13

AUTOMATIC SPEECH RECOGNITION: A PROBLEM FOR MACHINE INTELLIGENCE

DAVID R. HILL
STANDARD TELECOMMUNICATION LABORATORIES LTD
HARLOW

SUMMARY

Speech recognition by machine has not yet been achieved because no suitable specification of the recognition process has been formulated for the machine. The author outlines the disturbances and constraints found in speech, and goes on to a description of the structure implied by the constraints. This description is a prerequisite of speech recognition for two reasons, first to describe general speech structure in terms which allow knowledge of it to be built into the machine as an aid to the recognition process, secondly to allow a good enough description of the input signal for it to lead to a minimum set of recognition possibilities which includes likely alternatives. The outline is drawn of a hypothetical machine to recognise speech, comprising a basic recogniser working on short segments of acoustic waveform only, on to which may be added further structures to use knowledge of speaker characteristics, speech statistics, syntax rules, and semantics, in order to improve the recognition performance. Some detailed examples of possible structures are given. Finally there is a brief description of work in progress at Standard Telecommunication Laboratories towards implementing a basic recogniser of the type suggested.

INTRODUCTION

In a recent survey paper Lindgren (1965) says, '. . . the immediate aim of building automata that can recognise speech seems somewhat in abeyance'. Miller, on the same subject, says that the engineers concerned have a right to be discouraged. Why have two or more decades of intensive research been rewarded with such apparent lack of success? It is not due to lack of means of analysis for the acoustic signal. What is difficult is telling the machine what to do with the results of the analysis, for most machines working today work because someone has not only been able to tell them what to do but has been able to do so within the limitations of the machine. Suppose one tried to implement a recogniser by telling the machine to store every new pattern it encountered together with a label telling it what word or words the pattern represented, with the intention of recognising an arbitrarily large vocabulary for an arbitrarily large proportion of the total population of speakers. It would never work for a number of good reasons. It would have an endless need to store new data, and would therefore be so enormous that, even if it could be built and even if a sufficient number of the parts functioned correctly at the same time, it would be quite uneconomic, and would constantly be unable to make decisions through lack of data. The problem is to find a description of the recognition process which is sufficiently economical for it to be built into a machine to render the machine capable of performing a useful job at a competitive price.

Much early work proceeded on the assumption that there were, in the acoustic signal representing speech, invariant data groupings representing the phonemes used by linguists to categorise speech sounds in each particular language. It was hoped that these could be extracted to allow straightforward classification into the same categories, and hence into words. One important result of the work on the problem has been to show that this is not true, that in many cases there simply is not sufficient information in the acoustic signal representing a word to determine completely the word it is intended to represent (for example, Miller 1962). This is another reason why our hypothetical monster recogniser would not work, unless it worked on comparatively long portions of the input signal at something like the sentence level which would aggravate the storage problem. Miller (1964a) has conservatively estimated that 10^{20} sentences could be constructed, and these would take 1000 times the age of the earth even to utter! He (Miller 1965), Fry (1956) and Denes (1959), as well as others, have taken great pains to emphasise that the recognition of speech by human beings depends not only on the acoustic signal reaching the ear but on the whole structure of language. That man is the only biological system constructed to use language the way he uses it is suggested by the failure of attempts so far to teach our most intelligent non-humans to use language. It is very likely that our intelligence arises from our ability to use language rather than the other way round (Miller 1964b). One is tempted to suggest that the title of this paper could read, 'Intelligence—a problem for machine language' and still remain of great interest to the assembled company, for language is the 'stuff' of thought, and without thought we should be incapable of reaching those higher levels of abstraction, analogy, and generalisation which are the root of our biological eminence.

This then is the problem—what constitutes a sufficiently economical specification of the recognition process to enable a machine to be built to implement it? The real reason for this paper is the assumption that economy requires an ability to generalise from incomplete data, an ability to adapt to new environments, an ability to abstract from large amounts of data, an ability to retrieve information with minimum cost, an ability to use to advantage the constraints of the environment, and an ability to benefit from mistakes and successes. These are key problems in machine intelligence studies. The virtue of the automatic speech recognition problem (referred to henceforth as the ASR problem) is that it brings together a machine for which there is a real commercial requirement, and principles of machine intelligence which will be needed for the implementation. This happy conjunction provides a basis for discussion.

AUTOMATIC SPEECH RECOGNITION—THE PROBLEM

We have said that the problem is to specify the recognition process economically. This involves knowing about the speech signal we wish to recognise in order to specify the processing required by the machine. The most striking aspect of speech, which achieves its full impact when ASR is considered, is appalling complexity. The most effective overall picture of speech in relation to the machine and the recognition process is by means of a block diagram such as appears in Fig. 1. The connecting arrows stand for 'entails', and the circles, with numbers in, are conventional threshold gates, used to compound the entailments. In the top left area of the figure is shown the nature of the speech we wish to recognise. The bottom left area shows the main characteristics of a machine which might be built to recognise the speech. The right-hand side of the figure outlines some of the operations that the machine would need to perform.

The purpose of the next two sections is to amplify the diagram. The fifth section will describe briefly some work which is being carried out at Standard Telecommunication Laboratories towards implementing a real machine, and the final section will contain conclusions.

SPEECH

General

Speech is generated by a human being for the purpose of communicating with another human being, and is usually transmitted and received as an acoustic signal. This, like most other signals, is subject to various sorts of noise and distortion. These disturbances degrade the information in the signal. Because it is generated by a human being it is subject to a number of constraints due to his physiology, his intention and his previous linguistic experience. Finally, that a human being can recognise speech, despite the disturbances, is evidence of redundancy in the transmission/recognition process.

Disturbances

The first disturbance is channel noise, which comprises reduction of information by added noise, by attenuation of the signal (equivalent to

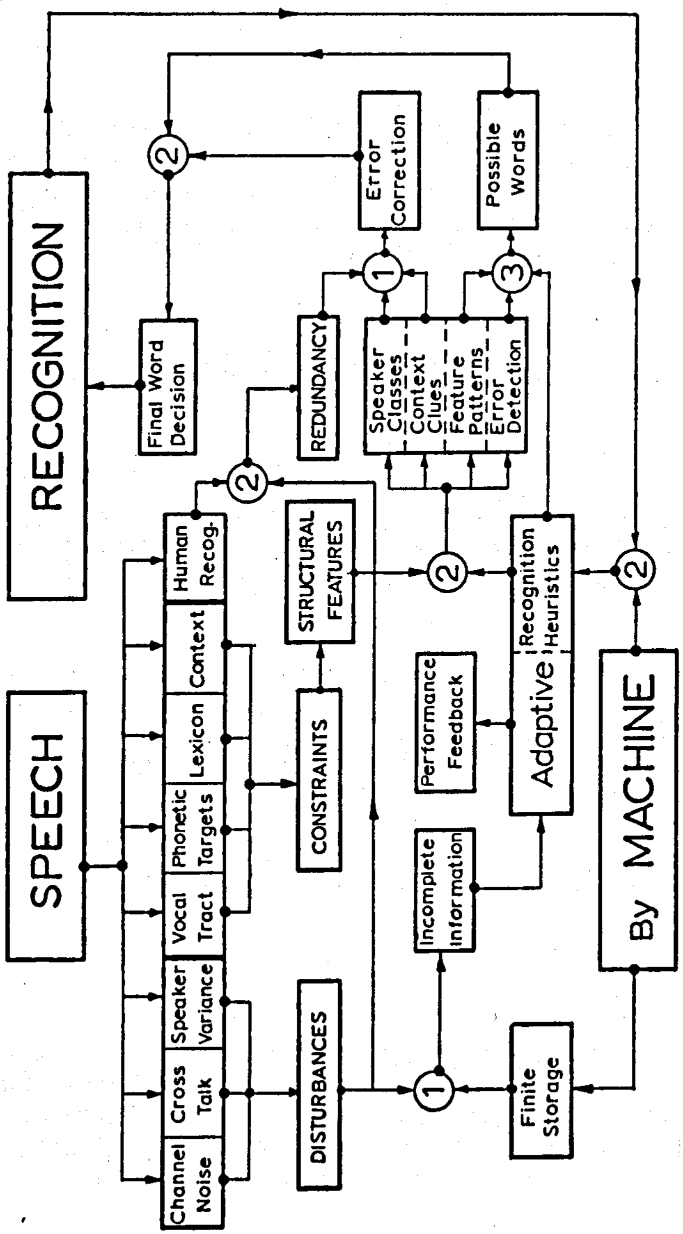

FIG. 1. The ASR problem.

adding noise), or by distortion of the signal because of the nature of the transmission path. If, as is highly probable for ASR, the speech is transmitted through a telephone link the problems of noise and distortion can be quite severe and include noises due to handling the handset, clicks and hisses in the speech band, limitation of the bandwidth to the range between 300 and 3400 cps, and pre-emphasis of the signal. With an air-only transmission path there is distortion because of the acoustics of the environment, and attenuation.

