical approach, but we know much les about speech perception. Most of the work in this latter category has consisted of identification experiments.

It is possible to generate synthetic speech in a variety of way (see Section 4.2). Using synthetic speech, it is possible to produce systematic variation in cues believed important for certain speech discrimination judgments, and use these in psychophysical experiments to find out how perception changes. This was the approach adopted by workers at the Haskins Laboratory in New York and their work has given us considerable insight into the nature of speech cues, and experimental methods for investigating speech.

A summary of some of the Haskins' results [85] indicates the importance of the spectral properties of noise produced at constrictions: the importance of the formant transitions in the perception of fricatives. affricates, and stops; and shows the close relation between nasal and stop sounds, which may be classified according to place of articulation and manner (see above, Section 3.2). Lisker and Abramson [92] have shown the importance of relative time of onset of voicing in the perception of stops. Acoustic cues for the perception of /w, j, r, l/ have been investigated [108]. Important work has been carried out at other centers: M.I.T. Research Laboratory of Electronics, for example; and in Fant's Laboratory at the Royal Institute of Technology in Stockholm. Both these places produce quarterly reports of their work, in addition to papers in the literature. To attempt a comprehensive list of published work would be foolish, and do grave injustice to the work which would, of necessity, be omitted. Not all work depends on synthetic stimuli. Analysis of speech is also important. For example, Strevens, [144] data on fricatives has already been mentioned; Wells, at University College, London, has published data on the formants of pure vowels [166]; and Lehiste has published data on acoustical characteristics of selected English consonants [83]. It is, perhaps, worth mentioning a few other centers where the experimental study of speech is a primary concern. Among these are Bell Laboratories, Speech

Communication Research Laboratory (Santa Barbara, California), the Department of Linguistics, at University of California at Los Angeles, the Air Force Cambridge Research Laboratory (Bedford, Massaeliusetts), the Department of Phonetics and Linguistics of Edinburgh University, the Institute for Perception Research (Soesterberg, Netherlands), and the Department of Phonetics of the University of Bonn (Germany). That work carried out in industry generally tends to be directed more at the solution of particular engineering problems for automatic speech recognition, speech synthesis, or analysis-synthesis speech communication systems (vocoders, see Section 3.5).

It is exceedingly difficult to give an adequate summary in such a complex ongoing field as speech perception. It is perhaps best to leave the subject, while re-emphasizing that further information may be obtained from Flanagan's book [37], from Meetham's Encyclopaedia [102], and the literature. We return to the subject, from a different viewpoint, in Section 5. There is, as yet, no generally accepted theory of speech perception, though there is a growing feeling that it is all active, constructive process—learned during early childhood on the basis of more primitive perceptual ability. Knowledge is still fragmented among a number of disciplines. Many of the unanswered questions in speech perception constitute unsolved problems in automatic speech recognition. The dominating question at the present time seems to be concerned with the means whereby such widely varying stimuli as different allophones rendered by different people can be perceived as "the same sound" while fine distinctions can be drawn in other cases. This may be regarded as the question "How can one adequately characterize phonemes?"—though this implies that phonemes are the "units of speech" in some basic sense. If the implication is accepted, then the question arises as to whether phonemes are recognized by segmenting the continuous acoustic waveform or not. Stevens has described the segmentation problem as "the problem that you can't." [See also Section 5.2.2(b).]

The problems of automatic speech recognition (ASR) are considered in Section 5. However, it should be noted that *an ASR machine is an analog of the human perceptual mechanism. Insofar as such a machine models the human process, it can be a test of hypotheses about the human perceptual process for speech*. This author has never seen this given as a reason for ASR research, but believes that it is an important reason, and one which should be given greater recognition.

3.5 Vocoders

Vocoders are important, in the context of this section, for the impetus they have provided to speech research. A vocoder was originally conceived as a device to reduce the amount of information required in the transmission

of speech messages over telephone circuits, and later (as radio frequency channels became filled) over radio channels. Vocoders are also valuable as a means of secure transmission of information for, say, military purposes. In information theory terms about 40,000 bits per second are required for transmission of the 300 Hz to 3.4 kHz bandwidth speech signal common in telephone circuits. High fidelity speech would require 100,000 hits per second, or more. A generous estimate of the amount of information contained in speech, based on the frequencies and rates of phoneme production, gives a figure of about 50 bits per second. The huge diserepancy suggests that a more efficient coding of speech would allow very many more conversations to be carried on a given telephone circuit, with consequent saving in cost. Of course, this degree of reduction (to only 50 bits per second) would convey only the message. Some additiona information would be required to transmit speaker-dependent cues The original VOCODER (derived from VOice CODER), invented by Homer Dudley, was designed to achieve a significant reduction in the amount of information transmitted. Incoming speech was analyzed in terms of the amount of energy present at various frequencies (to derive the spectral envelope) and the pitch or form of excitation (to determine the spectral fine structure). This information could be transmitted over about one tenth the bandwidth required for full speech and then reconstituted at the other end. Such a scheme is shown, diagrammatically, as Fig. 7. Because speech was analyzed into different frequency channels (typically fifteen) the device has been described as a "channel vocoder." The chief difficulty with the channel voeoder is in analyzing for "source," that is to say, in deciding whether

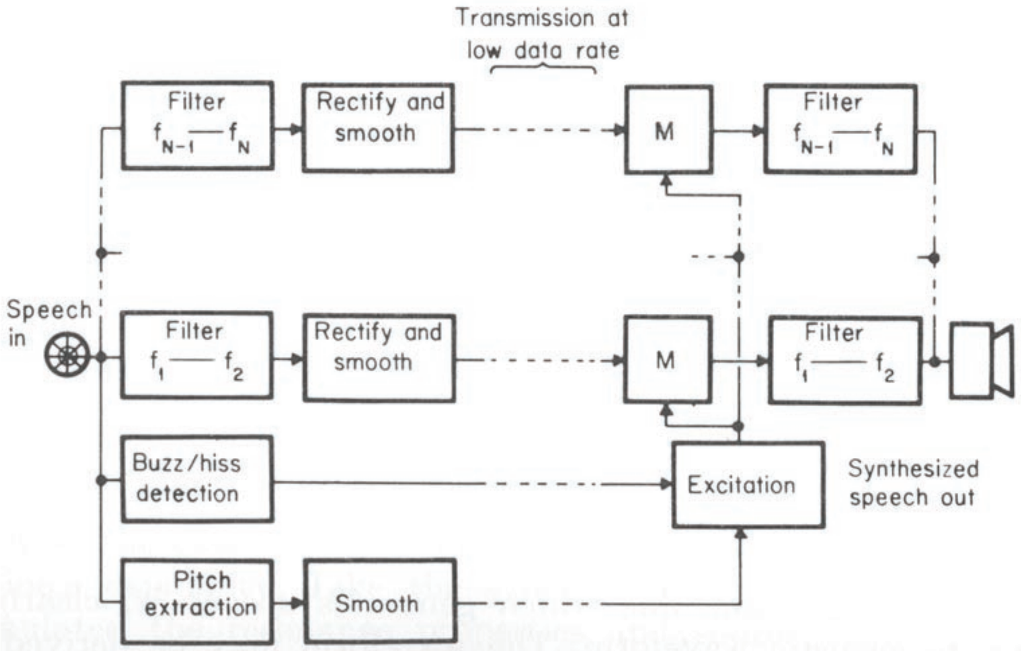

FIG. 7. Block schematic of a channel vocoder

the source is voiced or not, and, if so, what the pitch frequency is. Schroeder [134] gives an excellent review of vocodcrs, and of attempts to solve the "pitch extraction" problem. Suffice it to say here that another type of vocoder—the resonance vocodcr—was developed on the basis of a resonance model of speech. The channel vocoder was based on what may be termed an acoustic model of speech [see also Section 4.2.2(b)]; i.e..,that speech may be regarded as a time series of spectra having defined shape and fine structure. The resonance model assumes knowledge of the constraints implied in the production process at the level of the vocal tract within the domain of an acoustic theory. Analysis is typically performed, therefore, in terms of the first three formant frequencies, the amplitude and frequency of voicing, the amplitude of aspiration, and the amplitude and frequency of the main incoherent noise spectral peak. The values of these eight parameters may be transmitted over quite a narrow band, and then reconstituted by means of a parametric synthesizer, e.g. [2, 81] [also see below. Section 4.2.2(b)]. Such a synthesizer, constituting the synthesizing half of a resonance (or formant) vocoder, is also commonly called a parametric synthesizer. although in theory synthesizers based on other parameters (e.g. physiological parameters) could be developed. The main problem, apart from "pitch extraction," is "formant tracking." The parametric synthesizer is important because it is simple to control, even compared to a channel synthesizer, and it reflects the speech production process much more directly. However, both are of value in producing talk-back from a machine, as well in vocoder systems.

4. Talk-Back

4.1 Introduction

A machine that "talks back" must he equipped with some equivalent of the human speech production mechanism. The degree of sophistication may be slight, as when, parrot-like, the machine simply regurgitates a previously recorded message, when it is deemed appropriate; or the degree of sophistication may be such as to allow unlimited spoken output, composed—without special methods of any kind—by users, or even by the machine itself, and output by the vocal equivalent of a line-printer or graphical display. This section examines and discusses some of the practical methods within this range.

4.2 Methods of Generating Speech by Machine

4.2.1 Recording

The output of a machine which generates speech is, clearly, an appropriate acoustic waveform. The waveform mav be derived in a number of ways. The simplest method is to record a magnetic or other physical representation of the actual waveform of a desired utterance, and use this to recreate the acoustic waveform when required. This can be thought of as a direct representation. The physical representation need not be continuous, it may be discrete, and consist of amplitude samples (to the required accuracy) taken at a rate of at least twice the highest frequency component present in the waveform. If the samples are digitized, then such a discrete representation of the original analog waveform may be stored in conventional computer memory locations. Assuming 6-bit accuracy of amplitude samples, and a highest frequency of 3.4 kHz (to give telephone quality speech) leads to a storage requirement of about 40 kilobits for each second of speech. High-fidelity speech would require in excess of 100 kilobits of storage for each second of speech. Messages could be retrieved easily and replayed using a standard digital-to-analog converter into an audio-amplifier/loudspeaker system.

Conventional audio-recording techniques, such as magnetic drums, records, audio tape, or wire recorders, may be used to store and replay messages in the continuous direct representation. Such apparatus would constitute special peripheral equipment, and there could be problems of access time, and message synchronization. There is an interesting third alternative. Wilishaw and Longuett-Higgins [168] have described a device called the holophone, which is analogous to the holograph, but operates in time rather than space. Such a device may store many waveforms, and if part of a stored waveform is played into the device, the continuation of the waveform emerges. It would be possible to start each waveform with a unique sound, the sound being one the computer could generate without special equipment, as a pulse stream perhaps. Messages could then be retrieved simply and precisely when required.

4.2.2 Synthesizing

(a) General The alternative method of producing the required waveform for all utterance involves synthesizing the wayeform using suitable ingredients and an appropriate recipe. The recipe is the set of rules implied in a particular way of describing, or modeling, speech. The ingredients are the particular values assumed by the units or parameters required for the model, in order to produce the required utterance. Von Kempelen, the ubiquitous Baron, is generally credited with the first speaking machine of the synthesizer type [161]. It could, with the aid of a human operator, produce (synthesize) all Latin, French, and Italian words, except those containing about five sounds beyond the machine's capability. Like the synthesizer of a resonance vocoder, it simulated the resonance properties and source characteristics of human vocalization, but it did so mechanically. Another machine of

note, which also required a human operator, was Dudley's VODER—demonstrated at the World's Fair in New York in 1939 [27]. Trained operators (whose training took a year or more) could play this machine to produce intelligible speech. Like von Kempelen's machine, it was a resonance analog of the vocal apparatus, but was electrical rather than mechanical. Synthesizers have been many and varied. Dunn and Barney [29] provide a comprehensive review of synthesizers up to 1958, Flanagan [37] is again a most excellent reference. He devotes some 125 pages to a comprehensive treatment of speech synthesis, and the related topic of bandwidth reduction, tracing it from its earliest history. Much of the work on synthesizers has been aimed at improving speech communication and at attaining a better understanding of speech, the synthesizer providing a flexible medium for testing hypotheses. and gathering data on the cues for different perceptual effects. With the advent of computers, which are able to manipulate the information needed to control synthesizers, they have assumed a special importance as a means of speech output from machines.

(b) Models and Synthesis The very simplest model of speech assumes a series of concatenated units (phonemes?) and attempts to synthesize new utterances by cutting and splicing the trace from a direct representation of other utterances. This model is too simple. principally because of the variability of speech and the effect of interaction between neighboring speech sounds, both of which lead to poor matching at the joins, unnatural effects, conflicting cues, and the like. Nevertheless, useful results have been obtained using this model in experiments on speech perception ([47, 51, 54, 124], for example). To account for the many interaction effects between phonemes, segments which are not phonemic are usually chosen. Wang and Peterson [165] suggest that as many as 8500 segments may be required. A bandwidth compression scheme was initiated at RCA Laboratory [110. 111] which identified segments at the analyzing end, and transmitted only codes. Synthesis was then achieved by retrieving prerecorded segments from a drum store. There were problems of timing and in dealing with suprasegmental features. More recently, a highly successful method of segmental synthesis has been developed at IBM [26, 30] by Dixon and Maxey, whose demonstration of synthetic speech was the most impressive that was heard at the IEEE/AFCRL Boston meeting in 1967. Matching (of amplitude envelope. formant frequencies, etc.) at segment boundaries is achieved by using a parametric (resonance analog) synthesizer to produce the segments. The segments themselves are carefully chosen to have boundaries where interaction effects. are least, and they carefully adjust their segments on the basis of repeated listening trials with naive listeners. This operation is made easier because the required parameter tracks are stored in discrete form in computer memory. They note, in this connection, the impossibility of one operator listening critically to very much synthetic speech, and part of their success undoubtedly stems from recognition of this fact, and the use of relays of fresh naive listeners, to highlight the inadequacies of their intermediate attempts. They estimate 1000 segments to be the basic minimum for highly stylized speech (as compared to about 40 phonemes required for a "synthesis-by-rule" approach, see below). The weakness of the scheme, as with any scheme producing speech by concatenating prerecorded segments—no matter how sophisticated the segments—is an inability to handle suprasegmental features economically. To produce real flexibility in these suprasegmental features would multiply a prerecorded segment inventory by a large number. The alternative is to handle the suprasegmental variation at the time the speech is synthesized, which would require an on-line parametric synthesizer (see below). Speech synthesis by segment assembly, of segments produced synthetically, is a two-tiered synthesis, in which the first stage could be replaced by the one-time use of a highly skilled, obedient, and completely standardized human being, if one existed. This fact is emphasized purely to clarify the distinction from approaches described in the next part of this section. The parametric synthesizer involved for two-tiered synthesis is used, in a sense, merely as a convenience.

Other forms of synthetic speech require a synthesizer which models more general aspects of speech, or speech production, and in this way embodies enough of the constraints or rules of speech (the recipe for synthesis) to simplify the ingredients (units or parameters). Ladefoged [77] has distinguished four approaches to speech synthesis: the use of an acoustic analog, the use of acoustic parameters, the use of a physiological analog, the use of physiological parameters. Each requires control information, suited to the form of the implied model.