Another form of disturbance is cross-talk. In the most severe form this is referred to as the cocktail party problem. The effect on recognition performance is considerable, for the competing noise has the same general form as the signal. Miller (1947) has shown that a babble of four or more voices is one of the most effective masking noises available, given a single communication channel. The cross-talk mean level on commercial telephone systems (assuming no cocktail party at sending or receiving end!) is normally no worse than 30 dB (voltage) below the signal, which could give a target for machine recognition.

Finally, and this form of disturbance is the one most often considered in ASR studies, there is a great deal of speaker variation. Even the same speaker attempting an identical series of utterances will produce signals which differ noticeably from utterance to utterance (Fry 1959), and the words he utters in isolation will be different from the words he utters as part of connected speech (Truby 1958). Between different speakers there is considerable variation, and, for example, the vowel and fricative categories of one speaker are very likely to overlap different vowel and fricative categories for other speakers (Fry 1959, Strevens 1960). He may also use the same vowel categories in a different way because he has a different accent.

Constraints

At the lowest, most general, level there are constraints on the signal structure because of the nature of the generating apparatus. The human vocal apparatus consists in essence of a main tube through which a modulated stream of air may be blown, and a secondary tube which may be connected in parallel with the second half of the main tube. The configuration of the main tube may be varied by constricting the walls, moving the tongue—or parts of it—back and forth, and up or down. The secondary tube may be connected by lowering the velum. The tongue may produce a sufficiently narrow constriction to produce turbulent noise (fricative sounds) or even cut off the flow of air altogether (stop sounds). If the velum is lowered nasalisation occurs which, with the main air flow stopped by the tongue, leads to nasal consonants. The modulator at the lower end of the main tube may impress periodic (voiced) or aperiodic (aspirated) variation in the air flow, or may stop it altogether to give a glottal stop. In the case of voiced modulation the energy distribution, in terms of frequency versus amplitude, will show a harmonic series which falls off at about 12 dB per octave above

500 cps. The acoustics of this relatively complex vocal system have been the subject of classic studies by Fant (1960). Briefly, however, they are as follows. The source energy excites standing waves in the tract just as standing waves are formed in an organ pipe. These resonances modify the energy distribution in the basic spectrum leading to bands of enhanced energy in the frequency domain called *formants*. In theory there is an infinite series of these *formants* but in practice only five are important (neglecting nasal sounds) and of these only the lowest three are significant for intelligibility. At the frequency of F1 (the first and lowest *formant*) the standing wave pattern is $\frac{1}{4}$ of a full wavelength. For F2, and F3, the standing waves are $\frac{3}{4}$ and $\frac{5}{4}$ of a full wavelength respectively. For an average vocal tract, 17.5 cm long, this leads to formant frequencies of about 500 cps, 1500 cps and 2500 cps. Inside the tract pressure and velocity have a phase difference of 90°, and at the frequency of a formant, there is always a pressure minimum at the lips, while the opposite is true at the larynx (the modulator). The simplest rule governing the relations between vocal tract configurations and F-patterns is that if a uniform tube is constricted at a place where one of its formants has a velocity maximum the formant frequency will decrease; a constriction at a velocity minimum leads to an increase. In addition to the static acoustic constraints on the signal due to the vocal tract, there are dynamic and neurological constraints. The parts of the tract have inertia and particular forms of attachment, and the messages which activate movement have limitations. Thus the rate at which changes occur, and the possible configurations, are restricted in a manner general to human speakers. The fastest changes are due to the velum and larynx, next those due to tongue movements, next those due to pharyngeal constriction, with the slowest changes resulting from lip and jaw movements.

Constraints at a less general level result from the acquired articulatory habits of the speaker, ignoring the differences due to speaker physiology. In attempting an utterance, the speaker will be aiming at certain phonetic targets related to the language and dialect he speaks. These habits of utterance are acquired at an early age, and are difficult to break, as evidenced by the difficulty that speakers of English have in producing really good French vowels.

A further constraint on the speaker is his vocabulary, or lexicon. Speakers of the same language, even the same dialect, will use different vocabularies. Miller (1951) has shown that, taking 50 per cent success as the intelligibility criterion, there is an 18 dB variation in speech power threshold when the choice set for intelligibility testing is changed from 2 to 256 words. Normal people have vocabularies well in excess of this, some estimates put the size as high as 30 000 words, but this figure is drastically reduced if derivatives are not counted as different words.

The last set of constraints are contextual. Some are due to the syntax and semantics of particular utterances, some are situational in that the topic of conversation will restrict the likely use of the vocabulary. The contextual constraints are hard to partition, since they range from quite general rules about how utterances may be constructed from lexical items, and what is meaningful, to quite particular restrictions on what is likely to be talked about next, and what interpretation should be placed on what was said a few moments ago in the light of what is said now.

Redundancy

That a human being can recognise speech despite the disturbances is evidence that there is redundancy in this form of communication. The redundancy is embodied in all aspects of speech and results from the constraints which exist. Certain signal-component structures simply are not allowable because of the generative and phonetic target constraints, and certain recognition possibilities are not likely due to the lexical, syntactic and semantic constraints. The redundancy can be increased by restricting the lexicon and imposing stricter rules on the construction of utterances from the lexical items.

A very simple recognition machine might work on input strings of the form *number*, *operator*, *number*, allowing only twenty or so recognition possibilities. Syntactic constraints would then inhibit the recognition of an operator for the last item, while semantic knowledge would allow the machine to reject the recognition of 'zero' as the last item if the operator happened to be 'divide by'. This level of using syntactic and semantic redundancy already exists in computer programs—it is extending the principle to the whole of language which is likely to prove so difficult.

Structural Features

If there are constraints on the signal, it will have structure; the more restrictive the constraints the less will be the variety of the structure. In speech we wish to describe the structure for two reasons; first so that a general knowledge of the structure may be built into the recogniser to allow it to utilise the redundancy to aid the recognition process, secondly so that the input may be adequately described in terms relevant to the recognition process. The term *structural features* refers to the descriptors or 'clues' which define the structure, so this section is concerned with the description of speech, and the form and nature of the descriptors. Whatever level of description is aimed at, the evidence for the description of a particular utterance must either come from the acoustic signal itself, or be built into the recogniser and be accessed by clues from the acoustic signal. However useful and easy it is to describe an utterance in terms of non-acoustical features, it does not help ASR if these cannot be related to the acoustic signal.

There are two ways of arriving at a description of the structure (analytic and synthetic) and two ways of formulating the description (segmental and parametric). The analytic determination of structure depends on analysis and classification of data from signals exhibiting the structure; the synthetic approach depends on setting up models of the structure to form the basis of synthesis, followed by testing the goodness-of-fit of the results of synthesis with reality. Very often both methods are used together. The difference between the segmental and parametric formulations is that the former divides the signal with respect to time, using multi-dimensional descriptors, whereas the latter considers temporal variations in unidimensional descriptors. In general the two will be strongly related, for the segmental descriptors may be compounded from the parameters.

In speech there are only two significant segmental approaches below the level of words, one relating to phonemic segmentation, and the other to the detection of distinctive features. Let us start by considering these two segmental approaches in relation to phonetic targets. In the former, sounds of a language are grouped into categories called phonemes, and any utterance in the language may then be represented by an appropriate series of these. Individual sounds are classed in the same phoneme category if the substitution of one sound for the other never distinguishes two words in the language, and so a particular phoneme structure is specific to the language it describes. The phonemes may be classed into a number of broader categories, for instance voiced/unvoiced. During the whole of voiced sounds the larynx generates periodic excitation to the vocal tract, which impresses its configuration on the signal radiated from the talker by modifying the energy distribution in the frequency-time domain. From the point of view of intelligibility the emitted signal has three main peaks of energy (in addition to the peak around the fundamental frequency) somewhere in the range from approximately 300 cps to 3500 cps, and these are formants (see also 'Constraints', p. 203). Higher formants may be observed, and, if the nasal passage is connected to the main resonant tract by lowering the velum, more than three formants may be observed in the lower range, but in what follows *formants* refers to the three main formants. The voiced sounds in normally spoken English include all the vowels together with some of the consonants, such as /m, b, w, z/—these latter being examples of nasal, voiced stop, semi-vowel, and voiced fricative consonants respectively. The vowels are characterised by the steady state values of the formants (which in connected speech may never be achieved) and the consonants by the transitions of the formants to or from a characteristic steady state value (this latter, in the case of the stops, being of negligible or non-existent duration) together with the amplitude, type and duration of concurrent hiss or aspiration noise. The other ends of the transitions will depend on the adjacent phonemes. The voiceless sounds, in general, consist of a period of hiss type noise, generated by forcing the breath stream through some constriction in the tract, preceded and/or followed by formant transitions in the tract modified signal. There may be two distinctly different types of hiss, as in the case of /p/ for instance where one type could occur due to the release and relate to the constriction at the lips during the release, and the other could occur after the release and relate to the constriction at the larynx prior to the onset of voicing for the succeeding sound. The place of constriction can vary between these two extremes, and the further down the tract the hiss is produced, the more the tract will be able to impress its configuration on the signal by modifying the energy distribution. The basic spectrum of the hiss is characteristic of the place of constriction. The only consistent difference between voiced stops and voiceless stops lies in the time of onset of voicing (Lisker 1965), though the release is usually much more pronounced in initial and medial voiceless stops than in the corresponding voiced stops. There is a close correspondence between the series of voiced consonants and the series of voiceless consonants, of the stop and fricative varieties, but reasonably intelligible speech may be produced without making the distinction, as in whispered speech. This is another piece of evidence for the existence of redundancy in speech, and is in part due to the redundancy of coding with respect to voicing.

For this reason among others, many workers prefer a sub-phonemic segmental analysis of speech sounds. An important contribution to this approach was made by Jakobson, Fant & Halle (1961). Since the phonemes may be dichotomised in a number of ways, the features used for the dichotomies may be used to classify the phonemes and fewer units are likely to be necessary. These units they termed *distinctive features.* Each comprised the opposition of two polar qualities of the same category, or the opposition between presence and absence of a certain quality. A concurrent bundle of distinctive features defined a phoneme. The difficulty with the original distinctive features is that they were largely based on articulatory considerations and are not easy to relate to the acoustic signal. However, variations on the theme of distinctive features lie behind quite a few approaches to ASR, and the engineer, free from linguistic inhibitions (but, hopefully, with some linguistic knowledge) happily modifies the idea to his own ends, and can also use the features to classify whole utterances. Hughes (1961), now at Purdue University, developed a machine which, under certain constraints, could recognise nonsense syllables a little better than a human being, using a modified series of distinctive features. These new features are the most explicit example of the structural features with which his section is concerned, and a particular starting set is suggested in 'The hardware' (p. 219).

Parametric descriptions at this level also exist. The best example of acoustical parameters occurs with Lawrence's (1953) Parametric Artificial Talker (PAT) (Antony *et al.* 1962, Ingemann 1960) and its successors, which approach the problem from the synthetic side. The acoustical parameters used are the three formant frequencies, the rate and amplitude of periodic excitation to the tract, the amplitude of aspiration, and the amplitude and frequency region of hiss noise, a total of eight parameters. In their work on speech by rule, using slightly different parameters, Holmes *et al.* (1965) have shown that generative rules in terms of such parameters can lead to highly intelligible speech. He also showed, while at Stockholm working with OVE II—a synthesiser very like PAT—in Fant's laboratory, that real speech could be copied parametrically accurately enough to be almost indistinguishable from the original. Holmes' speech by rule, despite the fact that it was parametric, was based on phonemic segmentation, and the rules took care of the continuity. This supports the view expressed on p. 206 that segmental and parametric approaches are strongly related. Particular configurations of, or changes in, the parameters may constitute an event, and the events may provide a more economical description than the parametric description.

Use of the event approach, which is a variant of the distinctive feature approach, has proved the most successful means of attack on the determination of structural features. By systematically varying the events in speech synthesised partly on the basis of events and partly on the basis of parameters, the workers at the Haskins Laboratories have added extensively to our knowledge of the essential structure of speech. A classic interpretive paper by Liberman (1957) summarises the work up to 1957, and nominates events such as spectral quality of sounds at constant constriction, spectral quality in transient sounds—at or near the time of maximum constriction, transitional events in the parameters—indicating movement of the articulators (transitions), and events characteristic of the introduction of the nasal passage. Further data appears in other papers (for example, Liberman *et al.* 1952, 1955, 1957, Lisker *et al.* 1965).

On the analytic side surprisingly little data has been published. The classic study is that by Potter and his colleagues (1947) at Bell Telephone Laboratories. Wells (1963) has made a study of the formants in British English vowels, Lehiste (1962) published a monumental study of allophonic variations in /l, r, w, y, h/ and included whispered speech, Strevens (1960) has investigated the spectra of fricative sounds, and Green (1958) has made an extensive study of second formant transitions. These are a sample of the most informative.

A further class of features deals with the sequence of events. The incorporation of the time clues has provided a constant hindrance to ASR schemes and at a phonemic level of recognition results in the 'segmentation problem'. Most efforts to take time into account have either done what is effectively a template matching procedure, with some allowance for expanding and contracting the time template, or have attempted to quantise time in a manner exactly related to the spectral changes on the grounds that this leads to an identical scaling regardless of the rate utterance. This is by far the most popular basis and is termed segmentation by a stability criterion. When the signal changes from a previous state, to a sufficient extent, it is presumed that some new phonetic target should be evaluated. What seems to escape most workers is that what is really important are the sequential relationships between the parts of the message. If one expects a sequence ABC, and the B gets lost or changes to D, the sequential relationship 'A before C' is still there, which may be sufficient to recognise the group. This strategy, as well as others, is used by people attempting to decipher illegible hand-writing, but here the process becomes a little more explicit.

One can also notice separation of content and order in children learning to speak. Sometimes parts of the intended sound are lost—but what remains is in the right order, and sometimes all the right sounds are there—but the sequential relationships are distorted. Moray at Sheffield University is investigating the relation of sequence and content information to the information processing abilities of the human, though at the slightly higher level of words. Almost no work has been done on sequential features of words. Marril & Bloom (1963, 1965) have tackled the analogous problem of 'rules for combination' for the primitive features used in their 'CYCLOPS' picture-pattern recogniser. Their work stimulated this author to the idea of sequential features for ASR.

We are now in a position to return to a consideration of the features which describe the action of the vocal apparatus. A considerable upsurge of interest in this topic has taken place recently. What has been said before about approaches and formulation is applicable, but the idea behind the interest is that utterances may be more usefully described in terms of articulatory descriptors than acoustic descriptors. On the analytic side, by relating electromyographic measurements to the production of utterances, and classifying the data, it is hoped to formulate descriptors. Truby (1959) is responsible for some of the best cineradiographic pictures taken of the articulatory process, while MIT, in conjunction with the Haskins Laboratories, are examining both kinds of data. Similar work is also being started at Fant's Laboratory at the Royal Institute of Technology in Stockholm. On the other hand, Ladefoged and his colleagues at the University of California, Los Angeles, are not only taking measurements on speakers (see, for example, Fromkin 1964), they are also attempting to build a controllable physical model of the real vocal apparatus of the human. MIT are using a computer controlled electrical analogue of the physiology of the vocal tract, which has about 15 parameters (Lindgren 1965). In England, Abercrombie (1965) has suggested a preliminary set of parameters for describing the vocal generating process and emphasises the difference between the parametric and segmental formulations. He suggests three broad divisions—respiratory parameters, phonatory parameters and articulatory parameters—and, in the last category, includes velic valve action, tongue body movements, tongue tip movements, lip movements and jaw movements.

The descriptors for the lexicon fall into several categories—statistical, syntactic, and semantic. First there is a straight specification of the words in the lexicon, which for the average talker changes with time. Next there is a frequency of usage of those words, also changing with time. The lexicon also specifies phoneme transition probabilities and in addition there will be word transition probabilities. These comprise the statistics. The last two feature or descriptor categories reach up to the last level of description, namely context, for the words in the lexicon may be categorised into a relatively small number of *parts of speech* and they have *meaning*. The statistics may be measured reasonably easily, though again there is a lack of recent data. Hultzen (1964) has published the most recent study of the transition probabilities of phonemes, based on General American. Ungeheuer, at the University of Bonn, who has made one of the most interesting recognisers so far (Tillman *et al.* 1965), is working on the statistics of German using the whole of Kant's work reduced to punched cards. One difficulty here (working from the printed word) is that there is no ready conversion from the orthography of a language to the spoken word, though dictionaries do specify standard pronunciations, which has enabled Bhimani (1964) to develop rules for the conversion process for some dialects of English. All this is limited to a phonemic segmental formulation and I do not know of any similar work on distinctive features or parameters, except that a dynamic plot of F1 against F2 shows little structuring of data even when duration is used as a weighting factor (Holmes *et al.* 1961).