The acoustic analog, which is the least restrained, requires the most arbitrary kind of control information. It models the acoustic domain of speech in terms of time-frequency-energy distribution. The input may be thought of as a two-dimensional distribution of points of varying intensity, specifying the amount of energy present at a given frequency and time, The Pattern Playback synthesizer [21], used in so many of the Haskins Laboratory experiments, and the channel vocoder synthesizer are both acoustic analogs, the latter being perhaps towards one extreme (the peak-picking vocodcr, a variant of the channel vocoder, is in fact rather close to an acoustic parameter approach [134]). DECIPUS, a machine at University College Department of Phonetics in London, which allows a small number of spectral patterns of varying durations to be played in succession, also comes into this category. Control information for such synthesizers is, typically, derived from some form of continuous spectral analysis. A spectrogram is one such analysis, consisting of a two-dimensional pattern of variable intensity marking on special paper. The marks are made according to the output of an analyzing filter which is scanned over the frequency range of interest while the marking stylus moves along the frequency axis. By wrapping the paper around a drum, and rotating this in synchronism with a repeated recording of the utterance, the time axis is generated. Figure 8 shows a spectrogram of the word "zero".'

Pattern Playback accepted its input in almost exactly this form, which allowed simplified spectrograms (embodying systematic variation of selected attributes) to be "played back" for the purposes of their experiments. There is a time-frequency tradeoff in spectrographic analysis. The more accurately the frequency is determined, the more the time structure is blurred, and vice versa. Figure 8 was obtained using an analyzing filter bandwidth of 300 Hz. Reducing the bandwidth would emphasize the harmonic structure at the expense of the spectral envelope. The original papers on the sound spectrograph appear in [117]. A paper by Presti [119] covers recent developments, including increased speed of operation, and the generation of intensity "contours" to clarify the marking.

An interesting, and virtually unexplored method of synthesis is that based on acoustic features, or events. This is a variant of acoustic analog synthesis in which speech is regarded as a mixture of events, the events being perceptually significant data groupings within the acoustic domain. Such an approach would handle transient phenomena as a single event, rather than a time series of changing spectra, and take specific account of a feature like "rapid onset." The approach involved in Schouten et al. [135] is certainly of this kind. Haggard has made some unpublished, related experiments in which, for example, using a parametric resonance analog synthesizer (see below, this section), the acoustic events associated with the initial fricative and stop of an utterance of the word "steel" were mixed with the acoustic events associated with the vowel and final lateral. The word "steel" was still heard (*sic*).

Synthesis in terms of acoustic parameters implies a model embodying all that is known of the constraints in the acoustic domain, and controlled by parameters directly representing the significant characteristics of the acoustic signal. Such a model is the resonance analog, the synthesizing device in a resonance vocoder system (see Section 3.5). Such a model should only generate noises within the range of a human vocal system, and indeed, this author's experience was that even

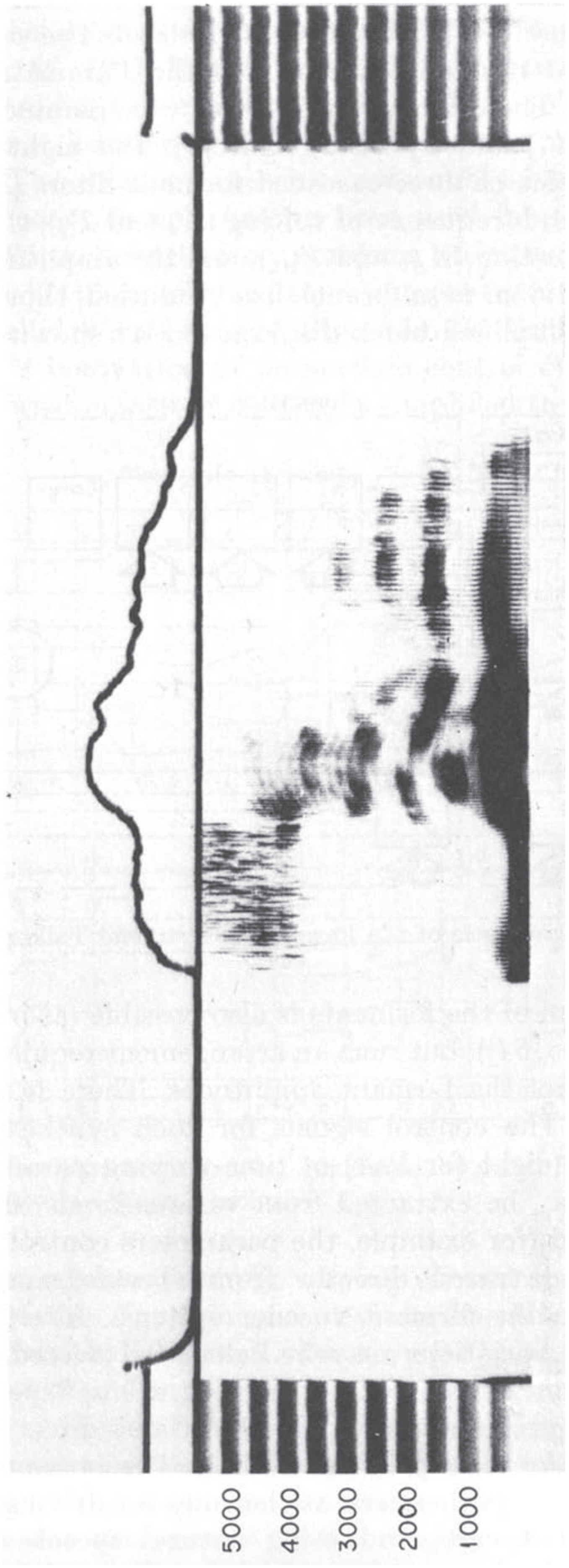

FIG. 8. Spectrogram of the word "zero"

uncontrolled variation of the parameters of one such synthesizer (PAT, see below) produced a very natural "belch." The original parametrically controlled resonance analog was the Parametric Artificial Talker, or PAT [81]. The original 1953 version used four parameters, but by 1958 it had six [143] and by 1962, eight [2]. The eight parameters control the frequencies of three cascaded formant filters (F_1, F_2, and F_3), the amplitude and frequency of voicing (A_x and F_x), the amplitude and frequency of frication (A_{H2} and F_{H2}), and the amplitude of aspiration (A_{H1}). An additional nasal branch has been used, though this does not appear in the simplified block diagram of PAT shown in Fig. 9.

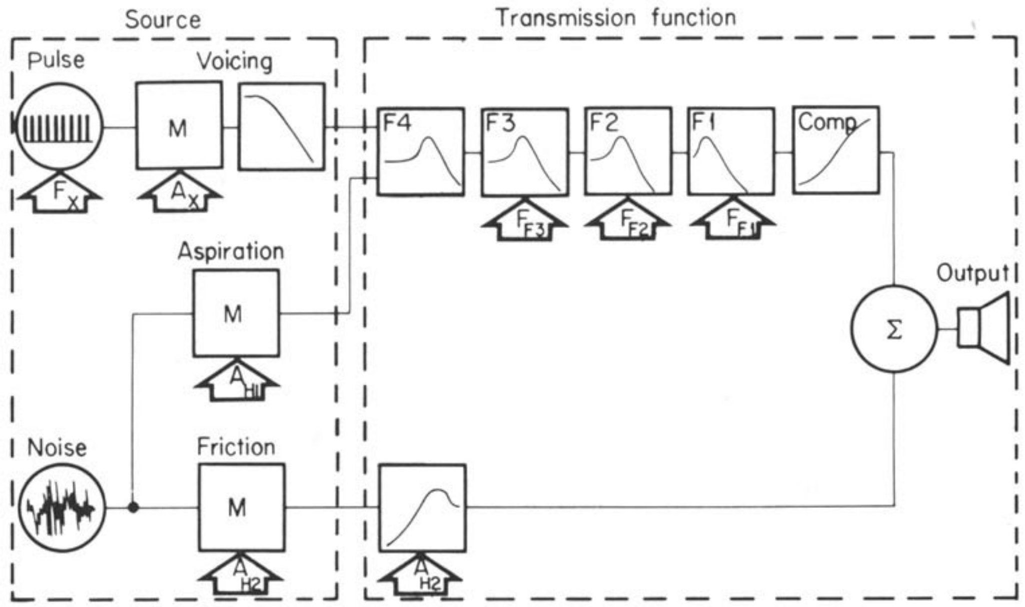

FIG. 9 Block Schematic of the Parametric Artificial Talker (PAT).

Parallel connection of the formants is also possible (as in the synthesizer used by Holmes [64]), but such an arrangement requires additional parameters to control the formant amplitudes. There is, of course, a gain in flexibility. The control signals for such synthesizers consist of a small number (eight for PAT) of time-varying parameter values. The information may be extracted from various kinds of analyses ol real speech, by hand (for example, the parameters controlling formani frequencies may be traced directly from spectrograms [153]), or automatically (as in the formant vocoder system). Alternatively, the required parameter variations may be calculated according to rules using phonetic script as the only ingredient. This "speech-by-rule" can be quite intelligible, and is produced without direct reference to any real utterance (for example, [50, 64, 95]). The present limit seems to be not so much the synthesizers, as the rules for driving the synthesizers. With sufficient care, and using natural speech as a model, Holmes has synthesized a sentence which is of very high quality in terms of intelligibility and naturalness [63]. This author has hand simulated the production of speech by rule (based on the Haskins findings [85, 86. 108], British vowel formant data published by Wells [166], and some departmental "rules-of-thumb" of the Phonetics Department at Edinburgh) and the parameter tracks produced appcar as Fig. 10. The resulting speech is reasonably intelligible, but the intonation is poor, no rule having been used for the pitch parameter. Mattingly [95] has worked on problems of suprasegmnental rules, and recently Iles [71] has achieved some success with adequate intonation based on Hallidav's theories of intonation for British English [53].

Lawrence's innovation of parametric control of a resonance analog synthesizer was important because it simplified the problem of control

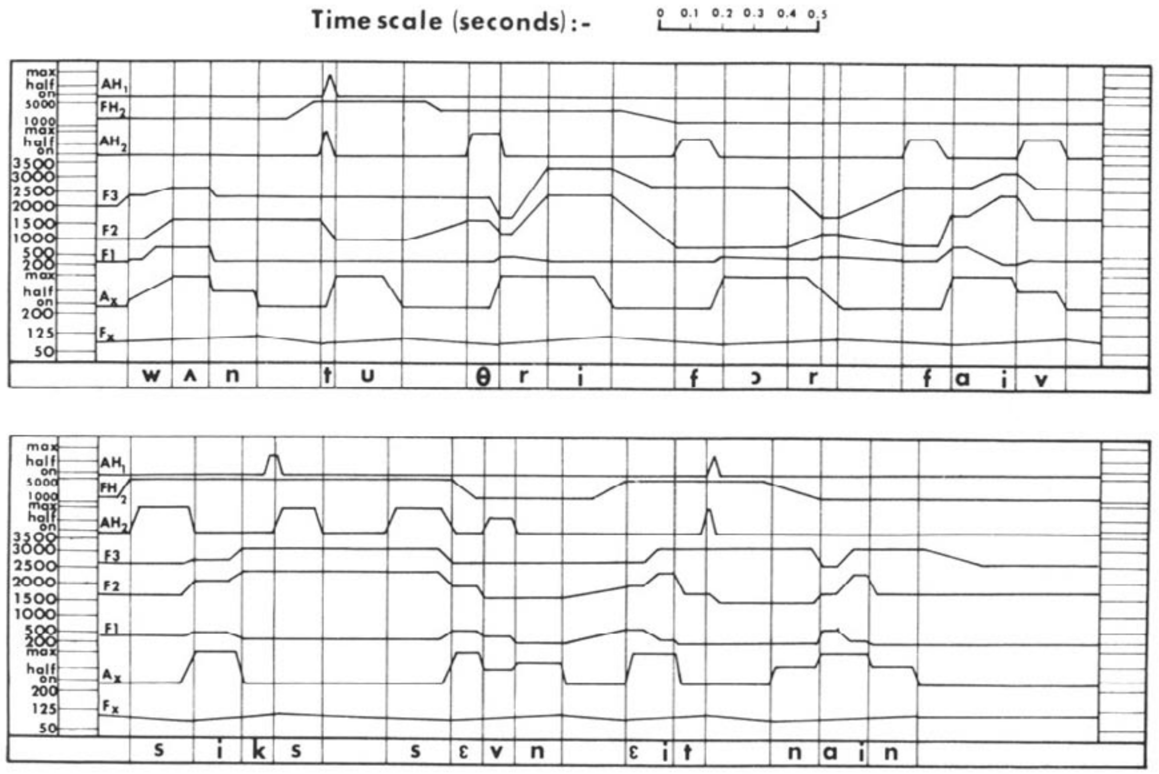

FIG. 10 PAT parameter tracks, generated by rule, for the synthesis of the nine digits "one" through "nine," inclusive.

A few meaningful and simply varying control signals suffice to determine the acoustic signal in such devices. The latest synthesizer of this type, in use at the Haskins Laboratory [96], uses nine parameters and three binary channels. A similar synthesizer (the one used by Dixon and Maxey) at IBM Laboratories, North Carolina, uses seven parameters and two binary channels. OVE III at Fant's laboratory at the Royal Institute of Technology in Stockholm [34, 88]—which includes a nasal resonator and control of the voicing spectrum—and the latest PAT represent developments of earlier versions, circuit and control improvements representing the main changes. All the latest machines or models have further simplified the control problem by allowing for digital control of the parameters, which has also reduced the practical difficulty of drift in the dc amplifiers required when using analog parameter-track control, and has simplified computer control of their behavior. Figure 11 illustrates the means of controlling PAT, fron a PDP-8/I computer, as developed at The University of Calgary. Two modes of operation may he selected. One packs two 6-bit channel samples per computer word, and loads all channels sequentially, at each sample time. The other mode allows an individual channel to be changed, storing the channel address, and one 6-bit channel sample in a computer word.

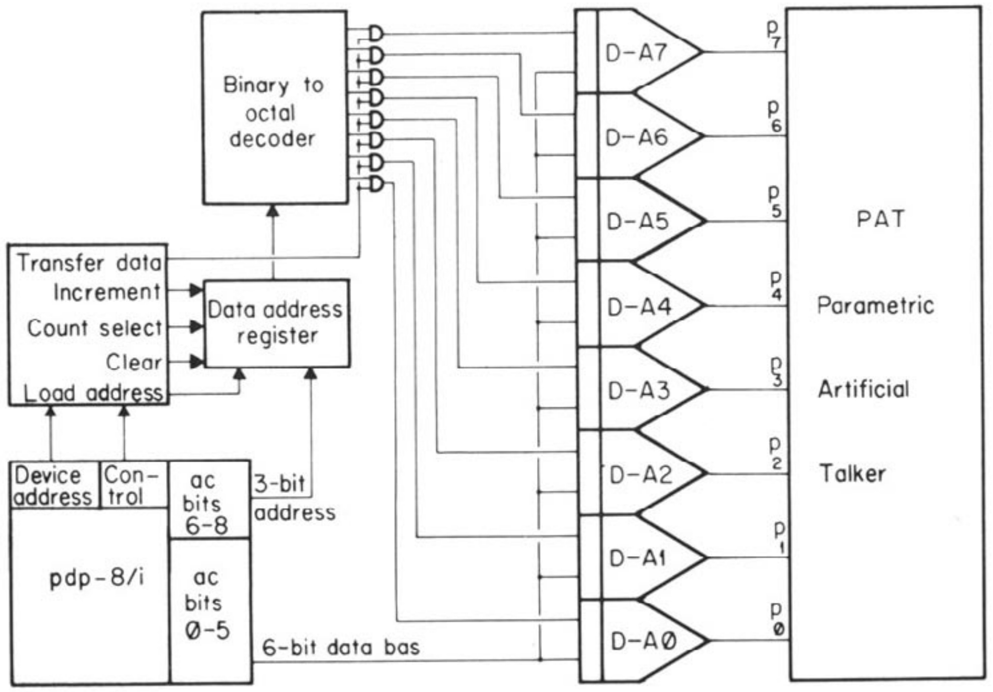

FIG. 11 Block Schematic of the means of attaching a PAT to a PDP-8/I computer.