An interesting illustration of the reality of the characteristic structure of the words of a language is given by the ability of someone like Michael Bentine to produce nonsense utterances which are readily 'identifiable' as belonging to particular languages. Fry *et al.* (1959) and Denes (1959) built such statistics into their phonetic typewriter to demonstrate the resultant improvement in phoneme recognition, with considerable success.

This is the most appropriate stage at which to consider speaker features. Some very successful speaker identification has been achieved by examining details of contour spectrograms (where intensity is given by contour lines rather than intensity of marking), looking, for example, for characteristic shapes of transitions, overall energy distribution and fricative energy distribution, Kersta (1962) originated this 'voice-printing' technique. Sebesteyen *et al.* (1962), at Litton Systems, says features including formant 3 and formant 4 frequencies, rates of change of the formants and first order probability function of the pitch interval are good clues to the speaker. Pruzansky (1963) describes a talker recognition procedure based on comparing the array points in a quantised spectrogram for chosen words, which is literally 'voice-printing'. Speaker features also include the words he uses, and the overall statistics of his acoustic feature production. The speaker features even extend to the syntactic and semantic levels, in the sense that certain speakers tend to say certain things in certain ways. If any features can be used to identify the speaker (or more likely, for reasons of economy, the speaker category) then knowledge of the other features of the speaker, or speaker category, may be used to aid recognition.

Finally, we consider contextual features, which are the descriptors of the rules governing the allowable sequences of words which will constitute *grammatical*, *meaningful* and *relevant* utterances. There are two major ways of looking at the grammar or syntax of a language, either in terms of a phrase structure grammar, or in terms of a transformational grammar. The former is analytic, in the sense that observed data is taken as it stands, but it still has to be fitted to a modelling of the presumed rules of the language syntax and there is a predictive element in it. The latter is much more closely synthetic, for the syntax of the utterance is described in terms of the transformations required to produce the utterance from a *kernel sentence*. Work by Hanne at the University of Michigan, and by Newcomb at General Dynamics, Rochester, New York (neither published), is typical of the application of phrase structure techniques to ASR. In both these studies the string of sound elements in an utterance is used in conjunction with the lexicon to produce possible strings of words, and with each word is associated the part(s) of speech into which the word can be categorised. The various strings of parts of speech which are thereby possible are compared to the general schema of phrase structures and the impossibilities eliminated. Some ambiguity may still remain which can only be eliminated by semantic considerations. The transformational approach can lead to more accurate descriptions of the syntax of the utterance, but it seems much less suitable for actually applying to ASR directly, as there is the problem of deciding what the kernel sentence should be. Work is in progress at MIT Research Laboratory of Electronics under Chomsky (1957) and Yngve (1960), and under Thorne at Edinburgh University in Britain, to quote only a few examples. The meaning of language is, for the ASR worker, the most forbidding hurdle of all, for it implies not just a knowledge of the real world and what sort of things can and cannot happen in it, but also the whole human ability to use a language as a means of thought. Miller's sentence illustrates one of the difficulties. It starts (and this should be spoken rather than written) 'Writing it rapidly with his undamaged hand he saved from destruction the contents of the . . .'. Thus far we have detected no error, though we may be a little puzzled. However, when it ends with 'capsized canoe', we realise that the word at the beginning of the sentence should have been spelt 'righting', at least this is the most probable intention of the speaker. In attempting to use semantic features, which are the descriptors of the world we live in, machine intelligence will face its greatest test.

THE MACHINE AND THE TASK

General Characteristics

At this point the purpose of this report has perhaps been achieved, and the problem for machine intelligence has been formulated. It would not be fair, however, to pose a problem without at least some suggestion of how the problem might be solved, and what steps are being taken towards implementing the proposed solution. It will be necessary to become rather more particular in order to make these suggestions, and it is with some trepidation that a scheme is put forward at all, for there are many question marks and much that ought also to be considered will be ignored. The scheme will consist of suggestions for machine operations, some described in more detail than others. If the underlying assumptions are correct, and the parts could be designed to perform the operations suggested, this would deal with a fairly complex input of connected speech, with allowance for reversion to word by word operation when in difficulties. A primary objective has been to make each descriptive level of operation of the machine as independent of the operation of higher levels as possible. Some degree of recognition may be achieved with a simpler version of the machine, and the higher levels could be added progressively to improve the performance. 'Level' is perhaps a misleading designation for much of the processing will be parallel.

The input to the machine will consist of an acoustic signal representing an utterance. The machine's output will be a code representing the word(s) in the utterance. The input signal is characterised by structural features, embodying redundancy, bearing some relation to the words uttered. The machine may be characterised by a task structure embodying recognition heuristics, and by finite size. If a machine having any degree of generality is to be built with finite storage, especially when there are disturbances on the signal, it will need to be adaptive. Adaption requires performance feed-back to control the adaptive process, so this must be given to the machine, and probably sets a minimum requirement of the recogniser prior to adaption. A block diagram, on which the description is based, is given in Fig. 2.

Description

A commonly used division of pattern recognition machines is into a fixed observation-taking pre-processor, followed by an adaptive or non-adaptive decision-taker. This is a useful division, provided it is understood that where the boundary is drawn depends on which decision is being considered. The only requirement is that adaption in what is being considered as the pre-processor should either consist of logic switching according to previously acquired knowledge, or should be slow enough to be quasi-static with respect to adaption in the decision taker. Each decision stage utilises more knowledge of the constraints in the speech, and, in so doing, reduces the number of bits of information passed on to the next stage.

The speech to be recognised enters the machine by means of a linear transducer (microphone) which reproduces the pressure variations in terms of voltage variations. This signal is analysed by means of filters and special-purpose circuits which determine when certain predetermined features are present and when they are not. These we may call *primitive acoustic features*, or *PAFs* (see 'Structural features', pp. 205-11). They will comprise such features as 'silence', 'relative time of voicing onset', 'vowel quality', 'hiss', 'friction quality', 'relative amplitude', 'relative duration', 'occurrence of transitions', 'voicing', together with some features relevant to speaker identification (see p. 210). There is some evidence that features of this type are extracted by biological systems, though this is not necessarily a good reason for doing the same in a machine. Evans & Whitfield (1961) have demonstrated single cell responses in the primary auditory cortex of the cat to such events as 'tone present', 'tone absent', 'tone starts', 'tone ends', 'tone frequency rising' and 'tone frequency falling'. Hubel & Wiesel (1963) have demonstrated analogous responses in the visual system of cats, and Lettvin and his colleagues (1959) have shown some feature extraction in the frog's visual processing system. Effects have also been observed in psychophysical experiments on humans which could be explained using a similar model.

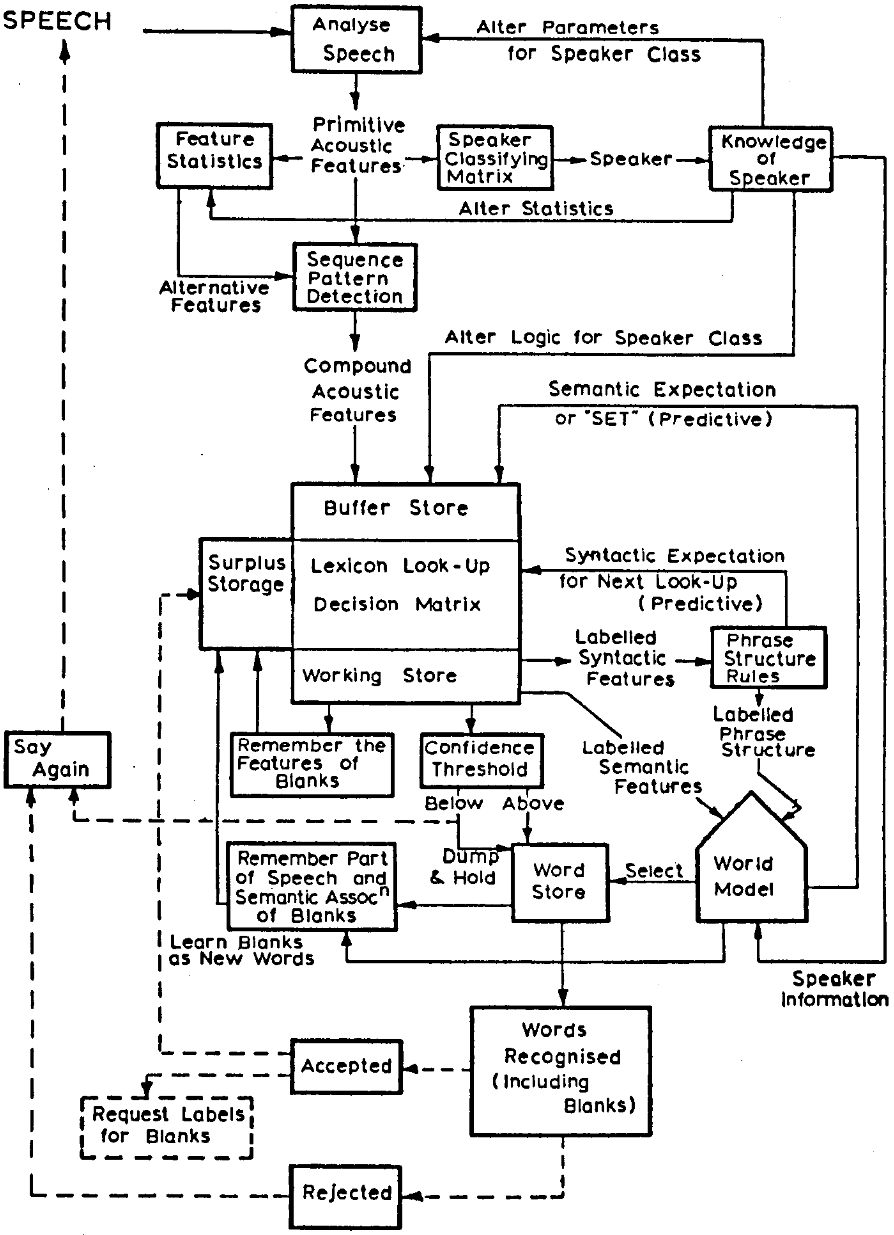
SPEECH
Analyse Speech
Alter Parameters for Speaker Class
Primitive Acoustic Features
Feature Statistics
Speaker Classifying Matrix
Speaker
Knowledge of Speaker
Alter Statistics
Sequence Pattern Detection
Alternative Features
Alter Logic for Speaker Class
Compound Acoustic Features
Semantic Expectation or "SET" (Predictive)
Buffer Store
Surplus Storage
Lexicon Look-Up
Decision Matrix
Syntactic Expectation for Next Look-Up (Predictive)
Labelled Syntactic Features
Phrase Structure Rules
Working Store
Labelled Phrase Structure
Say Again
Remember the Features of Blanks
Confidence Threshold
Labelled Semantic Features
Below
Above
Dump & Hold
Remember Part of Speech and Semantic Assocⁿ of Blanks
Word Store
Select
World Model
Learn Blanks as New Words
Speaker Information
Words Recognised (Including Blanks)
Accepted
Request Labels for Blanks
Rejected

FIG. 2. A hypothetical ASR machine.

The PAFs could provide evidence for recognising speakers. In parallel with the feature extraction process, there could be a predictor. This would embody the statistics of the features, and its predictions would be combined with the features detected in an attempt to take advantage of the known properties of the language to correct errors. The predictor would be very slowly adaptive, if at all, but recognition of the speaker, or speaker type, could lead to a modification of the information used in the predictor. Speaker recognition could also lead to slight alterations in the feature extraction circuit parameters as well as switching of vowel quality channel connections between the detectors and the outputs to allow for dialectical variation in the use made of the vowels. These changes would be based on acquired knowledge of the speaker category characteristics.

Sequential patterns would then be detected in the PAFs on which word decisions were to be based; the output of the detector would be called *compound acoustic features*, or *CAFs.* It seems likely that the CAFs would be rather syllabic in nature and would form the main basis of word decisions. The relation between PAFs and CAFs would be determined as a result of five basic operations on the PAFs—occurrence, simultaneous occurrence, *X or Y*, *X before Y* and iteration. The output would include an indication of the number of occurrences of the CAFs. As indicated on p. 208 some caution would be required in the selection of the particular CAFs to be output. If they are chosen to be too simple they will not be discriminating enough, if they are made too complex then they will be so discriminating that each will be evidence for the occurrence of only one word; either alternative is uneconomical and therefore undesirable. Ideally each should occur in about 50 per cent of the vocabulary, but this is unlikely to be achieved in practice. A further constraint of the CAFs chosen as categories for recognition is the 'necessary and sufficient' condition for word recognition, for however well they are chosen there can still be a very large number of them, and this may be a critical area for the use of adaption to keep the working set as small and useful as possible. What strategy could be used for this is simply not known at present and a great deal of work is required even to assess the feasibility of a manual determination of a sub-set for recognising a small vocabulary. The basis for generating a new CAF would be comparison of the PAFs of a word with the stored CAFs of the words with which it was confused. The merit of the CAF would then have to be evaluated in the light of further new word discriminations which it enabled.

Information concerning the occurrence of CAFs would be passed to the buffer store of the word decision taker. It would be necessary to assume at this stage that any important order information was explicit in the CAFs, and the word decision would therefore be a straightforward maximum likelihood estimate. This is a degenerate form of the Bayesian strategy resulting from the assumptions that correct decisions cost nothing, incorrect decisions all cost the same, and the *a priori* probabilities of the decisions are known. If $P(A)$ represents the absolute probability of A occurring, $P(A/B)$

represents the probability of A occurring—given that B has occurred, E_j is the occurrence of the jth event, and X is a set of elementary observations having binary outcomes, then the maximum likelihood strategy may be summed up in the formula:

$$P(E_j/X)=\frac{P(X/E_j)\ P(E_j)}{P(X)} \tag{1}$$

The realisation of this is covered in a little more detail on p. 221. If the speaker had been recognised, and there was stored knowledge concerning the lexicon of the speaker—either required changes in the items, or required changes in the features expected for the items—the necessary adjustments could be made. In addition certain semantic features would be fed in concerning the associations of the preceding utterance, to bias the decisions taken in favour of words associated with the previous utterance. These would be provided by a later stage (see p. 216). When there was space in the working store, the new information would be fed successively into the working store and continue the process of generating likelihood estimates on the words. When the likelihood of a word or words exceeded a confidence threshold then a maximum detection would be made and the most likely word would be checked by comparing the stored features 'expected' with the features actually in the working store; those items in the working store actually used would be marked. If the marked items could form part of a longer word (and this information could be stored with the word recognised) then a check would be made of the CAFs immediately following to determine whether the longer word was present or not. If the longer word were detected the process would repeat until the longest word admissible had been determined. Finally a check would be made to see if the longest word determined could be *exactly* split into smaller words. If it could then the alternatives would also be considered as responses and the word(s) would be put into the output store, with label(s). At the same time information concerning the part(s) of speech of the word(s), and the semantic features if any, would be passed to other sections of the machine, suitably labelled to correspond with the word label(s). The marked items would then be discounted for recognition purposes and the process would continue. If there was a gap of unmarked features preceding the recognised item and these were sufficiently unlikely to be associated with a word then a labelled blank would be put into the output store, prior to the recognised word, and the unused features put into a secondary working store with the same label as the blank. These features would then also be marked. If these unknown features led to an equivocal decision then a partial output of the words recognised so far would be made with a request to repeat the doubtful section. If the output included the correct words thus far the operator could merely repeat as requested and the process would continue, the new features being inserted in place of the doubtful string. If the partial output were not accepted the input would have to start afresh. If the input stopped for more than a certain period a definite boundary would be inserted. Unused features between the last word recognised and the boundary would be treated in the same way as those between two recognised words.

When a definite boundary occurred, or if the output store filled up, a selection of the words in the output store would be made on the basis of information from a parsing program, using in addition any relevant semantic constraints. (At this point the ice becomes even thinner, if it is there at all.) The parts of speech put out to the parsing program would lead to the determination of a phrase structure which fitted one of the limited number of *consecutive* combinations of parts of speech derivable from the given alternatives. Newcomb's program is of this type. The labels associated with the parts of speech would then allow the semantic relationships to be checked on the basis of the phrase structure against the allowable relationships in the machine's model of the world. The only purpose of both these activities would be for the resolution of ambiguities. If no phrase structure could be fitted to the given parts of speech, then there would be no selection from the alternatives in the output store of the machine on this basis. Likewise if there were no semantic homogeneity obtainable by selection, or no semantic inconsistencies detectable, then again there would be no selection on this basis. Both these activities would only be aids to the recognition, and in the absence of any aid all the words in the output store would be output, and the operator would be left to rephrase or correct the utterance in the event of there being ambiguity still present. The machine's model of the world need not be very complex to be of assistance, but even a modest complexity is outside the present state of the art, for reasons both of size and nature of the storage required. (One can only postulate the most trivial constraints for a small vocabulary, such as the one suggested on p. 205.) Any information concerning the likely part(s) of speech and semantic associations of blanks could, however, be stored in an auxiliary store, together with a knowledge of who said it (if known). Any semantic associations would be passed back to the input to the word decision taker, to bias the recognition of succeeding utterances, and the parsing program could pass back syntactic expectation to bias the recognition of the word immediately following the string compiled up to that point.