The third approach to synthesis, by means of a physiological analog. involves a controlling apparatus which represents the physiological aspects of vocalization. Such analogs generally take the form of an electrical network consisting of sections, each representing a short cylindrical section of the vocal tract, of variable cross section (see above Section 4.2). DAVO (Dynamic Analog of the Vocal Tract) at the Research Laboratory of Electronics,M.I.T., was one such analog. A nasal branch (DANA) was added later [56]. There are problems with dynamically variable hardware analogs in the form of a tendency to instability [79]. Another problem arises in the control. First, the control signals, though parameteric in form, are arbitrary, and reflect too little of the real constraints of the real vocal tract; secondly, data on the area functions required is very difficult to obtain, and to date no satisfactory method has been invented for obtaining vocal tract area data, though various methods, including that of filling the mouth cavity with moulding compound. have been tried. Height, in saggittal section X-ray data, is usefully related to the requited data, but there seems little progress of immediate practical importance in this area, partly because of this lack of data, partly because of the difficulties experienced with dynamic analogs, and partly because parametric resonance analog approaches have proved so simple, practical, and economical.

The fourth approach to synthesis would require parameterization of the physiological aspects of vocalization, to give a few slowly varying signals related to the important features of physiological variation. This could be supplied by an articulatory model, requiring parameters such as “position of tongue hump.” Even less progress has been made in this area, which has important implications for our understanding of the speech process. Work in the Department of Linguistics at the University of California, Los Angeles, is specifically directed toward obtaining the data needed for the required characterization of speech in terms of physiological parameters [77, 79] and a more fundamental understanding of the whole speech production process.

4.3 Discussion

Synthesis-by-rule, using a parametric resonance analog synthesizer, seems to offer the best prospect for talk-back from machines, at present. Such an approach, besides being immediately practical, seems to offer clear advantages in terms of the amount of information to he stored (at the most, say 250 bits per second instead of 40,000 to 100,000 bits per second required for direct representation) and in terms of flexibility; by changing the rules one could, perhaps, change the “accent” of the machine. In practice, however, the techniques for handling speech information may he combined in almost any proportion. One may store a representation of the low information rate signals required to drive a parametric resonance analog synthesizer giving a bit rate of perhaps 2400 hits per second; or one may “vocode” the original speech. and store the vocoder channel symbols. These symbols it may be remembered. are typically binary codes which represent the varying energy levels in various frequency bands, togethcr with some additional information about the presence/absence of voicing and pitch. Such vocoding may again reduce the bit rate to about 2400 bits per second, though this figure represents the state-of-the-art on channel vocoders, and the quality is poor. It may be seen that the two methods—recording or synthesizing—are not so far apart as might be imagined. They are, perhaps, closest when an utterance is synthesized by copying the control signals from a real utterance. This is essentially what is done by using a channel vocoder synthesizer in the manner suggested. The control signals for a parametric synthesizer may also be extracted automatically from real speech as in the resonance vocoder, though, so far, with less success. In general. one should distinguish what is recorded, and what is synthesized. Storing vocoder samples is more economical than storing the original waveform, but a synthesizer is required to reconstitute the speech. Of course the synthesizer may he simulated, but this takes a lot of computing power and real-time operation, and becomes impractical if not impossible. Whatever form is used, storage in normal computer memory simplifies the data access problem prior to synthesis.

If direct representation is used, then the recording may he either continuous (which is familiar to most people) or discrete (digital), which requires analog-to-digital conversion during recording, and digital-to-analog conversion during play-back. Digital recording again offers the advantage of being able to store the waveforms in ordinary computer memory from whence they may he retrieved at will and replayed. However, in the departure from straightforward recording, it has again been necessary to introduce a rudimentary analysis (A-D conversion) and synthesis (D-A conversion).

Problems of random access to large numbers of stored messages in analog form make it seem rather unlikely that straight recording will ever be of any great importance, except in situations where fidelity or other special considerations are paramount, or where the vocabulary is so small that analog drum storage, or something similar. is economical. An attractive, but unexploited, method of storing and replaying direct recordings of isolated words. or short phrases, would use the principle of the Mellotron organ, pulling spring loaded strips of ordinary one-quarter inch magnetic recording tape past replay heads, by means of solenoids. Such a system offers immediate, random access, interchangeability, ease of expansion, and parallel capability.

At present, IBM offers several versions of a basic drum talk-back system (the 7770 Audio Response Unit [70]) which, for example, deals with forty-eight outgoing lines, and stores one word per drum track; also a channel-vocoder based system (the 7772 ARU [69]), which is able to deal with eight lines per synthesizer, and up to four synthesizers The quality of this low-bit rate (2400 bits per second) vocoded speech is considerably improved since it is not analyzed in real time, and arbitrary adjustments may be made to improve quality and intelligibility [11, 12]. These two audio response units allow spoken messages to be generated by suitably equipped computers. Stock quotation have been generated using such a system in New York. An advantagc of talk-back is that an ordinary telephone may be used as a remote enquiry terminal (above). At present the query must he keyed or dialed in, but a recognition input would allow a true speech interface. Speaking clocks often use the drum method of "speaking."

On long-term considerations, speech output by rule synthesis from a phonetic, or similar low-bit-rate, storage format seems to offer many advantages. A prime advantage is that the spoken output can he completely unrestricted, at least in a better sense than the sense that a present line-printer output is unrestricted. There would not be "characters" which were unprintable. Indeed, instead of being restricted to one type font, many different accents could be produced, the equivalent of an infinitely variable font, by suitable rule changes Only a small amount of information need be stored for each message. The output rate would be greater than that for the printed output from a Teletype. Faster output still could be achieved by increasing the output rate, and using normal rate playback later, but it would. of course, involve recording. It is not foreseeable that printed output can be done away with, indeed one must avoid the suggestion that speech output (or input) is a panacea, but one can foresee that edited recordings could be "typed up" by a special purpose recognition machine, whose recognition task would be considerably simplified because of the formalized nature of the machine-produced speech. Such a system might offer advantages for some secretarial work.

At present, offline synthesis and online assembly of diphone segments is a good approach, but this author believes it is not sufficiently general and will be displaced by rule synthesis as soon as the techniques are perfected. It is likely, however, that the diphone will replace the phoneme as the basic unit for synthesis by rule, since considerable saving in the required complexity of rules, and hence computing time needed for synthesis, could result.

4.4 Cost

To give an idea of the cost—it is now certain that speech parameters may he synthesized for a parametric resonance analog synthesizer in at least real time, and soon (using new techniques such as diphone-based rules, and faster computers) in better than real time. This will he CPU time. Being rather pessimistic about the size of computer, required one might suppose that speech output would cost $600 per hour at present, or about fifteen times the cost of a university lecturer.

Being optimistic, however, such output means are likely to be used in interactive situations, when the computer will only be talking half the time. Furthermore, in a multi-access environment, the computer may be "i/o bound" and the central processing unit (CPU), say. 50% idle. In such a case, the rule synthesis might simply mop-up available CPU time and "cost" nothing. These represent extremes. In these days of parallel processing, it could be worth having a second CPU in such a system, simply for speech synthesis. There is also the cost of the synthesizer hardware—a few hundred dollars.

A final word on the economics with regard to a particular application. With a multi-access system, the old choice of "store-versus-compute" comes up quite clearly. Many common words probably should not he completely resynthesized each time needed, and very common messages may be stored in parametric form for multi-access. For a Computer-Aided Instruction situation, messages may be synthesized when first needed, and then kept until the slowest student has passed the region in the course where the message may be used. Whatever view is taken. speech output is unlikely to be ruled out on economic grounds.

4.5 Conclusion

The practical steps taken towards speech output for machines are thus seen to be many, varied, and effective. At present, voice output is available by the expedient of partial or complete storage of speech originally generated by a human. In general the less the information stored for messages, the more the information required as a supplement, either in terms of hardware which embodies "rules" of some sort, or extra computer program which again embodies "rules" With larger message sets, the reduced storage achieved by more sophisticated approaches becomes increasingly attractive. A particular advantage of parametric synthesis is the potential ease and economy in dealing with the intonation and rhythm of speech. For connected speech, this advantage would be an overwhelming one. For many practical purposes in the future, it is likely that speech output will be generated by rule from a diphone representation, using an on-line, parametric resonance analog synthesizer.

5. Speech Recognition

5.1 Introduction

We now turn our attention to recognition. In his report on establishing voice communication with computers, Lea [82] dismisses consideration of the synthesis problem on the grounds that recognition is a more difficult problem. He states that work must concentrate on recognition, since the "lead" time required to get an operational recognizer exceeds very considerably the lead time needed for adequate synthetic output. The previous section draws a picture supporting Lea's view. It is important, however, to define what is meant by an operational recognizer, and to what end is research in recognition most effectively pursued. Some reasons for the greater difficulty of recognition have already been discussed. The basic problem is the involvement in "perception." The higher level processes associated with perception are not so well understood as the acoustics of speech production Thus the problem is one of representation, or how we specify the task of recognition, and the input to the machine, in such a way that words which are the "same" (defined as the "same" by a panel of listener who speak the appropriate language) produce sufficiently similar representations within the machine to be judged the same, by the machine; while, at the same time, allowing the machine to discriminate between words that are different. The requirements for the representation prior to synthesis are not only better understood, but are also much less severe, because there is no strong requirement for generality. This, probably explains the "synthetic" quality of all machine-produced speech to date, as well as explaining the fact that it has been possible at all.

In Flanagan's book [37] the topic of Automatic Speech Recognitiom (ASR) and the related topic of automatic speaker recognition, receive cursory attention—eight pages. This perhaps reflects to some extent the smaller amount of progress, but probably also represents the reputation of such work at the time the book was written. Nevertheless the whole book is a handbook on ASR. However, it cannot and does, not treat certain important experiments in neural processing of sensory input data in animals [42, 65, 84, 97, 167] which tend to support idea of relative measures and feature extraction as mechanisms in neural processing; general problems of pattern recognition; the huge subject of generative grammars, and other linguistic topics of importance; nor the equally huge topic of decision theory. It is concerned with speech research *per se* not language, decision-taking, pattern recognition research, or neurophysiology. It is worth noting that Lindgren [89], writing in the year of the book's publication, says the engineer who attempts to build a speech recognizer must "drink deep of the linguistic mysteries." This may be amplified. He must become well versed in such diverse topics as the visual perception of frogs, and the resonant properties of organ pipes; the study of phonetics, and the study of decision strategies; the physiology of the human ear, and problems of generative grammars; the meaning of meaning, and frequency analysis. What is, perhaps, worse—he must continue to be well versed. Lindgren also notes that " … the question, 'What is the present state of automatic speech recognition?,' transforms itself into the question. 'What is the present state of speech research?'," which is the subject of Flanagan's book. There is, however, no "neural theory of speech perception" to complement Fant's work on speech production, nor are there any clear divisions by which the field of automatic speech recognition may he partitioned for study. Much of this section, therefore. will rely on particular examples, and citation, rather than any clear signposting of the field. Ottcn [113] has reviewed some of the problems in automatic speech recognition, from a traditional standpoint. Hyde [68] has also surveyed work on automatic speech recognition up to 1967. There is an excellent and wide-ranging collection of earlier papers relevant to automatic speech recognition available as two volunles used for the University of Michigan summer course in 1963 [156]. and Hill [58] describes a hypothetical recognition machine which constitutes a review of the problems. Hyde's report includes a bibliography. Other recent ASR bibliographies include Hill [57] (noted above) and Paulus [114]. the latter including a second section covering linguistically oriented material.

5.2 Approaches to Machine Perception of Speech

5.21. Waveform Matching

The input to a machine which recognizes speech is most convenienly thought of as an electrical representation of an acoustic waveform. The transformation invokes the familiar microphone, and the only word of caution needed is to remark that some microphones do no t produce very precise electrical copies of the original. In particular, the carbon microphone (used, for example, in British telephone handsets) produces an electrical waveform which is approximately the time derivative of the original acoustic waveform, and of restricted bandwidth. An "obvious" mechanism, to allow machine response to selected words would he to match the incoming waveform with a stored set, in either continuous or discrete representation, and assign a response or label to the incoming waveform according to the stored waveform which most nearly resembled it. Although this approach, on the face of it, is straightforward. and of guaranteed effectiveness, it is completely useless in practice because of the variability of speech. Even one speaker, trying hard, cannot reproduce a waveform with sufficient precision. for a given word. A particular waveform version of an utterance is satisfactory for machine-generated speech; it is not nearly general enough for purposes of machine perception. Storing manv versions of the waveform for each utterance offers little hope either since, to maintain discrimination between utterances that were different, a match would need to be quite precise to allow "same" judgement, and the machine would spend more time storing new waveforms than in producing responses.

5.2.2. Analysis, Segmentation, and Pattern Recognition

(a) General Analysis is the converse of synthesis, and just like synthesis it assumes some underlying structure for the object being analyzed. The more complete the model in terms of which the analysis, is performed, the simpler (for example, in terms of bits of information) the results of the analysis. This fact is entirely analogous to the fact that the more complete our model of speech, the simpler the ingredients required for synthesis—as in synthesis by rule, where only about fifty hits per second suffice to specify intelligible speech. There is one important qualification. Models for speech synthesis need account for nothing beyond the acoustics of speech to make noises that are meaningful to humans. Synthesis need not be concerned with meaning, at least at this stage. On the other hand, a voice responsive machine is inevitably concerned with meaning, in that it must associate a meaningful response with a given meaningful input. Even a phoneme recognizer is responding on the level of meaning, rather than mere identity, for assignment of phonemic identity to speech sounds must involve meaning (see above, Section 3.3—this is the underlying problem in the need for generality). Unless a voice-responsive machine is simply to produce a phonetic transcription of the input, then it must produce responses at a higher level of meaning than phonemes—it must *do* something appropriate to the meaning of the input (for example, stop when told to "stop").

(b) Segmentation Speech is a continuous waveform punctuated by pauses (to allow breathing, thinking, or reply) and sprinkled with short gaps, representing some stop sounds (voiced stops may exhibit voicing all through the "gap"). No simple way has yet been devised to place boundaries consistently at all phoneme boundaries (see below, Sitton [139]). The reason is simple—the latter boundaries do not, in general, exist. One part of the rules needed for synthesis-by-rule, on a phonetic basis, is concerned to calculate how to merge individual phonemes smoothly into each other. Boundaries are purely conceptual, and as such can only represent the "midpoint" of the transition from one phoneme to the next. Truby highlighted this fact as long ago as 1958 [151]. The desire to put in segment boundaries where any distinction can only be based on compromise, has led to what is variously

called "the segmentation problem"—which, as noted above, Stevens suggests "is the problem that you can't"—or to the "time normalization problem," which arises from an inability to segment, or from attempts to segment which do not succeed. This author believes that it is only *necessary* to segment at the level of meaning, and that up to this level as many possible relevant clues as can be handled should he preserved, relying on the excellence of the model used to ensure economy. A phoneme recognizer is only necessary insofar as phonemes have meaning to some system. It may not be necessary to make a phonetic inventory of an utterance before ascribing meaning at a higher level. To draw an analogy with reading—it is possible to read by identifying letters, and hence deducing words, but an efficient reader recognizes (and relates to meaning) whole words. He even skips large numbers of words, if letters are in doubt—as they might he in a handwritten message—it may be necessary to determine the word, and the letters composing the word, by parallel processing, using context, yet for common words, a squiggle will suffice.

Machines intended to produce unique responses to words must segment into words at some stage, it would seem, at least in the sense of producing a unique output for each. Otten [113] distinguishes two varieties of segmentation: direct segmentation, in which an explicit segmentation procedure cuts the continuous speech into segments which are then classified; and indirect segmentation which requires continuous classification of the waveform input, with segmentation partly implicit in the classification, and partly in *post hoc* procedures operating on the (tentative) segments produced. Sitton [139] has recently developed a new approach to direct segmentation, at the phonemic level, which looks promising. Much current work on voice responsive machines assumes that segmentation into phonemes. at some stage, is necessary, or desirable, or both. One common argument suggests that only a small alphabet is then required to represent any word in a language. However, to produce a unique response (other than a spelled output) still requires a unique output line for the response, and a further stage of pattern recognition. To make the step to word or phrase recognition, in terms of a unique output for each, from unsegmented, noisy phoneme strings is decidedly nontrivial. Newcomb [105], Petrick [115], Morgan [104]. Alter [1], and Vicens [158] are among those who have worked on some of the problems. Work by McElwain and Evens [100] on a degarbler of machine-transcribed Morse code is also relevant. Even spelled speech, if it involves transforming phoneme strings directly into correctly spelled words without identifying the words, is less than straightforward, since this assumes that people all speak as they should, in terms of the phoneme strings they utter for a given word—which they do not; or it assumes some rather sophisticated error correction procedure operating on the phoneme strings: and finally, a phoneme-to-spelled-speech conversion is required.

In conclusion to the topic of segmentation, it may be said that the rules used in speech synthesis by rule, and implied in the speech production process, effectively scramble the egg that is speech. Segmentation is a procedure to help in unscrambling the egg—if it works, The most widespread strategy to assist in this unscrambling process, adopted by nearly every worker to date, has been to require speakers to speak the words or phrases (meaningful units), which their machines were designed to recognize, in isolation, thus *imposing* cues for direct segmentation at the level of meaning. One strategy which does not seem to have been considered at all is to aim at diphone segmentation akin to that used for synthesis by Dixon and Maxey. Fant has suggested that segmentation should he imposed at places where the time derivative of spectral measures is greatest. Perhaps an approach based on the opposite idea, segmenting only where the rate of spectral change was minimum, would lead to more readily identifiable segments of a "diphonemic" kind, This author favors indirect segmentation at the word or phrase level for current work on voice-responsive machines. However. in order to recognize any segment, the segment must first he described, and this is another great problem.

(c) Description and the Problem of Feature Selection Description of an object implies an analysis in terms of the important features of the object. The important features of an object may be its function, its composition, or any number of attributes. If two objects differ in terms of an important feature, then presumably the two objects are not the same. Conversely, if two objects are the same, then presumably they do not differ in any important feature. The problem in deciding whether two speech utterances are the same is that acoustical descriptions refer to the composition of the utterance, while they are judged the same on the same functional basis that serves to define phonemes (see section 3.3). The problem of choosing acoustically derived features to represent what is, at least in part, a functional equivalence leads to what has been called "the feature extraction problem." The possibility of success in solving this problem is presumed on the basis that every listener must solve it; but, clearly, the analysis must be carried out in terms of a model that includes parts at a higher level than the acoustic level, One is concerned not with machine classification so much as machine perception; a process by which acoustic evidence is interpreted and evaluated in terms of a total model. Perception seems to be a constructive, organizing process, Which requires a period

of learning at a level above that of the raw data. This is why people confronted by an utterance in a new language cannot even hear the words, which would at least allow them to ask a bilingual companion 'What do — and — and — mean?" There is parallel evidence in visual perception. Persons, blind from birth, given sight by an operation, have been appalled by the confusing, disorganized patches of light and color confronting them, on removal of the bandages, and some return completely to the security of their world of touch and sound by wearing opaque glasses [162]. Current psychological theory of perception assumes an active process, which consists largely of rules—for organizing the data input—learned from experience [145]. This need for an active component in speech perception has led in the past to "the motor theory of speech perception" (for example, [87]). Halle and Stevens [52] and Stevens [141] proposed "analysis-by-synthesis" approaches to automatic speech recognition on this basis. Hill and Wacker [60] reveal a scheme studied in 1963 which used a learning machine as part of the active process. Denes [23] attempted to obtain direct confirmation of the motor theory of speech perception by psychophysical experiments, using a speech distorting system The control group learned the distorted auditory patterns related to printed text; the experimental group had the same task, but could hear their own voices after being through the same distorting system. Little confirmation of the motor theory was obtained.

An analysis-by-synthesis approach to speech recognition my be necessary, but it does not solve the problem of feature extraction, or for that matter the problem of segmentation/time-normalization. Speech still has to be described in terms of the important features in order to judge whether or not the internally generated construct matches the incoming acoustic stimulus. To try to carry out this matching process in the acoustic domain merely consigns the two major problems to the comparison machine. It seems likelv that the constructive process must be directed upward, from acoustic primitives to recognizable pattern descriptions, rather than outward, from abstracted acoustic descriptions to an acoustically matching response. This is where the requirement for information at a higher level than the acoustic level arises, for out of all the pattern groupings that could be learned, only those that are meaningful are learned. Sutherland [145] argues the points involved in such a model of perception very cogently, for the visual case. We do not, at present, know enough about the process of speech perception to duplicate the feat. Sutherland's model for visual perception, which is still highlv general at present, arose because, as a psychologist obtaining results from experiments on visual perception. he found that information

from work on the neurophysiology of vision and on machine procedures for automatic pattern recognition (particularly Clowes [18, 19], but also well exemplified by Guzman [48, 49]) could be combined into a model of visual pattern recognition which offered great explanatory power in terms of the results obtained. A start on the explanation of speech perception may yet well involve a similar interdisciplinary approach, based on the same combination of psychology, machine methods, and neurophysiology. An appreciation of the need for such an explanation, on the part of those attempting machine perception of speech, and an appreciation of the source of relevant information, including related work in the field of visual pattern recognition, is probably essential to real progress in machine descriptions of speech patterns, of which the feature extraction problem is but part.

(d) Models and Analysis The development within this subsection will follow, as far as possible, the lines of the section on models and synthesis [Section 4.2.2(b)] in an attempt to show the relationships, and illustrate more clearly some of the unexplored areas. The picture is complicated somewhat by the need to account for a time scale which can no longer be chosen arbitrarily, and which varies considerably; by the ingenious variations that have proved possible, in describing components of the models involved, in specific realizations; and by the fact that, whatever model is adopted, there ultimately must be some form of segmentation, which interferes with any attempt at clear distinctions between the approaches, in the former terms. Also physiologicallv based approaches are almost unrepresented.

The very simplest model of speech, then, assumes a series of concatenated units. In the analysis case these may be interpreted as segments of the waveform which can he identified. Since segmentation at some stage is inevitable, this approach hardly differs from that of waveform matching, except that by trying to match shorter sections. the process should presumably be somewhat easier. An approach which identified phonemes on the basis of autocorrelation analysis of waveform segments would be in this category, but no significant devices appear in the literature. Such an approach would almost certainly require reliable direct segmentation.

Other approaches to analysis attempt to retrieve the ingredients that would need to be used, by some particular model, in producing the unknown utterance. These ingredients may be used as descriptors of segments at particular levels in order to classify them. Almost always, segments compatible with phonemic analysis are used at some stage. although for convenience in computer analysis the raw data may initially consist of shorter segments arising from some fixed sampling scheme. Early approaches to automatic speech recognition worked almost entirely in the domain of the acoustic analog, and the simplest version thereof. The approach is characterized by some form of spectrographic analysis, giving a two-dimensional array of data points representing energy intensity at some time and frequency. Time, frequency, and intensity are usually quantized to have a small number of discrete values—usually two for intensity, energy present or absent. Thus analysis produced a binary pattern which could be matched against stored patterns derived from known words, and a decision made as to which of the stored patterns most nearly resembled the unknown input, hence naming the input. This was the method of Sebesteyen [130], Uhr and Vossler [154, 155], Purton [120]—who actually uses multi-tap autocorrelation analysis rather than filter analysis, Shearme [137], Balandis [3]—who actually uses a *mechanical* filter system, and Denes and Matthews [24], as well as others. A rather similar kind of analysis may be obtained in terms of zero-crossing interval density or reciprocal zero-crossing interval density, the latter being closely related to frequency analysis [127, 128, 132, 139]. The major difficulties with either analysis lie in the time and frequency variability of speech cues. Time-normalization on a global basis (squashing or stretching the time scale to fit a standard measure) assumes, for example, that a longer utterance has parts that are all longer by the same percentage, which is not true. One simple way around the difficulty was adopted by Dudley and Balashek [28]. They integrated with respect to time for each of ten selected spectral patterns, and based their word decision on matching the ten-element "duration-of-occurrence" pattern, for an unknown input, against a master set. The more usual approach adopted has been to "segment" the input in some way, usually into phonemic segments, so that a series of (supposedly significant) spectral patterns, or spectrally based phonetic elements, results.

Segmentation is either indirect—segments beginning when a named pattern is first detected and ending when it is no longer detected—or it is based on significant spectral change, much as suggested by Fant [see Section 5.2.2(b)]. Segmentation is frequently two-stage in computer-based approaches, the short (10 msec) raw data segments due to the sampling procedure being lumped to give the larger segments required. Vicens [158] describes one such scheme of great significance, using three stages, followed by "synchronization" of the input segments detected with the segments of "candidate" recognition possibilities stored in memory. His approach may be considered a "head-on" attack on the related problems of segmentation and time normalization and is the only scheme with a demonstrated ability to handle words in connected speech. The work is important for other reasons, as well. Other examples of the indirect approach include Olson and Belar [109], Bezdel [5], Fry [43], and Denes [22]—who also built in specific linguistic knowledge for error correction, and Scarr [133]—who incorporated an interesting scheme for amplitude normalization of the input speech. Examples of the direct approach include Gold [45]—who included other segmentation clues as well, Ross [125]—who included some adaptation, Traum and Torre [150], and Sakai and Doshita [128]. Needless to say either approach requires a further stage of recognition, to recogize words. Sakai and Doshita, and Bezdel used mainly zero-crossing measurements in their schemes. This may he significant in building machines entirely from digital components.

At this stage. the different approaches become harder to disentangle. Individual segments. usually phonemic in character, may be described partly in terms of the kind of parameters used by a parametric resonance analog synthesizer, and partly in terms of measures derived from these parameters, or from spectral attributes. Otten [112] and Meeker [101] both proposed that time parameters in a formant vocoder system should be adequate for recognition, but the results of any such approach are not generally available. Forgie and Forgie approached vowel and fricative recognition on the basis of formant values, fricative spectra and transient cues, which are related to such parameters, and were quite successful [39, 40]. They reported up to 93% correct on vowel recognition, and "equal to humans" on fricative recognition. Frick [42] pointed out the advisability of "not putting all the eggs in one basket," stating the M.I.T. Lincoln Laboratory aim at that time as being to define a set of cues which might be individually unreliable but, in combination, could lead to a reliable decision. This really indicates the general feeling for the decade that followed. Such a philosophy clearly leads to combined approaches. which are still in vogue. Here, really, is the nub of the "feature extraction problem." which now centers around the question of which particular blend of which features is best suited to describing individual segments. The not too surprising conclusion, according to Reddy, is that it depends on what particular kind of segment one is trying to classify. Thus in Reddy's scheme [121], segments are broadly classified on the basis of intensity and zero-crossing data, and finer discrimination within categories is accomplished on the basis of relevant cues. This seems an important idea, spilling over from research on the problems of search in artificial intelligence (and is the one developed by Vicens). He emphasizes the point that much automatic speech recognition research has been directed at seeking a structure, in terms of which the problems might he formulated, rather than seeking the refutation of a model or hypothesis. He further remarks that lack of adequate means for collecting and interacting with suitable data has held up this aspect of the work.

In 1951 Jakobson, Fant. and Halle published their *Preliminaries to Speech Analysis*. This work. recently reprinted for the eighth time [72] has a continuing importance for those working in fields of speech communication. The idea of distinctive features is based on the idea of "minimal distinctions," a term coined by Daniel Jones [74] to indicate that any lesser distinction (between two sound sequences in a language) would be inadequate to distinguish the sequences clearly. Distinctive features are thus tarred with the same brush as phonemes; they are functionally defined. In *Preliminaries to Speech Analysis* it is suggested that a minimal distinction faces a. listener with a two-choice situation between polar values of a given attribute, or presence versus absence of some quality. There may be double, or triple differences between some sound sequences. But the fact that there are interactions between adjacent phonetic elements in speech means that in practice a "minimal distinction" may not be confined to one phoneme. For example English "keel" and "call" are distinguished by more than the distinctions between the medial vowel sounds; for instance, the initial velar stops are markedly different, but the difference is not phonemically significant in English (though it is in some languages). However, distinctive features form a ready-made set of binary descriptors for phonemes, which fact has had practical and psychological effects on those working towards machine recognition.

In their book, Jakobson, Fant and Halle describe the acoustic and articulatorv correlates of their distinctive features, but in qualitative terms. They remark that use of distinctive features for acoustic analysis would make the analysis easier, and supply the most instructive (acoustic) correlates. The search for acoustic correlates is not yet ended, for rather the same reasons that phonemes have resisted description. Daniel Jones' discussion of minimal distinctions in [74] reveals some of the difficulties. Nevertheless, approaches in terms of "distinctive features," whether following the original set, or merely using the concept, but instrumenting features more easily detected by machine, have been made, and the concept has had a systematizing effect. One noteworthy scheme, which follows the original distinctive feature set closely is that of Hughes and Hemdal [67]. Vowel recognition scores of 92%, and consonant recognition scores of 50%. giving over-all recognition scores for words comparable with those of listeners, are reported for a single speaker. Their results indicated that hopes of a nonadaptive recognition scheme for more than one speaker arc ruled out. An earlier approach. based on the original distinctive features, was that due to Wiren and Stubbs [169].

Work by Bobrow and his colleagues [7, 8] is the only reported truly parametric approach, though using rather different parameters from those required for a resonance analog synthesizer. Various binary features are detected, on the basis of the outputs of a nineteen-channel spectrum analyzer, and these are then treated independently, as binary parameters of individual utterances. For each feature (which at any given time may he "1" or "0") the time dimension resulting from fixed sampling is collapsed, to give a sequence of single l's and 0's (i.e., 101 rather than 1110001). Each feature sequence present votes for any recognition possibility in which it has appeared, the output comprising the most popular possibility. Two sets of features were used, a linguistically oriented set and an arbitrary set. Both performed comparably, but the latter set degraded more when the parameter patterns for one speaker were tried on another. Single speaker recognition scores of 95% were achieved for single speakers on a fifty-four word vocabulary. The developments reported in the second paper were concerned with the testing of some new features in conjunction with an increased vocabulary size to 109 words. Recognition scores of 91-94% were achieved. This approach is a departure from the con - ventional approach of segmenting into phonemes, classifying the phonemes on some basis, and then recognizing words on the basis of phoneme strings. However, it throws away a certain amount of information about the relative timing of events.