When the output occurred from the machine the operator could either accept or reject it. If he accepted it, and it contained blanks, the machine would take it that the blanks were new words and would request, one by one—referring to the labels—for the blanks to be named. This operation could consist of spelling out the word, specifying an output channel, or manually interfering with the machine, but in any case the features and other relevant information which had been held in temporary storage would be put into surplus locations in the word decision store (thus displacing least useful items if the store were full) together with the appropriate output connection. Finally the whole output would be made available for whatever purpose the recognition had been intended. If the output were rejected the machine would ask for a repeat, and it would be up to the operator whether to try the whole utterance again, or proceed on a word by word basis.

Discussion

This section has done little towards proposing concrete ideas on the solution of the ASR problem; it would not be a research topic if it were possible to do so. Instead an attempt has been made to indicate the general lines of a solution to the total recognition problem in the hope that, however tentative, ambiguous and ill-defined the general scheme may be, it will at least give an idea of the complexity of the machine required to recognise a usefully large vocabulary spoken by speakers who differ significantly and use connected speech to converse with the machine. The core of the machine comprises the analyser, the sequence detector and the word recognition matrix with an output requiring verification. This could be non-adaptive at the start and would probably allow recognition of a limited vocabulary for a selected sample of speakers. The first task is to build these parts, or simulate them on a digital computer in order to show that they are feasible, and to allow an evaluation of the recognition performance. Work is already in progress on all these parts and an early version, lacking the sequence detector, was described at the IFIP 1965 Congress in New York (Hill 1966), and is described briefly in 'Building the machine', p. 219. Until the feature extractor is working there is no data which can be used to program the CAFs detection logic and therefore none to train the decision matrix on. The building of the feature extractor is therefore a first goal, though the design may well have to be modified in the light of recognition trials. Some of the features, such as silence, are reasonably easy to detect, but there are very serious difficulties when it comes to the treatment of spectral energy information for vowel quality detection and fricative quality detection. It seems that the processing for a number of features, particularly these, must proceed on a relative basis. Forgie† has recognised both fricatives and vowels for a fair spectrum of talkers using a straightforward statistical decision technique which required two dimensions for the fricatives and three for the vowels, but this has not been repeated by other workers, and the newcomer finds that he must start from scratch.

The other parts of the machine, with the exception of the parsing program rules, are highly speculative, and indicate the general lines of research for many years to come. They show the lines that machine intelligence studies will have to pursue in order to improve the speech recognition machines which will exist in five years' time.

Very little has been said about the details of adaptive processes for inclusion in the scheme. Probably the design of adaptive algorithms represents the most unexplored problem of all for the designer of an ASR machine.

† The two and three dimensions referred to came out during personal discussions with J. W. Forgie.

Adaption will be required both to allow the machine to benefit from past experience, and also to enable it to generalise from the experience over a wide range of potential input. The first objective could be achieved simply by recognising the re-occurrence of the previous situation and making parameter adjustments according to stored knowledge as suggested for speaker identification in the description of the machine. As an example, suppose there was a vowel quality categoriser which, given information about the spectrum of a sound, would classify it into one or other of a number of vowel categories. If it were known from previous experience that a particular speaker had an accent which substituted one vowel quality for another, then, when the speaker was recognised, the substitution could be made for the whole vocabulary simply by changing the connectivity between the categoriser class and the output line. If, on the other hand, it was desired to build an adaptive categoriser, which could take examples of the vowel sounds and build up a generalised categorisation, a far more subtle strategy would be required.

Let us consider a possible strategy. Suppose that the information pertaining to the vowel qualities was presented to a machine in the form of a binary pattern, which we shall call an I-pattern. And suppose that the machine comprised a number of storage rows containing a number of weighted digits equal to the number of digits in the I-pattern together with a category label, and means for comparing the I-pattern with the rows—to detect differences between the weighted patterns stored and a given I-pattern—together with a device which summed the weights of the digits which differed. The device might then operate in the following manner. Initially examples of each vowel quality would be given to the machine, and the patterns would be stored with the weights equal to unity and labels corresponding to the given labels. Further examples of the vowel quality would then be presented to the machine which would attempt to classify them. With each row would be associated a sum of the weights which differed. Some rows would differ, or deviate, more than others and the machine would try to choose the row most similar to the I-pattern. To do this a set latitude would be allowed for the deviation. The labels of the rows whose deviation fell within the latitude would be called 'similar' and would be considered as responses. First the label of the row with the least deviation would be given as a response. If this were correct then the weights of the digits which differed would be decreased and the categoriser would be more likely to make the same generalisation the next time. If it were incorrect then the weights which differed from the I-pattern would be increased, so that the response would be discriminated against, and then the row having the next smallest deviation would be tried, iterating the process. If none of the rows within latitude had the right label, then a new pattern-label pair would be stored in unoccupied storage. If the store were full then the least useful row would be displaced and the new row stored in its place. The usefulness of a row would be determined by two factors. First, how often it had been used successfully (useful age) and, secondly, how often it had been used successfully compared to the number of times it had been used unsuccessfully (reliability). This would lead to competition between the rows and ensure a tendency to keep the stored patterns up to date while retaining reliable useful patterns. The amounts by which the row pattern weights were decremented and incremented, and the balance between the measures of the usefulness of a row would be critical parameters for the success of this adaptive strategy, which was inspired by one of those in Andreae's STELLA machine (Gaines *et al.* 1966) and is related to the maximum likelihood strategy.

BUILDING THE MACHINE

General

A recogniser is being implemented at Standard Telecommunication Laboratories. It represents a first step towards the sort of recogniser outlined above, and will comprise the analyser, the sequence detector, and the word decision matrix. Parallel work on adaption is also being carried out, some directed at specific ASR problems. What is being built is the greater part of the analyser. The remainder, including adaptive vowel recognition strategies, and similar complex feature extractors required for the analyser, will be simulated. The machine has been described (Hill 1966), but without the sequence detector. Until the feature analyser is working little can be done towards implementing the sequence detector, since some knowledge of the features statistics is required as a basis for design.

The hardware

The analyser may be regarded as a very special-purpose analogue to digital converter. There is a more or less permanent basic system (i) for providing power, (ii) for sequentially storing the outputs of the feature extracting circuits, (iii) for providing a real time, broad band frequency analysis of the input signal, and (iv) for providing other ancillary equipment such as amplifiers, a tape punch for output, a visual display of stored information, monitoring of signal levels and the like. Using this framework, various feature extracting circuits may be tried out, and their effectiveness evaluated in conjunction with the computer simulation. It will also be possible to use the machine to provide binary coded data to the computer in the cases where the complexity of the feature decision makes it more profitable. The sequence of the features and data will not be determined in detail, instead various sampling techniques may be tried out and evaluated. For instance pulses could be generated at a relatively slow rate according to the broad structure of the word and note taken only of the pulse interval into which a feature, or group of features, fell. Thus sequence information within the pulse intervals would be lost, and also the effects of variation in the rate of utterance would be reduced before any decision-taking started. Two trial schemes of this nature are being evaluated at first, one based on the amplitude envelope of the word, pulses being generated at the beginning and end of a rise, and at the beginning and end of a fall in amplitude, and the other on the distinction between high and low frequency energy sounds. Any of the pulses may be used in combination to control the sampling.

The features being extracted at first are relative duration, voicing, friction noise, fricative quality, vowel quality, transitions of energy bands, silence and relative amplitude. Fig. 3 shows a block schematic of the feature extractor.