Vicens' scheme carried out segmentation, primary classification, and detailed matching in terms of six parameters based on the amplitude and zero-crossing rates of the signals derived in three frequency bands related to vowel formant domains. it is thus, in a sense, a parametric approach, but is concerned with segments as descriptors (of units of meaning), so is not truly parametric. Recognition performance for isolated words and single speaker (98% on a fifty-four word vocabulary in English, 97% on a seventy word vocabulary in French, and 92% on a 561 word vocabulary in English) and for words embedded in the connected utterances of strings of words in simple command language statements (96% on words, 85% for complete commands) is impressive. He makes a number of important points concerning ASR machines: that the features or parameters used are less important than the subsequent processing used; that accurate phoneme-like classification may be unnecessary in practical applications using a suitably restricted language; and that the techniques of artificial intelligence are important in ASR research. He also points out that Bobrow' and his colleagues [7, 8] have an approach which does not extend easily to longer connected utterances in which a division into smaller segment. (words or phonemes) is necessary, though this is clearly understood by those authors, who called their machine "LISPER" (Limited SPEech Recognizer). One reason is that information concerning relative timing of feature sequences is not preserved.

Another new approach to the handling of the kind of information derived from binary feature extractors is described by this author in earlier papers [59, 60]. The approach is based on the notion of treating the start and end of binary feature occurrences as events. The input may then be described in terms of the occurrence and non-occurrence of various subsequences of these events, which provides a binary pattern output for decision taking. There are means of applying various constraints on what events are allowed, and what events are not allowed in the sub-sequences. The scheme has certain general properties in common with that of Bobrow and his colleagues, but it specifically includes information about relative timing of individual features. In this respect, it perhaps has something in common with the model of speech which regards speech as a mixture of events [see Section 4.2.2(h)]. It assumes the necessity of analyzing speech explicitly in terms of both content (events) and order, a distinction noted by Huggins [66] as long ago as 1953. The scheme (at time of writing) has had no testing with a reasonable set of features. Experiments on an early prototype hardware version, using a feature set consisting of just the two features "presence of high frequency" and "presence of low frequency," gave recognition scores of 78% correct. 10%, reject, and 12% misrecognized, for a sixteen-word vocabulary of special words (chosen with the feature limitations in mind) and twelve unknown speakers [59]. In the new machine [60] the determination of subsequences is independent of any absolute time scale. As a consequence of the strategy used to achieve this time independence the machine may he able to say what has occurred at some level of analysis. but it may be unable to say at all precisely when it occurred, since the irrelevant aspects of absolute time are discarded at an early stage in the processing. This failing it seems to share with the human [78]. The detection of subsequences is carried out by a "sequence detector." This is essentially an implementation of a simple grammar-based description machine, and operates in the dimensions of time and auditory primitives in a manner entirely analogous to the way, say, Guzman's figure description language operates in the dimensions of space and visual primitives (see below. Section 5.3). The bit pattcrn output provides a staticized commentary on the history of the input . By arranging for deactivation of those binary outputs from the sequence detector which lead to a particular decision at some level of meaning (initially words), or which are too long past to he relevant to the current situation, the system is readily extended to dealing with the recognition of successive parts of connected utterances.

Other approaches, each of interest in its own way, can be listed. An exhaustive survey is out of the question. Dersch [25] described a recognizer based on waveform asymmetry, as a measure of voicing and as a crude vowel discriminant. Recognition depended on identification of a voiced segment, which was crudely classified and which could be preceded and/or followed by friction. A recognition rate of 90% to the ten digits is quoted, using both male and female speakers. Scores could be improved with practice. Martin *et al*. [94] describe a feature based system using neural logic operating on spectral information. Following recognition of phoneme segments, words are recognized on the basis of particular phoneme sequence occurrences. Teacher *et al*. [147] describe a system aimed at economy. Segments are classified according to the quantized values of only three parameters—the Single Equivalent Formant (SEF), voicing, and amplitude. The SEF is reported as a multi-dimensional measure of vowel quality based on the zero-crossing interval under the first peak of the waveform segment in a pitch-synchronous analysis. This is similar to work by Scarr [132]. Tillman *et al.* [149] describe an approach based on features convenient to machine analysis, but linguistically oriented. There are five features and (due to redundancy) seven possible combinations of these, which lead to segments by indirect segmentation. Recognition is based on the occurrence of sequences of segments. Von Keller [159, 160] reports an approach which involves formant tracking as one of the required operations. Initial segmentation into voiced fricative, unvoiced fricative, plosive, nasal, and vowel-like is achieved on the basis of two spectral slope measures (over-all slope, and slope below 1 kHz), amplitude, rate of change in amplitude, and a nasal indicator. Part of the decision pattern consists of discrete features, referring to the segment pattern (somewhat like Dersch's machine) and part is continuous, based or key values of F_1 and F_2. Recognition rates around 95% are reported. Velichko and Zagoruyko [157] describe a scheme which recognizes segments on the basis of five frequency band parameters, and uses another form of segment synchronization (*cf.* Vicens, above) in the decision process. Single-speaker recognition scores of around 95%, for a 203-word vocabulary are reported.

There is one final study that falls in a class almost by itself. Hillix and his colleagues [61, 62] report the use of six nonacoustic measures in a speech recognizer. These measures were designed to reflect physiological descriptions of speech and used special instrumentation. Single-speaker recognition rates were reported as 97%. Recognition of one speaker, on the basis of data obtained from three others, gave recognition scores of 78-86%. Approaches based on physiological models of any kind are rare. It is difficult to manage the required characterization, or the alternative of direct measurement. Analysis, at the physiological model level has been confined to basic speech research. As Jakobson, Fant, and Halle remark in their book, there is a hierarchy of relevance in speech transmission: perceptual, aural. acoustical and articulatory. The last level is least relevant to identification of sounds. There may, in fact, be more than one way of operating at the articulatory level to produce the aural or perceptual effect. A good example of this is the perceptual similarity of the effect of lip rounding and pharyngealization.

5.3 Discussion

Richard Fatehchand [35] was able to write that there was, in 1960, no machine capable of dealing with continuous speech. Lindgren [89] five years later was able to say the same. It is not possible to repeat the statement a third time. The scheme developed at Stanford by Vicens [158] based on carlier work by Reddy [121-123] undoubtedly deals with connected speech and has been used for connected utterances forming commands to a robot arm, and to a desk calculator program [58, 99] in both of which individual words had to be recognized in order to understand the command. The acoustic task is eased by using tightly defined simple language syntax and easily segmented "key-words." such as "block." Vicens notes one of the major subproblems to be error recovery following failure to find a word boundary. A film (*Hear, Here* available front Cine-Chrome Laboratories Inc.. 4075 Transport Avenue, Palo Alto, California) shows a speaker telling a computer equipped with an eye and an arm to pick up blocks (the "robot arm" referred to above).

The progress is more in terms of excellence than in terms of a radical new approach though "segment synchronization" is both novel and effective. Perhaps our definitions are progressing. A short utterance is different only in a small degree from the utterance of a long word, and a number of' recognizers have been constructed which essentially recognize phoneme strings rather than, or as well as, words. The extended version of AUDREY [28] recognized a small number of continuant phonemes, though the output was in terms of words. Sakai and Doshita's [128] system was intended to accept "conversational" speech using unlimited vocahulary, and designed to type a phonetic representation, thus segmenting into phonemes, This machine has subsequently led to a digit recognizer [15]. Vicens dismisses the segmentation controversy saying that, rather than arguing as to whether phonemes, syllables. or words should be chosen, he uses all of these at various stages of the recognition process. The present author suggests that the only essential "segmentation" is into a unit of meaning, be it a word, phrase, or (for some purposes) a phoneme .

It is difficult to show, evidence of consistent progress in automatic speech recognition comparable to that in automatic speech generation. There is a lack of unity still, 10 years after Fatehchand [35] first noted it. The practical steps are best implied in terms of a "then and now" comparison, and sometimes it is hard to see that the steps have been forward, or even very practical. Real progress has been, and largely still is, confined to fundamental speech studies, which often have no direct, immediate significance to the engineer's goal of building a viable ASR machine. McLucas, President of the Mitre Corporation, speaking at the IEEE Systems Science and Cybernetics Dinner. 1967, remarked that an engineer, is "driven by a desire to see something work." A friend of mine has an equally apt definition, "an engineer is someone who can do for a quarter what any fool can do for a buck." Machines may or may not work, but few are engineered in the sense that my friend implies they should be. Industrial organizations have become very wary of speech recognition as a profitable exercise. Dreaming of profits, or at least product lines, within a few years, they feel hurt when their investment in research fails to turn up the final answer, and hurriedly stop work, trying to forget it was ever started. When they succeed in forgetting. they start again, and have to start again at the beginning. Surprisingly few workers concern themselves with practical economics on top of all their other problems. Vicens suggests a figure of $50,000 as a goal for a speech input terminal. One machine already mentioned [Sections 5.2.2(d)], that has repeatedly been overlooked in the literature, deserves mention in this connection. DAWID I, built in 1 year—1963—at the University of Bonn [149] represented a significant advance in terms of practicality versus performance. This machine, consisting of a surprisingly small amount of hardware, could recognize twenty Italian words—ten digits and ten commands. Recognition scores are unpublished, it seems, but even a nonspeaker of Italian could achieve a satisfying 90% or more on first acquaintance with the machine.

Probably the most important recent step has been not in ASR, as such, but in the related field of visual pattern recognition, typified by the work of Bloom and Marril [6], Guzman [48, 49], Clowes [18, 19], and Evans [31, 32]. Their contribution to pattern recognition in general terms is important, because it. is aimed at the problem of representation; and it is certainly practical because it works. They have approached the problem of describing visual fields, and visual patterns to be recognized, in terms of a common "figure description language." Such a language is able to describe scenes and objects in a general way by identifying primitives (the basic visual features, analogous to the features that may be extracted by neural systems), and expressing relationships between these primitives in a hierarchical manner. Such a hierarchical system allows any desired complexity to be described, as required by the recognition task, in terms of a small set of rules. We have already noted the generative power of rules in speech synthesis. This approach is used for the visual pattern recognition (of blocks), in the Stanford Hand-Eye project noted above [99], and is now starting to be used in ASR [59, 60], as well as in psychological modeling of human pattern recognition processes [145].

One final problem has been touched on above, and was strongly underlined by Lindgren in his admirable (though somewhat American) survey of the machine recognition of human language [89]. Quoting Sayre [131], he refers to a deep obscurity,

> ... an unclarity about the nature of the human behavior we are trying to simulate. We simply do not understand what recognition is. And if we do not understand the behavior we are trying to simulate, we cannot reasonably hold high hopes of being successful in our attempts to simulate it.

Recognition implies some degree of understanding, or at least some specific response related to the identity of the spoken input. If we build a phoneme recognizer, we are building an automatic phonetician or at the very best, a typist who cannot spell. Phoneticians require considerable training, in addition to that required to recognize and generate speech. If the machine is to do something, as a result of our speaking to it (other than to make a representation of the noises we made), it must, at the very least, recognize words. In practice it wil need to recognize phrases, sentences, and ultimatelv, if Licklidcr's "chemical thing" is to fulfill its intended function, whole conversations The latter surely implies real understanding—a process far deeper than mere identification of words or phrases. Understanding is an active process of reconciling objects, ideas, and relations in the real world with the objects, relations, and ideas implied by the denotations of the words and phrases. The difference between what we may term "identification-type recognition" and "understanding-type recognition" is similar to the difference between provability and truth. Understanding and truth both require an appropriate representation of the real world, and the ability to test it against reality. This is the monumental problem we must ultimately face.

5.4 Conclusions

Although it is difficult to show evidence of consistent progress comparable to that in machine-generated speech, a "then and now" comparison reveals that progress has actually occurred. In 1961 Marri1 [93] suggested four main problem areas in "speech recognition": segmentation, choice of measurements, choice of decision procedure, and handling large amounts of data. It is perhaps in the last area that least progress has occurred, despite the upsurge of computer processing since, in order to automate data handling, one must know precisely what must he done with the data—which is just the whole point of the research; we are not sure. However, recent machines have been more thoroughly tested, perhaps, than their predecessors. Decisiom procedure problems seem to have lost their sting, perhaps partly because decision theory has advanced in the period, and in any case many schemes use decision procedures tailored to their special conditions that are simple yet effective. Simple voting schemes have been found very useful. Adequate segmentation has been demonstrated and other schemes have managed to bypass conventional segmentation This latter looks a promising approach. The variety of features researched is imnpressive—every scheme having its own set. The performance of linguistically oriented and arbitrary features has been compared. and the difference found to be slight, for single speaker recognition. Perhaps this indicates that we have not yet solved the problem of measurement, reinterpreted as the problem of feature selection. We do, however, have usable means of controlling machines by human voice, though cost reduction will prove an important problem for the immmediate future in some approaches.

Perhaps the most significant progress has been not so much a step forward, as a change of direction allowing new attitudes and definitions Instead of talking about "measurement" and "decision procedures" we now talk about "features" and "problems of representation." It is here, in the area of information handling, and data representation (including the use of interactive systems), that the next major set problems lies, and progress will have a more *general* significance than in the past. Progress in speech recognition procedures is of importance in its own right, but problems are now being faced which are common to larger areas of computer science.

6. Progress toward Man-Machine Interaction Using Speech at the University of Calgary

6.1 Introduction

There are three projects at The Universitv of Calgary directly related to the development of a man-machine interface using speech. Two of these are ASR research projects, and the third a Computer Aided Instruction (CAI) project. One ASR project is aimed at two goals: the investigation of basic procedures for use in ASR machines, and the production of a limited objective voice-operated controller (see Section 5.2.2(d), and also Hill [59] and Hill and Wacker [60]). The project is presently concerned with developing computer-based methods for speech analysis, using a PDP-8/I computer; with the construction of a simple voice-operated controller, ESOTerIC II (Experimental Speech Operated Termina1 for Input to Computers II), which represents a linear and digital microcircuit hardware embodiment of the theoretical developments arising from ESOTerIC I; and with synthesis of speech. The second ASR project is presently concerned with establishing a theoretical basis for, and extending the application of, the phase-dependent assymetry measure, first used in SHOEBOX as an indication of voicing, and found to have value as a vowel discriminant [20]. Recent results of this work are available as a thesis [13]. and represent a theoretical advance which so far has not been put into practice. The third project, Computer-Aidcd Instruction, under Hallworth. is a practical applications project, with one speech-related aim—the use, and development of a practical man-machine interface using speech. Clearly, there is an intimate relationship among the three projects, but this section will emphasize the third.