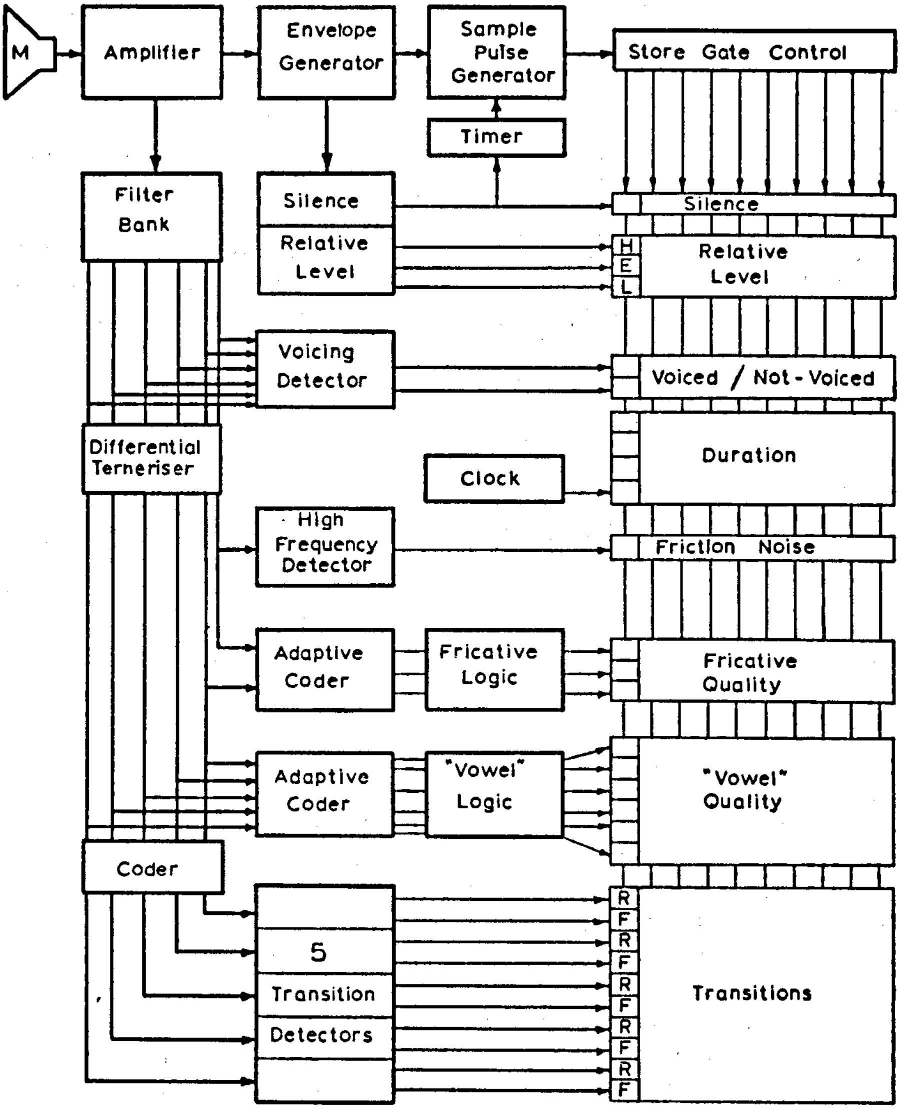

FIG. 3. A processor schematic.

The computer simulation

The rest of the machine will be simulated on a digital computer. One important variation from the scheme described at the IFIP congress concerns the manner in which the sequential features are taken into account, the revised scheme operating along the lines suggested on p. 214.

Where decisions are to be taken on binary data a variant of the maximum likelihood estimate will be used. Referring to equation (1) (p. 215) we derive a second equation similar to the first, concerning the probabilities when the event E_j does not occur:

$$P(\bar{E}_j/X)=\frac{P(X/\bar{E}_j)\ P(\bar{E}_j)}{P(X)} \tag{2}$$

then dividing (1) by (2) $P(X)$ is eliminated. Making the usual independence assumption on the features, and using an abbreviated notation we may derive the following formula:

$$R_{post}=R_{prior}\overset{n}{\Pi}(W_{ij}) \tag{3}$$

where R_{prior} is the *a priori* probability ratio of the event to the NOT-event, R_{post} is the *a posteriori* probability ratio (i.e., that, given the set of observations), and W_{ij} is the ratio of the probability of observing the ith feature—given the occurrence of the event—to that given the NOT-event. These probabilities may be estimated empirically. Thus the *a priori* probability that the event will occur, rather than not occur, is modified by such positive evidence as is available, on the basis of previous observations of the probability of observing elements of that evidence under the occurrence and non-occurrence of the event, to give the *a posteriori* probability that the event has occurred rather than not occurred. This is the general decision strategy which is employed, and the strategy outlined on p. 218 is related, given suitable incrementing and decrementing rules for the weight changes. The results obtained from the computer simulation could be embodied in hardware as indicated in Fig. 4, where logs are taken to facilitate the computation.

CONCLUSIONS

The first set of conclusions to be drawn concerns speech. We conclude that, although individual utterances can be described very accurately by means of present analytical techniques and synthesis can be equally accurate, there is a lack of means for general description of *all* utterances of the same class, where 'same class' means 'leading to the same, or a strongly similar, phonetic transcription'. For this reason there is currently much emphasis on physiologically based description of speech, on the grounds that a description of the generating process should be simpler than describing in general terms what may be generated. This approach will not be fruitful for ASR unless either physiological parameters are given to the recogniser (Hillix 1963) or such physiological parameters can be obtained from the acoustic signal. The former is not reasonable for most ASR applications, and the latter is likely to prove as difficult as recognising words using the acoustic signal. Approaches at the physiological level may, however, lead to a better understanding of the underlying structure of the acoustic signal,

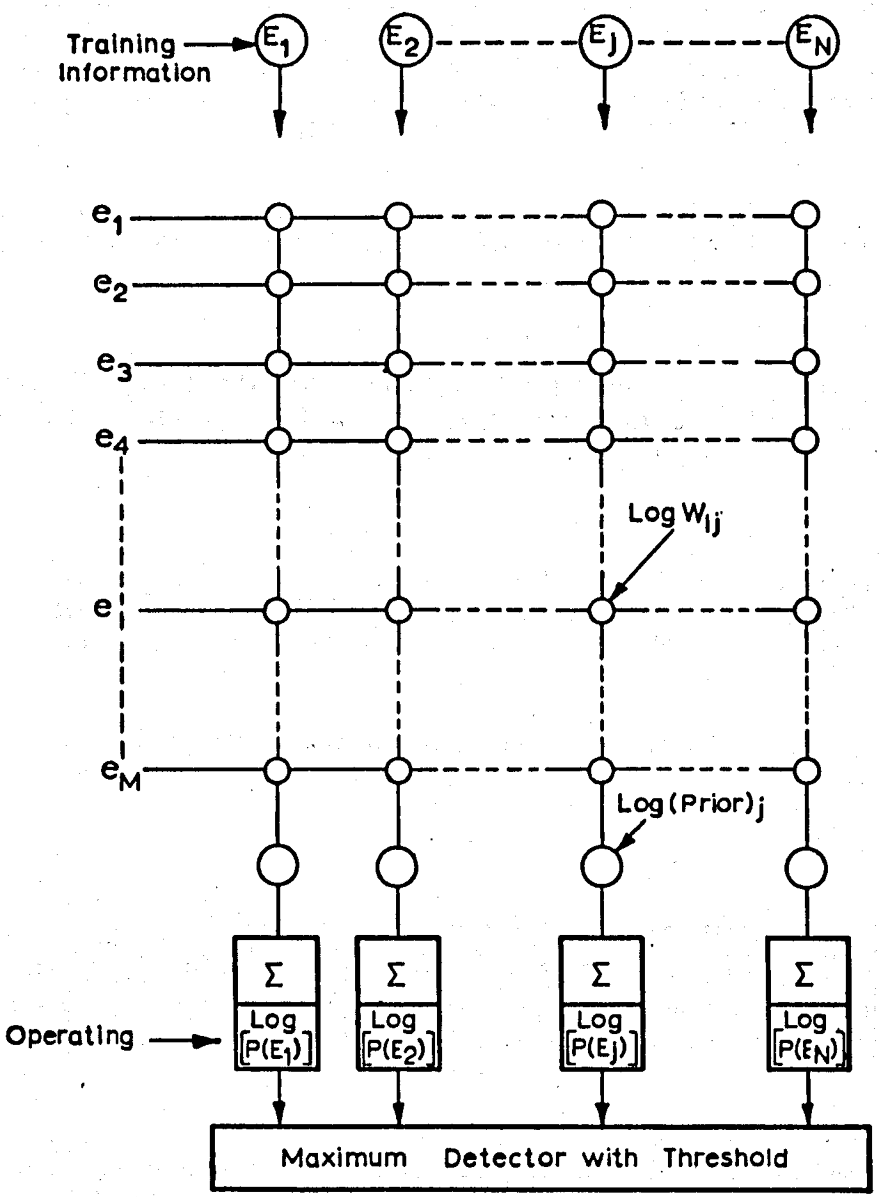

FIG. 4. Decision matrix.

and hence improve our acoustic description. Better descriptions of speech at all levels are essential to progress in ASR, both to describe the input signal prior to recognition, and to describe the general structure of speech so that useful knowledge of it may be built into an ASR machine. Above all there is an immediate need for quantitative work on speech data and statistics at all levels of description, based on descriptors considered important at present. Only by attempting recognition on present knowledge can the inadequacies of this knowledge, with respect to ASR, be fully illuminated.