6.2 The Computer-Assisted Instruction (CAI) System

The requirement for speech output at a CAI terminal raises many problems general to speech output. Such a terminal must be inexpensive; it is only one among perhaps fifty terminals serviced by the same computer; it requires random access to a large number of messages, preferably composed of connected speech; it must work in real time; and it must be as natural as possible. A random access tape recorder was considered, but there are disadvantages associated with such a solution. The message set is limited; the correct tape must he mounted and indexed prior to a lesson; access may be delayed during rewind—an uncontrolled factor in the learning situation; and, although the capital cost could, perhaps, be kept down (but a comparison of the cost of random access slide projectors with standard slide projectors suggests it could not), there is the considerable reliability/maintenance problem common to most electromechanical devices. These disadvantages are multiplied by any increase in the number of terminals. For this reason it has been decided to experiment, from the start, with messages synthesized from information stored within the computer system. Ultimately speech messages will be synthesized entirely rule, at or just before the time they are needed. Initially, however. speech will be synthesized in a manner similar to that used for the IBM 7772 Audio Response Unit; i.e., using stored control signals and a similar bit rate, but using a parametric resonance analog synthesizer instead of a channel vocoder synthesizer. The synthesizer to be used is the latest PAT, built in linear and digital microcircuits and driven digitally. The means of connecting PAT to the PDP8/I computer has appeared above [Fig. 11, Section 4.2.2(b)]. The use of such means of speech output overcomes all the disadvantages listed above for the tape recorder approach. A number of important advantages also accrue. The same basic information may be used to generate messages in which the intonation pattern varies. Since the intonation pattern is an essential component of the meaning of an utterance, this advantage is important quite apart from the improvement in naturalness. Furthermore, the speech is held in digital form like the rest of the CAI program, so that the distribution and exchange of CAI programs between different centers is simplified—cards or digital tape alone will suffice. Clearly, some national, or better still international. agreement on synthesizer standards is economically important for this advantage to be fully realized. (In this connection, it should be mentioned that there is a user group for the Glace-Holmes resonance analog synthesizer. Details are available from Weiant Wathen-Dunn at the Air Force Cambridge Research Laboratory, Bedford, Massachusetts.)

The present CAI setup at Calgary is a pilot system to be used for research into some unsolved problems of computer-aided education and man-machine interaction, many of which arise from our lack of knowledge concerning the educational process in general; to gain experience and gather data, which will be invaluable when a larger, operational, system is built; to discover problems which have not yet been thought of; and to show what can be done with present techniques. The system consists of a PDP8/I, in time-shared configuration, driving four terminals each having a Teletype keyboard/page-printer and a random access slide projector. Initially, a separate PDP8/I has been set up with a PAT synthesizer for experimental purposes, though others will be incorporated into the main system by the end of this year. Speech may be generated locally (at the computer) and, if required for a remote location, can be transmitted over the telephone lines being used for the data connection, preceded and followed by suitable control characters, for output control. Some work on speech by rule has already been carried out [see above, Section 4.2.2(b), especially Fig. 10]. Development of software for the new work is now in progress.

Speech input for the CAI system is already in pilot operation on the separate PDP8/I using an ESOTerIC II. In later work, an ESOTerIC II device may be placed at each terminal (whether remote, or local) and unique eight-bit codes assigned to each recognition possibility,, allowing voice output data to be handled in the same way as Teletype response codes. Many problems are anticipated, but it is expected that the existence of an established teaching situation, with adequate feedback, will allow problems such as imperfect recognition, lack of understanding on the part of the student concerning the operation of the system, and other similar problems. to be overcome. Some pilot experimcnts on recognizing unknown speakers, and training non-technical staff [59] have been encouraging. With a CAI system, and a true speech interface, more extcnsivc and more representative experiments may be carried out. The only unexpected result of this work would be no unexpected results! Figure 12 shows ESOTerIC I being used for voice control of a text -editing pprogram based on the PDP-7 and graphic display at Cambridge University Mathematical Laboratory. The figure shows a hand-held microphone, though a boom microphone was normally used. The component cost of ESOTerIC II is of the order of $1000, but this could be reduced for non-rescarch applications.

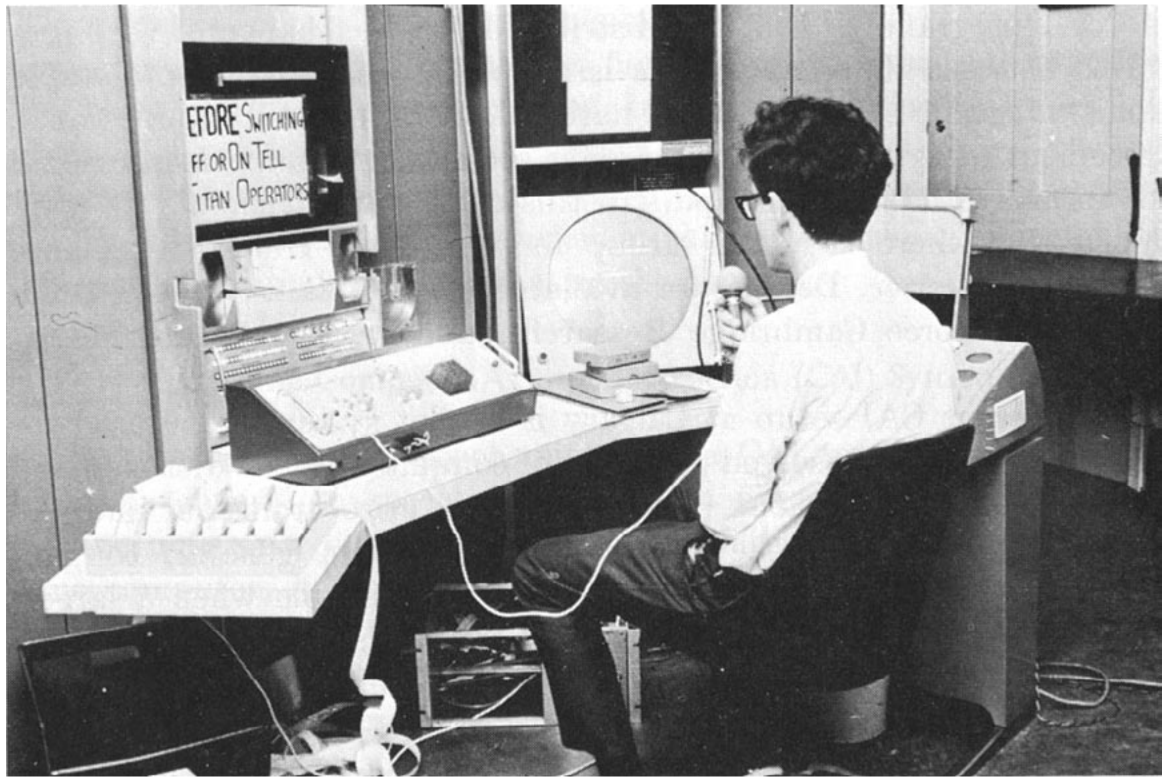

FIG. 12. Voice control of the PDP-7 text editor program

6.3 Conclusion

The foregoing remarks briefly recount the practical steps that are being taken towards a man-machinc interface using speech, at The University of Calgary. Work on both recognition and synthesis of speech is being carried out. The results of this work are being applied, in practical ways, to a man-machine-interface using speech for CAI work. In a CAI situation the shortcomings of such an interface, using our present speech capability, are most likely to be overcome. The human factors experience gained in this early system should prove invaluable in designing applications which place a greater demand on the speech interface.

7. Applications

7.1 Introduction

The previous section has preempted one major potential application of the man-machinc interface using speech—namely in automated audiovisual display systems. Such systems themselves may be applied in a wide variety of ways [107]: to teaching, to mass screening (census taking and medical screening for example), to interaction with computers, to Air Traffic Control (controllers must speak to aircraft, and such an interface would allow this speech to be a means of data. entry to the ATC computer, reducing the over-all load on the controllers), and to therapy or prosthesis for various forms of disability or injury. This is a good reason for conducting research in the context of such a system. Some general advantages and disadvantages of a speech interface have been covered in Section 2 of this article. Some particular advantages of using a speech interface in such audiovisual systems have been covered elsewhere [59] and also, in the context of a CAI situation, in the preceding section.

7.2 Specific Examples

7.2.1 Speech Input

A “speech-recognizer,” or “voice-operated controller” is a more or less sophisticated version of an acoustically operated controller. Probablv the first application of acoustical control was to switch a tape recorder on or off, depending on whether there was a signal to record or not. The difference between “something” and “nothing” was decided on the plausible basis of short-term mean power. Some forms of burglar alarm are operated by a similar device, on the grounds that noises in empty factory, at night, need investigating. These are not really different from Paget’s toy dog Rex. which responded to its name, because a sensor “recognized” the high frequency at time end of “Rex.” Two acoustically controlled telephone answering machines in the, United Kingdom use the human voice as a signal source. One, by Shipton Automation, is called the Telstor, and responds in much the same way as the noise switches mentioned above, except that a counter is incorporated. The correct code (three separate counts) allows messages to be replayed, and then erased, from a remote telephone. A device by Robophone is similar, but measures the duration of “silence” (as the operator counts under his/her breath) to obtain different numbers for use in the code. Some telephone-answering machines may be remotely operated by tones, but (in the absence of a TOUCH-TONE® phone) the operator, out on his travels, must carry a tone generator. Machines which allow the voice to be used offer a clear advantage for the present.

In Electronics for May 15, 1967, Dorset Industries Inc., of Norman, Oklahoma, were reported to be offering a teaching machine for $350, together with an optional voice input costing a further $100, which allows the eight inputs to be stimulated by eight different sounds. Chiba [15] reports that the phonetic typewriter, developed at the Nippon Electric Company, has led to a digit recognizer, which is now being used for voice dialing on a small telephone exchange. SHOEBOX (Dersch’s machine [25]), when demonstrated at the New York World’s Fair in 1962, was used to operate a calculating machine. Dersch, who left IBM—after working on SHOEBOX—to form his own company, offered a voice-sensitive switch, which would respond to a human voice, even when the voice was 95 dB below the ambient noise. A device called VOTEM (Newell [106]) at Standard Telecommunication Laboratories, in the United Kingdom, is a Morse-code operated type writer which uses the human voice as a signal source. Use of Morse-code makes the device cheap and universal. Though somewhat slow (about two-thirds the speed of a good typist) it could allow a seriously disabled person to type personal correspondence and perhaps even do useful work—both of these activities contributing immeasurably to the person’s confidence and self-respect. Vicens applied his recognition procedure to providing input for the robot arm, and a desk calculator. Lavoie [80] refers to an RCA voice operated “stockbroker” which learns 28 command words for each new speaker, and then allows him to use the words to carry out stock transactions by phone. Another RCA device for recognizing the digits of zip codes has just completed field trials in Philadelphia, in a parcel-sorting task.

These are some of the ways in which voice-operated acoustical controllers have been applied, it will be noted that the earliest devices are analogous to the first light-operated devices, which responded to the presence–absence of light; such devices subsequently progressed to counting and then coding. The analogous progression may now be observed in the development of sound-operated devices designed for speech. The analogy is incomplete because there is no general requirement to recognize machine-generated speech (the acoustic equivalent to printed words). Instead, sound-operated devices are in the process of jumping a logical gap, by going straight to the recognition of humanly generated symbols, without the benefit of experience gained on machines designed to recognize machine-produced symbols (synthetic speech). This phase of recognizing humanly generated symbols has only just begun for light-operated devices, even though they must have enjoyed a greater investment than ASR devices because of their success in the previous phase.

7.2.2 Speech Output

On the synthesis side, some devices have found more general application. The two IBM audio response units have been mentioned above. Speaking clocks, which may be dialed from any telephone and then generate a message such as, “At the next tone the time will be exactly three twenty-seven ... (peep),” have been widely used for years; the principle is exactly the same as that of the IBM 7770 ARU—namely the assemblage of prerecorded words. There was, on the London (United Kingdom) underground railway, a recorded message warning passengers to “Stand clear of the doors, please.” An unconfirmed report concerns a device in the United States of America which in corporates a voice output, printed display, and a typewriter. Apparently this device has been used with some success in helping autistic children to communicate at a very simple level. It produces a story, one letter at a time, chanting each letter until the appropriate key on the typewriter is struck.

7.3 Discussion

As we look to the immediate future, the requirement for communication, and hence two-way operation, looms large. The doors in the underground system still closed, even if a passenger shrieked. Mechanical sensors were used to avoid the obvious consequences of such lack of feeling, and perhaps the equivalent will always be necessary, as a fail-safe mechanism for most devices. In the future, however, in many applications of either synthesis, or recognition of speech, the use of either one will require the other. Anyone who has tried using a TOUCH-TONE® phone to pass information will appreciate the difficulty of remembering to what point, in a desired sequence, keypressing has progressed; the uncertainty about what question elicited the present answer; and the annoyance of being able to hear the answers in a replay of a recorded transaction, without being able to identify the sounds generated at the TOUCH-TONE® end. No doubt these problems diminish with practice.

One interesting class of applications for speech-operated and speech generating devices is in prosthetic aid for various disabilities. Speech output would be of obvious advantage to a blind person, whether as a computer-programmer, or as a "listener" in the audio-output section of a computerized library of the future. Speech recognizers are highly relevant to the deaf. It is a sad comment on the profit motive that so little of real value to the deaf has been produced in hardware, though instruments used in ASR research would at least be better than the present aids used in teaching the profoundly deaf. Two devices have been produced with the deaf in mind in Fant's laboratory in Stockholm. One gives a simple display of the varying energy spectrum of speech and may be "frozen" at any moment to give a section for study. The other gives a measure of pitch variation—of particular importance to deaf persons whose poor intonation conveys an impression of stupidity to uninformed listeners. These are indirect applications of ASR work. Cues, found important in recognition by machine, may be useful information for deaf persons learning to speak. As a more direct application, consider the utility of a device which responded to a deaf person's name. It could respond slightly to any sudden change in the acoustic environment, and much more to a sound like the deaf person's name—and perhaps to a motor horn as well. Coupled to a tactile communicator, worn round the waist like a belt, and designed to give indication of the direction of the sound, it would prove an immense boon to the deaf. Acoustic controls for severely disabled persons have already been mentioned, in connection with VOTEM (Section 7.2.1).

It is possible to summarize all applications of a man-machine interface using speech as being to "make machines more intelligent." An acceptable definition of intelligence is "the ability to handle information effectively." This implies goals, in order that effectiveness may be gauged in terms of goal achievement. It also implies gathering, sorting, selecting, structuring/storing (representing), and communicating information—these comprise "handling." This definition admits that any system having a goal, and the ability to handle information in a manner suitable for attaining that goal, is intelligent. Presumably handling more information to achieve more goals, or handling less information to achieve the same goal(s), or handling the same information to achieve more goals, would each count as demonstrating greater intelligence. However, a machine which was able to communicate more effectively would, by the above definition, be more intelligent. The main purpose of a man-machine interface using speech is to allow more effective communication. In thus making machines more intelligent, we are making them more useful. Perhaps, in the process, we shall make them more like ourselves, despite the opposing view expressed by Pierce [116].

7.4 Conclusions

It seems that a small number of applications of limited usefulness have been tried with ASR equipment; applications of machine generated speech have been wider and more successful, but such output means are not yet in general use. One significant reason for the lack of speech interface hardware is probably lack of demand, due to lack of awareness. A kind of "credibility gap" which causes potential users to dismiss the subject as too fanciful. The idea that machines have switches, knobs, keys, and means for printing or even plotting is generally accepted; the idea that machines might draw, or recognize pictures is gaining credence; the idea that machines might speak and obey commands is sufficiently new that it is seldom considered as a practical proposition. We are now in a position to experiment with prototype speech interfaces, say for CAI work, and attack the credibility gap.

8. Concluding Remarks

An attempt has been made to keep the various sections of this report fairly self-contained, and there is little that can be added by way of further specific conclusions. An optimistic picture has been painted, based upon a total range of progress, in areas relevant to the subject of the report, that is impressive. There remain serious obstacles to the kind of speech interface that would really facilitate communication between men and machines. Providing computers with adequate means of speech production and perception is a small part of the total problem. This is well illustrated by considering the situation where all computers were provided with speech output and with expert, incredibly fast typists, who could enter a perfect written copy of any utterance as it was made. There are still the problems of natural language processing, of techniques for computer-aided learning, and of information retrieval, to mention only the first that spring to mind. The really serious problems are now problems of information handling, even though we have not completely solved the problems at the acoustic level. In fact a solution of some of the problems at the higher levels is almost certainly essential to what might appear, superficially, as acoustic recognition of speech, since speech perception ultimately involves the whole of language. However, we must use the equipment we now have, for in some restricted situations we may already reap benefits. At the same time we shall gain an appreciation of the real restrictions on present techniques, and be able to formulate more pertinent questions on which to base further research.

ACKNOWLEDGMENTS

The author can only claim the errors and omissions as his own. Many people, knowingly and unknowingly, have contributed ideas and information, by publication and discussion, too many for individual recognition, but the author is deeply grateful to those concerned. The author wishes to thank especially Professor Abercrombie and the members of the Edinburgh University Phonetics Department (as it was then) for providing the facilities and experience necessary to the author's work on synthesis. The author also gratefully acknowledges the financial support given to his research by the National Research Council of Canada.

REFERENCES

1. Alter, R., Utilization of contextual constraints in automatic speech recognition. *IEEE Trans. Audio Electroacoustics* **AU-16**, 6-11 (1968).
2. Antony, J., and Lawrence, W., A resonance analogue speech synthesiser, *Proc. 4th Int. Cong. Acoust., Copenhagen, 1962*, G.43.
3. Balandis, L. S., Sceptron: a sound operated fiber-optic "Brain cell." *Electron. World* **69** (1963).
4. Bar-Hillel, Y., *Language and Information: Selecteds Esays and their Theory and Application*. Addison-Wesley, Reading, Massachusetts, 1964.
5. Bezdel, W., and Bridle, J. S., Speech recognition using zero-crossing measurements and sequence information. *Proc. Inst. Elect. Eng.* **116**, 617-624 (1969).
6. Bloom, B. H., and Marril, T., The Cyclops-2 system. Tech. Rept. TR65-RD1, Computer Corp. of Amer., Cambridge, Massachusetts, 1965.
7. Bobrow, D. G., and Klatt, D. H., A limited speech recognition system. *Proc. AFIPS Fall Joint Computer Conf., San Francisco, 1968* **33**, Part 1, pp. 305-318. Thompson, Washington, D.C., 1968.
8. Bobrow, D. G., Hartley, A. K., and Klatt, D. H., A limited speech recognition system II. Bolt, Beranek, & Newman Rept. No. 1819, Job No. 11254, under Contract NAS 12-138, Cambridge, Massachusetts, 1969.
9. Borko, H. (ed.), *Automated Language Processing—The State of the Art*. Wiley, New York, 1967.
10. Broadbent, D. E., *Perception and Communication*. Pergamon, Oxford, 1958; also Scientific Book Guild (Beaverbrook Newspapers), London, 1961.
11. Buron, R. H., Generation of a 1000-word vocabulary for a pulse-excited vocoder operating as an audio response unit. *IEEE Trans. Audio Electroacoustics* **AU-16**, 21-25 (1968).
12. Buron, R. H., Audio response unit connected to a digital computer. Div. 1.1/1. *Nat. Ass. Telecommun. Engrs. Conf., Madrid, 1965*, pp. 1-13.
13. Bryden, B., Speech recognition by machine. M.Sc. Thesis, Dept. of Electrical Engineering, The Univ. of Calgary, Alberta, Canada, February, 1970.

14. Cherry, C., *On Human Communication*. M.I.T. Press, Cambridge, Mas sachusetts, and Wiley (Studies in Communication Series), New York, 1957 also Science Edition, Inc., New York, 1961.
15. Chiba, S., Machines that "hear" the spoken word. *New Scientist* **37** (552) 706-708 (1967).
16. Chomsky, N., *Syntactic Structures*. Mouton, 's-Gravenhage, The Netherlands seventh printing 1968.
17. Chomsky, N., and Halle, M., *The Sound Pattern of English*. Harper, New York, 1968.
18. Clowes, M. B., Pictorial relationships-a syntactic approach, in *Machine Intelligence* (B. Meltzer and D. Michie, eds.), Vol. 4, pp. 361-383. Edinburgh Univ. Press, Edinburgh, Scotland, 1969.
19. Clowes, M. B., Perception, picture processing and computers, in *Machine Intelligence* (N. L. Collins and D. Michie, eds.), Vol. 1, pp. 181-197, Oliver & Boyd, Edinburgh, Scotland, 1967.
20. Corner, P., The use of waveform asymmetry to identify voiced sounds. *IEEE Trans. Audio Electroacoustics* **AU-16**, 500-506 (1968).
21. Cooper, F. S., Liberman, A. M., and Borst, J. M., The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proc. Nat. Acad. Sci. U.S.A.* **37**, 318-325 (1951).
22. Denes, P., The design and operation of the mechanical speech recogniser at University College, London. *J. Brit. Inst. Radio Eng.* **19**, 219-23 (1959).
23. Denes, P., On the motor theory of speech perception, in *Proc. Int. Cong. Phon. Sci.* (E. Zwirner and W. Bettige, eds.), pp. 252-258. S. Karger, Basle, Switzerland, 1965.
24. Denes, P., and Mathews, M. V., Spoken digit recognition using time frequency pattern matching. *J. Acoust. Soc. Amer.* **32**, 1450-1455 (1960).
25. Dersch, W. C., Decision logic for speech recognition. IBM Tech. Rept. 16.01.106.018, San José, California 1961.
26. Dixon, N. R., and Maxey, H. D., Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Trans. Audio Electroacoustics* **AU-16**, 40-50 (1968).
27. Dudley, H., Riesz, R. R., and Watkins, S. A., A synthetic speaker. *J. Franklin Inst.* **227,** 739-764 (1939).
28. Dudley, H., and Balashek, S., Automatic recognition of phonetic patterns in speech. *J. Acoust. Soc. Amer.* **30**, 721-732 (1958).
29. Dunn, H. K., and Barney, H. L., Artificial speech in phonetics and com- munications. *J. Speech Hearing Res.* **1**, 23-39 (1958).
30. Estes, S. E., Kerby, H. R., Maxey, H. P., and Walker, R. M., Speech synthesis from stored data. *IBM J. Res. Dev.* **8**, 2-12 (1964).
31. Evans, T. U., A heuristic program to solve geometric analogy problems. *Proc. AFIPS Spring Joint Computer Conf., 1964* **25**, 327-338. Spartan, New York, 1964.
32. Evans, T. G., A grammar based pattern analysing procedure. Proc. *IFIP Conf., Edinburgh, Scotland, 1968*. Preprints H152-H157, August (1968).
33. Fant, C. G. M., *Acoustic Theory of Speech Production*. Mouton, 's-Gravenhage, The Netherlands, 1960.
34. Fant, C. G. M., Martony, J., and Rengman, U., OVE II synthesis strategy. *Proc. Stockholm Speech Commun. Seminar* **II**, F5. Speech Transmission Lab., Royal Inst. Technol., Stockholm, 1962.
35. Fatehchand, R., Machine recognition of spoken words. *Advan. Computers* **1**, 193-229 (1960).

36. Feigenbaum, E. A., Artificial Intelligence: themes in the second decade. Stanford Univ. Artificial Intelligence Res. Proj. Mem. No. 67 (1968).
37. Flanagan, J. L., *Speech Analysis, Synthesis, and Perception*. Springer, New York, 1965.
38. Fodor, J. A., and Katz, J. J., *Structure of Language: Readings in the Philosophy of Language*. Prentice-Hall, Englewood Cliffs, New Jersey, 1964.
39. Forgie, J. W., and Forgie, C. D., A computer program for recognising the English fricative consonants /f/ & /0/, *Proc. 4th mt. Cong. Acoust., Copenhagen, 1962*.
40. Forgie, J. W., and Forgie, C. D., Results obtained from a vowel recognition computer program. *J. Acoust. Soc. Amer.* **31**, 1480-1489 (1959).
41. Foster, J. M., *Automatic Syntactic Analysis*. Macdonald, London, 1970.
42. Frick, F. C., Research on speech recognition at Lincoln Laboratory, *Proc. Seminar on Speech Compression and Coding*. PB 146624. Air Force Cambridge, Res. Lab., LG Hanscom Field, Bedford, Massachusetts, 1959.
43. Fry, D. B., Theoretical aspects of mechanical speech recognition. *J. Brit. Inst. Radio Eng.* **19**, 211-218 (1959).
44. Giuliano, V. E., Analog networks for word association. *IEEE Trans. Mil. Electron.* **MIL-7**, 221-234 (1963).
45. Gold, B., Word recognition computer program. Tech. Rept. 452, Mass. Inst. Technol., Cambridge, Massachusetts, 1966.
46. Green, B. F., Wolf, A. K., Chomsky, C., and Laughery, K., BASEBALL: an automatic question answerer. *Proc. Western Joint Computer Conf.* **19**, 219-224 (1961); also in *Computers and Thought* (E. A. Feigenbaum and J. Feldman, eds.), pp. 207-216. McGraw-Hill, New York, 1963.
47. Guelke, R. W., and Smith, E. D., Distribution of information in stop consonants. *Proc. Inst. Elec. Eng.* **110**, 680-689 (1963).
48, Guzman, A., Scene analysis using the concept of model. Sci. Rept. No. 1, Computer Corp. of Amer. Contract (to AFCRL) AF 19 (628)-5914, 1967.
49. Guzman, A., Decomposition of a visual scene into three-dimensional bodies. *Proc. 1968 AFIPS Fall Joint Computer Conf.* **33**, Part 1, pp. 291-304. Thompson, Washington, D.C., 1968.
50. Haggard, M. P., and Mattingly, I. G., A simple program for synthesising British English. *IEEE Trans. Audio Electroacoustics* **AU-16**, 95-99 (1968).
51. Halle, M., Hughes, G. W., and Radley, J-P. A., Acoustic properties of stop consonants. *J. Acoust. Soc. Amer.* **29**, 107-116 (1957).
52. Halle, M., Stevens, K., Speech recognition: A model and a program for research. *IRE Trans. Inform. Theory* **IT-8**, 155-159 (1962).
53. Halliday, M. A. K., I*ntonation and Grammar in British English*. Mouton, 's-Gravenhage, The Netherlands, 1967.
54. Harris, K. S., Cues for the discrimination of American English fricatives in spoken syllables. *Language Speech* **1**, Part 1, 1-6 (1958).
55. Hayes, D. G. (ed.), *Readings in Automatic Language Parsing*. American Elsevier, New York, 1966.
56. Hecker, M. H. L., Studies of nasal consonants with an articulatory speech synthesizer. *J. Acoust. Soc. Amer.* **34**, 179-188 (1962).
57. Hill, D. R., An abstracted bibliography of some papers relative to automatic speech recognition. Standard Telecommun. Labs., Harlow, England. Tech. Memo. 522, 1967.
58. Hill, D. R., Automatic speech recognition-a problem for machine intelligence, in *Machine Intelligence* (N. L. Collins and D. Michie, eds.), Vol. 1, pp. 199-266. Oliver & Boyd, Edinburgh, 1967.

59. Hill, D. R., An ESOTerIC approach to some problems in automatic speech recognition. *Int. J. Man-Machine Studies* **1**, 101-121 (1969).
60. Hill, D. R., and Wacker, E. B., ESOTerIC IT-An approach to practical voice control: progress report 69, in *Machine Intelligence* (B. Meltzer and D. Michie, eds.), Vol. 5, pp. 463-493. Edinburgh Univ. Press, Edinburgh, 1969.
61. Hillix, W. A., Use of two non-acoustic measures in computer recognition of spoken digits. *J. Acoust. Soc. Amer.* **35**, 1978-1984 (1963).
62. Hillix, W. A., Fry, M. N., and Hershman, R. L. Computer recognition of spoken digits based on six non-acoustic measures. *J. Acoust. Soc. Amer.* **38**, 790-797 (1965).
63. Holmes, J. N., Research on speech synthesis carried out during a visit to the Royal Institute of Technology Stockholm November 1960 to March 1961. Post Office Eng. Dept., England, Rept. No. 20739 (1961).
64. Holmes, J. N., Mattingly, I. G., and Shearme, J. N., Speech synthesis by rule. *Language Speech* **7**, Part 3, 127-143 (1965).
65. Hubel, D. H., and Wiesel, T. N., Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106154 (1962).
66. Huggins, W., A theory of hearing, in *Communication Theory* (W. Jackson, ed.), pp. 363-380. Butterworth's, London, 1953.
67. Hughes, 0. W., and Hemdal, J. F., Speech analysis. Purdue Research Foundation, Lafayette, Indiana, Tech. Rept. TR-EE65-9, 1965.
68. Hyde, S. R., Automatic speech recognition literature survey and discussion. Res. Dept. Rept. No. 45, Post Office Res. Dept., Dollis Hill, London, 1968.
69. IBM, IBM 7772 Audio Response Unit. Form A27-2711-0 IBM Systems Library.
70. IBM, IBM 7770 Audio Response Unit Models 1, 2, and 3. Form A27-2712-0 IBM Systems Library.
71. lIes, L. A., Speech synthesis by rule. Use of Holmes, Mattingly, and Shearme rules on PAT. *Work in Prog. No. 3* (E. Uldall and A. Kemp, eds.), pp. 23-25. Dept. of Phonetics and Linguistics, Univ. of Edinburgh, Scotland, 1969.
72. Jakobson, R., Fant, C. G. M., arid Halle, M., *Preliminaries to Speech Analysis*. M.I.T. Press, Cambridge, Massachusetts, 1951; eighth printing, 1969.
73. Jones, D., *An Outline of English Phonetics*. W. Heifer, Cambridge, England, 1918; ninth edition, 1960.
74. Jones, D., *The Phoneme: Its Nature and Use*. W. Heifer, Cambridge, England, 1959; second edition, 1962.
75. Kiss, G. R., Networks as models of word storage, in *Machine Intelligence* (N. L. Collins and D. Miehie eds.), Vol. 1, pp. 155-167. Oliver & Boyd, Edinburgh, 1967.
76. Kiss, G. II., Steps towards a model of word selection, in *Machine Intelligence* (B. Meltzer and D. Miehie. eds.), Vol. 5, pp. 315-336. Edinburgh Univ. Press, Edinburgh, 1969.
77. Ladefoged, P., Some possibilities in speech synthesis. *Language Speech* **7,** Part 4, 205-214 (1964).
78. Ladefoged, P., and Broadbent, D. E., Perception of sequence in auditory events. *Quart. J. Exptl. Psychol.* **12**, Part 3, 162-170 (1960).
79. Ladefoged, P., Private communication 1969.
80. Lavoie, F. J., Voice actuated controls. *Mach. Des.* **42,** 135-139 (January 22, 1970).
81. Lawrence, W., The synthesis of speech from signals which have a low

information rate, in *Communication Theory* (W. Jackson, ed.), pp. 460-471. Butterworth's, London, 1953.

82. Lea, W. A., Establishing the value of voice communication with computers. *IEEE Trans. Audio Electroacoustics* **AU-16**, 184-197 (1968).

83. Lehiste, I., Acoustical characteristics of selected English consonants. AD 282 765, Armed Services Tech. Inform. Agency, 1962.

84. Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H., What the frog's eye tells the frog's brain. *Proc. Inst. Radio Eng.* **47**, 1940-1951 (1969); also in Embodiments of Mind (collected works of W. S. McCulioch), pp. 230-255, M.I.T. Press, Cambridge, Massachusetts, 1965.