Conclusions concerning the machine are even more difficult to draw. Certainly we do not know at present how to build a machine to recognise even the digits for just any speaker who happens by, though perhaps on such a small set of words *ab initio* training of the recogniser for each new speaker would be acceptable. Admittedly this would not be recognition in the sense that humans recognise the words of new talkers. Such an *ad hoc* solution should, however, be completely independent of the language of the talker, which perhaps lies behind Tribus' claim† that Dartmouth College's GE 235 will have an ASR input for any language within two years. To recognise a large vocabulary, when, for one reason or another, *ab initio* training would not be possible, a more powerful approach is required. It is likely that a first, primitive, recognition scheme, deciding entirely on the basis of acoustic phenomena, must be implemented. This will require good descriptions of speech at the acoustic level, a good description of the sequential properties, or 'grammar', of these primitive acoustic features, and a means for taking decisions, on the basis of this evidence alone, as to what word(s) could have led to the evidence. On to this basic structure must be added procedures which utilise inbuilt knowledge of the general structure of the speech signal at all the levels suggested above, in order to detect and correct errors, and to resolve ambiguities. The gradual introduction of these higher level procedures, and the embodiment of necessary adaption, will form the basis of ASR research over the next decades and constitutes the problem for machine intelligence.

ACKNOWLEDGEMENTS

The author wishes to acknowledge his debt to the many people who contributed directly and indirectly to the completion of this manuscript. In particular he would like to thank his colleagues at Standard Telecommunication Laboratories for helpful criticism and guidance, especially Dr J. H. Andreae who read the original manuscript and made many valuable suggestions at all stages of the preparation. The author has had fruitful discussion with many of the authors quoted, and would like to acknowledge his debt to them, while accepting full responsibility for any misunderstandings and misquotations which remain.

† Verbal statement during time-sharing, multi-access, computer demonstration at the Heriott-Watt College, Edinburgh, September 21, 1965.

REFERENCES

Abercrombie, D. (1965). Parameters and phonemes. In *Studies in Phonetics and Linguistics*. Oxford: Oxford University Press.

Antony, J., & Lawrence L. (1962). A resonance analogue speech synthesiser. *Proc. 4th ICA Copenhagen.*

Bhimani, B. V. (1964). Multidimensional model for automatic speech recognition AD 437 324 (1964). (This is a rather general reference to his speech recognition work, specific papers on the orthography-to-pronunciation programme, apart from abstracts to the 1965 Washington meeting of the ASA which bear some relation, have not apparently been published. The author's main source was from personal discussion.)

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Clapper, G. L. (1963). Digital circuit techniques for speech analysis. *Instn. elect. Engrs. Trans. Comm. Electronics,* **110,** 296-305.

Denes, P. (1959) The design and operation of the mechanical speech recogniser at University College, London. *J. Inst. Rad. Eng.*, **19,** 219-234.

Evans, E. F., & Whitfield, I. C. (1961). Classification of unit responses in the auditory cortex of the unanaesthetised and unrestrained cat. *J. Physiol.*, **171,** 476-493.

Fant, G. (1960). *Acoustic theory of speech production.* 'S-Gravenhage: Mouton.

Fromkin, V. (1964). Parameters of lip position. In *Working Papers in Phonetics.* University of Los Angeles, California. (*Also* as 'Lip positions in American English vowels', *Language and Speech*, **7,** 15-21.

Fry, D. B. (1956). Perception and recognition in speech. In *For Roman Jakobsen,* pp. 169-173. The Hague: Mouton and Co.

Fry, D. B. (1959). Theoretical aspects of mechanical speech recognition. *J. Brit. Inst. Radio Engrs,* **19,** 211-218.

Fry, D. B., Denes, P., Blake, D. Y., & Uttley, A. M. (1959). An analogue of the speech recognition process. In *Proc. Symp. Mechn. Thought Processes,* **1,** 375-395. London: HMSO.

Gaines, B. R., *et al.* (1966). A learning machine in the context of the general control problem. (Not published.)

Green, P. S. (1958). Consonant vowel transitions. *A spectrographic study. Studia Linguistica* 12, pp 57-105.

Hill, D. R. (1966). STAR—a machine to recognise spoken words. *Proc. Int. Conf. Inf. Process. 1965 Congress,* Vol. II, Spartan/MacMillan.

Hillix, W. A. (1963). Use of two non-acoustic measures in computer recognition of spoken digits. *J. Acoust. Soc. Amer.*, **35,** 1978-1984.

Holmes, J. N., & Shearme, J. N. (1961). An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1-formant 2 plane. *Proc. 4th Int. Cong. Phonetic Sciences.* Helsinki 1961, Mouton 1962.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol.*, **160,** 106-154.

Hughes, G. W. (1961). The recognition of speech by machine. *MIT Tech. Report* No. 395. ASTIA AD 268 489 (does not cover subsequent results referred to in text but indicates Hughes' general approach).

Hultzén, L. S., Allen, H. D., & Miron, M.S. (1964). *Tables of Transitional Frequencies of English phonemes.* U. Illinois Press.

Ingemann, F. (1966). Eight parameter speech synthesis. *Edinburgh University Phonetics Department Progress Report*, Sept.-Dec. 1960.

Jakobson, R., Fant, C. G. M., & Halle, N. (1961). *Preliminaries to speech analysis.* MIT Press (4th printing).

Kersta, L. G. (1962). Voiceprint identification. *Nature*, **196**, 1253-1257. Also *Bell Monograph*, **4485.**

Lawrence, W. (1953). The synthesis of speech from signals which have a low information rate. In *Communication Theory*, ed. Willis Jackson. Butterworth.

Lehiste, I. (1962). Acoustical characteristics of selected English consonants. *ASTIA Report* AD 282 765 (1962).

Lettvin, J. Y., Maturana, H., McCulloch, W. S., & Pitt, W. (1959). What the frog's eye tells the frcg's brain. *Proc. Inst. Radio Engrs*, **47**, 1940-1951.

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

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of unvoiced stop consonants. *Am. J. Psychol.*, **65**, 497-516.

Liberman, A. M., Delattre, P., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.*, **27**, 769-773.

Liberman, A. M., Gerstman, L. J., Delattre, P., & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/. In *English Word*, 13, p. 24.

Lindgren, N. (1965). Machine recognition of human language (in three parts). *Instn. elect. Engrs, Spectrum*, April/May/June.

Lisker, L., & Abramson, A. S. (1965). Stop categorisation and voice onset time. *Proc. 5th Int. Cong. Phonetic Sciences*, pp. 389-391. Karger: Basle. 1965.

Marril, T., & Bloom (1963). CYCLOPS—a second generation recognition system. *Proc. AFIPS Fall Joint Comput. Conf.*, pp. 27-33.

Marril, T., & Bloom (1965). CYCLOPS—2 system. *Comput. Corp. Amer. Tech. Rep. RT65-RD1.*

Miller, G. A. (1947). The masking of speech. *Psychol. Bull.*, **44**, 105-129.

Miller, G. A. (1951). The intelligibility of speech as a function of the context of test materials. *J. Exp. Psychol.*, **41**, 329-335.

Miller, G. A. (1962). Decision units in the perception of speech, *Inst. Radio Engrs Trans. Inf. Theory*, **8**, 81-83.

Miller, G. A. (1964*a*). The psycholinguists (on the new science of language). *Encounter*, **23**, 29-37.

Miller, G. A. (1964*b*). Communication and the structure of behaviour. *Disorders of Communication*, **42**, 29-40.

Miller, G. A. (1965). Some preliminaries to psycholinguistics. *Am. Psychol.*, **20**, 15-20.

Potter, R. K., Kop:, G. A., & Green, H. C. (1947). *Visible speech.* New York: van Nostrand.

Pruzansky, S. (1963). Pattern matching procedure for automatic talker recognition. *J. Acoust. Soc. Am.*, **35**, 354-358.

Sebesteyen, G., *et al.* (1962). Voice identification, general criteria. *ASTIA Rept.*

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

Tillman, G. H., Heike, G., Schnelle, H., & Junghever, G. (1965). DAWID I—ein Beitrag zur automatischen 'Spracherkenung'. *Proc. 5th Int. Cong. Acoustics.* Liege.

Truby, H. M. (1958). A note on visible and indivisible speech. *Proc. 8th Int. Congress Linguistics*, pp. 393-400. Oslo University Press.

Truby, H. M. (1959). Acoustico-cineradiographic analysis considerations with especial reference to some consonantal complexes. *Acta Radiol. Suppl.*, **182**, complete. Stockholm.

Wells, J. C. (1963). A study of the formants of British English. *Progress Report of the Phonetic Lab.*, University College, London.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *M.I.T. Tech. Report*, 369, 444-466.