85. Liberman, A. M., Some results of research on speech perception. *J. Acoust. Soc. Amer.* **29**, 117-123 (1957).

86. Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P., and Cooper F. S., Minimal rules for synthesising speech. *J. Acoust. Soc. Amer.* **31**, 1490-1499 (1960).

87. Liberman, A. M., Cooper, F. S., Harris, K. S., and MacNeilage, P. F., A motor theory of speech perception. *Proc. Stockholm Speech Commun. Seminar II*, D3, Speech Transmission Lab,, Roy. Inst. Technol., Stockholm, 1962

88. Liljencrants, J. C. W. A., The OVE III speech synthesiser. *IEEE Trans. Audio Electroacoustics* **AU-16**, 137-140 (1968).

89. Lindgrcn, N., Machine recognition of human language; Part I: Automatic Speech Recognition; Part II: Theoretical models of speech perception and language; Part III: Cursive script recognition. *IEEE Spectrum* **2**, No. 3, 114-136 (1965); No. 4, 44-59 (1965), No. 5, 104-116 (1965).

90. Lindgren, N., Speech-man's natural communication. *IEEE Spectrum* **4**, No. 6, 75-86 (1967).

91. Lindsay, R. K., Inferential memory as the basis of machines which understand natural language, in *Computers and Thought* (E. A. Feigenbaum and J. Feldman, eds.), pp. 217-233. McGraw-Hill, New York, 1963.

92. Lisker, L., and Abramson, A. S., Stop categorization and voice onset time, in *Proc. 5th Int. Congr. Phon. Sci.* (E. Zwirner and W. Bettige, eds.), pp. 389-391. S. Karger, Basle, 1964.

93. Marril, T., Automatic recognition of speech. *IRE Trans. Human Factors Electron.* **HFE-2,** 35-38 (1961).

94. Martin, T. B., Zadell, H. J., Nelson, A. L., and Cox, R. B., Recognition of continuous speech by feature abstraction. Tech. Rept. TR-66-189, Advanced Technol. Lab., Defense Electron. Prod., RCA, Camden, New Jersey, 1966,

95. Mattingly, I. G., Synthesis by rule of prosodic features. *Language Speech* **9**, No. 1, 1-13 (1966).

96. Mattingly, I. G., Experimental methods for speech synthesis by rule. *IEEE Trans. Audio Electroacoustics* **AU-16**, 198-202 (1968).

97. Maturana, H. R., Uribe, 0., and Frenk, S., A biological theory of relativistic colour coding in the primate retina. *Arch. Biol. Med. Exptl. Suppl.* **1**, 1968.

98. McConologue, K., and Simmons, R. F., Analysing English syntax with a pattern-learning parser. *Commun. Ass. Comput. Machinery* **8**, 687-698 (1965).

99. McCarthy, J. Earnest, L. D., Reddy, D. R., and Vicens, P. J., A computer with hands, eyes, and ears. *Proc. Fall Joint Computer Conf., San Francisco, 1968*, pp. 329-338. Thompson, Washington, D.C., 1968.

100. McElwain, C. K., and Evens, M. B., The degarbler-a programme for correcting machine-read morse code. *Inform. Control* **5**, 368-384 (1962).

101. Meeker, W. F., Parametric recognition of speech sounds. *Proc. Seminar Speech Compression Coding*. PB146624, Air Force Cambridge Res. Lab., LG Hanscom Field, Bedford, Massachusetts, 1959.

102. Meetham, A. R., and Hudson, R. A. (eds.), *Encyclopaedia of Information, Linguistics, and Control*. Pergamon, Oxford, 1969.

103. Miller, G. A., *Language and Communication*. McGraw-Hill, New York, 1951.

104. Morgan, 0. E., Private communication, 1969.

105. Newcomb, W. B., A dictionary programme for speech recognition, Unpublished, private communication, July 30, 1964.

106. Newell, A. F., VOTEM-A voice operated typewriter. 53rd Exhibition, The Institute of Physics and the Physical Society, London, 1969.

107. Newman, E. A., and Scantlebury, R., Teaching machines as intelligence amplifiers. Rept. Auto 31, Nat. Phys. Lab. Teddington, England, 1967.

108. O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., and Cooper, F. S., Acoustic cues for the perception of initial 1w, j, r, I/ in English. *Word* **13**, 24-43 (1957).

109. Olson, H. F., and Belar, H., Phonetic typewriter III. *J. Acoust. Soc. Amer.* **33**, 1610-1615 (1961).

110. Olson, H. F., and Belar, H., Syllable analyzer, coder, and synthesiser for the transmission of speech. *IRE Trans. Audio* **AU-10**, 11-17 (1962).

111. Olson, H. F., and Belar, H., Performance of a code operated speech synthesiser. *Proc. 16 Ann. Meeting Audio Eng. Soc., October, 1964*.

112. Otten, K. W., The formant vecoder and its use for automatic speech recognition. *Proc. 3rd Int. Conf. Acoust.* (L. Cremer, ed.), Elsevier, Amsterdam, 1961.

113. Otten, K. W., Approaches to the machine recognition of conversational speech. *Advan. Computers* **11**, 127 (1971). (this volume)

114, Paulus, E., Bibliography on automatic speech recognition. Tech. Rept. TR 25.064, IBM Lab., Vienna, 1966.

115. Petrick, S. R., and Griffiths, T. V., On the relative efficiencies on context free grammar recognisers. *Commun. Ass. Comput. Machinery* **8**, 289-299 (1965).

116. Pierce, J. R., Men, machines, and languages. *IEEE Spectrum* **5**, No. 7, 44-49 (1968).

117. Potter, R. K., The technical aspects of visible speech. *J. Acoust. Soc. Amer.* **17**, 1-89 (1946), also Bell Monograph 1415, Bell Telephone Labs., Murray Hill, New Jersey, 1946.

118. Potter, R. K., Kopp, 0. A., and Kopp, H. 0., *Visible Speech*. Dover, New York, 1966.

119. Presti, A. J., High Speed sound spectrograph. *J. Acoust. Soc. Amer.* **40**, 628-634 (1966).

120. Purton, R. F., An automatic word recogniser based on autocorrelation analysis. *Colloq. Some Aspects Speech Recognition Man-Machine Commun.* Reprint, Inst. of Elec. Eng., London, England, 1968.

121. Reddy, D. R., An approach to computer speech recognition by direct analysis of the speech wave. Ph.D. dissertation, Tech. Rept. CS49, Stanford University, Stanford, California, 1966.

122. Reddy, D. R., Computer recognition of connected speech. *J. Acoust. Soc. Amer.* **42**, 329-347 (1967).

123. Reddy, D. R., and Vicens, P. J., A procedure for the segmentation of connected speech. *J. Audio Eng. Soc.* **16**, 404-411 (1968).

124. Reeds, J. A., and Wang, W. S. Y., The perception of stops after s. *Phonetica* **6**, 78-81 (1961).

125. Ross, P. W., A limited-vocabulary adaptive speech-recognition system. *J. Audio Eng. Soc.* **15**, 415-418 (1967).

126. Sager, N., Syntactic analysis of natural language. *Advan. Computers* **8**, 153-188 (1967).

127. Sakai, T., and Inoue, S. I., New instruments and methods for speech analysis. *J. Acoust. Soc. Amer.* **32**, 441-450 (1960).

128. Sakai, T., and Doshita, S., The automatic speech recognition system for conversational sound. *IEEE Trans. Electron. Comput.* **EC-12**, 835-846 (1963).

129. Salton, G., *Automatic Information Organization and Retrieval*. McGraw-Hill, New York, 1968.

130. Salton, G. A., Ph.D. program in Information Retrieval. *Commun. Ass. Comput. Machinery* **12**, 110-117 (1969).

131. Sayre, K. M., *Recognition: A Study in the Philosophy of Artificial Intelligence*. Notre Dame Univ. Press, Notre Dame, Indiana, 1965.

132. Scarr, R. W. A., Zero-crossings as a means of obtaining spectral information in speech analysis. *IEEE Trans. Audio Electroacoustics* **AU-16**, 247-255 (1968).

133. Scarr, R. W. A., Normalization and adaption of speech data for automatic speech recognition. *Int. J. Man-Machine Studies* **2**, 41-59 (1970).

134. Schroeder, M. R., Vocoders: analysis and synthesis of speech. *Proc. IEEE* **54**, 720-734 (1966).

135. Schouteri, J. F., Cohen, A., and 't Hart, J., Study of time cues in speech perception. *J. Acoust. Soc. Amer.* **34**, 517-518 (1963).

136. Sebesteyen, 0., Automatic recognition of spoken numerals. *J. Acoust. Soc. Amer.* **32**, 1516 (1960).

137. Shearme, J. N., and Leach, P. F., Some experiments with a simple word recognition system. *IEEE Trans. Audio Electroacoustics* **AU-16**, 256-261 (1968).

138. Simmons, R. F., Burger, J. F., and Schwarcz, R. M., A computational model of verbal understanding. *Proc. AFIPS Fall Joint Computer Conf., San Francisco, 1968* **33**, 411-456.Thompson, Washington, D.C., 1968.

139. Sitton, G., Acoustic segmentation of speech. *Int. J. Man-Machine Studies* **2**, 61-102 (1970).

140. Stevens, K. N., On the relations between speech movements and speech perception. *Z. Phonetik, Sprachwissenschaft and Kommunikationsforschung* **21**, 102-106 (1968). Presented by invitation at seminar on speech production and perception, Pavlov Inst. of Physiol., Leningrad, 1966.

141. Stevens, K. N., Toward a model for speech recognition. *J. Acoust. Soc. Amer.* **32**, 47-55 (1960).

142. Stevens, S. S., *Handbook of Experimental Psychology*. Wiley, New York, 1951; eighth printing, 1966.

143. Strevens, P. D., and Antony, J., The performance of a 6-parameter speech synthesiser (a lecture demonstration). *Proc. 8th intern. Cong. Linguists, pp. 214-215, Oslo Univ. Press, Oslo, Sweden, 1958.*

144. Strevens, P. D., Spectra of fricative noises in human speech. *Language Speech* **3**, Part 1, 32-49 (1960).

145. Sutherland, N. S., Outlines of a theory of visual pattern recognition in animals and man. *Proc. Roy. Soc. (London) B***171**, 297-317 (1968).

146. Taylor, E. F., A skimmable report on ELIZA. Educ. Res. Center Rept., Mass. Inst. Tech., Cambridge, Massachusetts, 1968.

147 Teacher, C. F., Kellet, H. 0., and Focht, L. R., Experimental, limited vocabulary speech recognizer. *IEEE Trans. Audio Electroacoustics* **AU-15**, 127-130 (1967).

148. Thorne, J. P., Bratley, P., and Dewar, H., The syntactic analysis of English by machine, in *Machine Intelligence* (B. Meltzer and D. Michie, eds.) Vol. 3, pp. 281-297. Edinburgh Univ. Press, Edinburgh, Scotland, 1968.

149. Tillman, H. G., Heike, G., Schrielle, H., and Ungeheuer, C., DAWID I- em Betrag zur automatisehen "Spraeherkennung," *Proc. 5th Int. Cong. Acoust., Liege, 1965.*

150. Traum, M. M., and Torre, E. D., An electronic speech recognition system. *Proc. 1967 Spring Symp. Digital Equipment Corporation User Soc.*, The State University, New Brunswick, New Jersey, pp. 89-92. Digital Equipment Corporation User Society. Maynard, Massachusetts,

151. Truby, H. M., *A note on invisible and indivisible speech*. Proc. 8th. Int. Cong. Linguists, pp. 393-400, Oslo Univ. Press, Oslo, Sweden, 1958.

152. Turing, A. M., Computing machinery and intelligence. *Mind* **59**, 433-460 (1950), also in *Computers and Thought* (E. A. Feigeribaum and J. Feldman, eds.), pp. 11-35. McGraw-Hill, New York, 1963.

153. Uldall, E., and Antony, J. K., The synthesis of a long piece of connected speech on PAT. *Proc. Stockholm Speech Commun. Seminar II F8, Speech Transmission Lab., Roy. Inst. Technol., Stockholm, 1962.*

154. IJhr, L., and Vossler, C., Recognition of speech by a computer program that was written to simulate a model for human visual pattern perception. *J. Acoust. Soc. Amer.* **33**, 1426 (1961).

155. Uhr, L., and Vossler, C., A pattern recognition program that generates, evaluates, and adjusts its own operators. *Proc. Western Joint Computer Conf.* **19**, 555-570 (1961), also in *Computers and Thought* (E. A. Feigenbaum and J. Feldman, eds.), pp. 251-268, McGraw-Hill, New York, 1963.

156. University of Michigan, Automatic Speech Recognition. Univ. of Michigan Eng. Summer Conf. Automatic Speech Recognition, Serial 6310, Vols. I and IT (1963).

157. Velichko, V. M., and Zagoruyko, N. G., *Automatic recognition of 200 words. Int. J. Man-Machine Studies* **2**, pp. 223-234 (1970)

158. Vicens, P., Aspects of speech recognition by computer, Ph.D. Dissertation, CS 127, Dept. of Computer Sci., Stanford Univ., Stanford, California, 1969.

159. von Keller, T. G., An on-line system for spoken digits. *J. Acoust. Soc. Amer.* **44**, 385 (1970).

160. von Keller, T. G., Automatic recognition of spoken words. *75th Meeting Acoust. Soc. Amer., Ottawa, 1968.*

161. von Kempelen, W., Speaking Machine, *Philips Tech. Rev.* **25**, 48-50 (1963/ 1964).

162. von Senden, M., *Space and Sight: The Perception of Space and Shape in the Congenitally Blind Before and After Operation*. Free Press, Gleneoe, Illinois, 1960.

163. Walker, D. E., English pro-processor manual (revised). Rept. SR-132, Inform. System Language Studies No. 7, MITRE Corp., Bedford, Massachusetts, 1965.

164, Walker, D. E., Chapin, P. 0., Geis, M. L., and Gross, L. N., Recent developments in the MITRE syntactic analysis procedure. Rept. MTP-1 1, Inform.

Systems Language Studies No. 11, MITRE Corp., Bedford, Massachusetts, 1966.

165. Wang, S-Y., and Peterson, G. E., Segment inventory for speech synthesis. *J. Acoust. Soc. Amer*. **30**, 1035-1041 (1958).

166. Wells, J. C., A study of the formants of the pure vowels of British English. Dept. of Phonetics Progress Rept., under Contract AF EOAR 62-116 USAF European Office OAR, Univ. College, London, 1963.

167. Whifield, I. C., and Evans, E. F., Responses of auditory cortical neurones to stimuli of changing frequency, *J. Neurophysiol*. **28**, 655-672 (1965).

168. Willshaw, D. J., and Longuett-Higgins, H. C., The Holophone-recent developments, in *Machine Intelligence* (B. Meltzer and D. Michie, eds.), Vol. 4, pp. 349-357. Edinburgh Univ. Press, Edinburgh, 1969.

169. Wiren, J., and Stubbs, H. L., Electronic binary selection system for phoneme classification. *J. Acoust. Soc. Amer.* **28**, 1081-1091 (1956).

170. Woods, W. A., Procedural semantics for a question answering machine. *Proc. AFIPS Fall Joint Computer Conf., San Francisco, 1968*, 33, Part 1, 457-471. Thompson, Washington, D.C., 1968.

Author Index

Subject index