Automatic recognition and speech research

Fant, G.

journal: STL-QPSR
volume: 11
number: 1
year: 1970
pages: 016-031

http://www.speech.kth.se/qpsr
Introduction

The purpose of this article is to collect some thoughts on automatic speech recognition (ASR). At the Speech Transmission Laboratory we have a very limited experience with actual recognition schemes but we have a need to formulate research objectives and outline alternative approaches.

Pierce (1969) has recently challenged the community of automatic speech recognition workers by an article in which he expresses rather critical view of the aims and methods of current work in this field. Most projects have a narrow scope and are based on primitive concepts, the practical value of ASR is limited, and the general recognition problem is beyond our present capacity to handle according to Pierce. However, I do not believe the Bell Telephone Laboratories have given up their work in this field and the purpose of the challenge should be regarded as a stimulant for proper planning rather than a general dissuasion.

Indeed, most of the work reported up till now is concerned with limited inventory word recognition schemes that allow a tolerable level of accuracy for his master's voice only and require special voice training. Some more sophisticated systems are speaker adaptive, some make use of linguistic and probabilistic constraints, but no system has yet been reported on that fully utilizes all of these constraints and in addition is designed for time normalized, low bit rate, information preserving processing at lower levels.

The typical primitive approach is to sample speech intensity within a few frequency bands at regular time intervals and with two or only a few levels of amplitude. This quantized representation is then correlated for maximum match with the standard patterns of the lexical items. It does not pay to introduce many amplitude and frequency levels since the lack of time normalization will anyhow introduce gross matching errors. The performance of such systems will be rather low but they have the benefit that speaker adaptive functions as well as new lexical
items can easily be added. No phonetic criteria are needed since a word is recognized as a single unit.

This type of recognition is similar to that utilized by a dog responding to simple oral commands. A decomposition of speech into a set of minimal phonetic categories and the handling of more complex sentences than single words is attempted in more advanced ASR-projects. Evidently, it will be hard to compete with the human brain and practical compromises have to be made in any design. This, however, is not the same as saying that computers cannot be taught to react in a way similar to that of human beings.

Future success in ASR depends on two factors, an increase in basic knowledge concerning speech and language, and the advance of computer technology. The need for research is unlimited and will continue to be so for the decades to come. Meanwhile, we should now and then test the status of the art and evaluate its practical applicability in man-machine communication systems. The objectives of ASR are thus in a wide sense the same as of speech research in general. Recognition algorithms are the image of synthesis-by-rule programs and ASR is accordingly just as much an organizing principle for speech research as a product of such research. At least this is how it should be. Primitive systems do not provide a feed-back to the investigators' facts concerning why mistakes are made and how models should be improved.

The demands for very large computer storage space for "unlimited" recognition of connected speech are basically the same as those in automatic language translation. The practical applicability will depend on future cost reduction, miniaturization, and feasibility of handling large data quantities, enabling the necessary storage of all a priori knowledge from the separate stages of the speech communication process, the semantic stage included.

In addition to these academical views I want to emphasize the need for developing reasonably optimized special purpose limited inventory systems and partial solutions, e.g. the technique of recognizing some basic phonetic categories that can be utilized in aids for the deaf, e.g. as supplements to lip-reading.
Previous work

It is not my intent to make a review of the literature. Readers are referred to the survey of Lindgren (1965) which provides a comprehensive presentation of ASR in relation to speech research. One of the most detailed recent surveys is that of Hyde (1968) who lists a large number of earlier contemporary projects and rounds up his study with a general discussion. His review of separate projects is rather condensed and does not give much room for evaluations. His discussion, however, is concerned with general principles and recommendations. These are well founded. I can accordingly agree with Hyde that "Feature abstraction is better than pattern matching in recognition at the acoustic level", "The speech signal cannot be directly segmented into phoneme units", "Time normalization by simply stretching or compressing one speech wave compared with another, is of rather limited value", "Probabilistic information must be used with caution", "Learning systems have a place in automatic speech recognition but cannot be valuable exploited until the basic principles of recognition have been determined", and most of his other themes, etc.

This issue of STL-QPSR contains summary articles on recent Soviet work at the USSR Academy of Sciences, Institute for Mathematics in Novosibirsk, Zagoruiko (1970), and at the University of Lvov, Ukraine, Derkach (1970). The Soviet work on ASR is well organized and has produced interesting results but was not very well referenced in the earlier Western literature, Falter and Ctten (1967).

Models of recognition

In principle it would be feasible to design speech recognitions without any knowledge of the perception processes providing the human production processes are sufficiently well known and a sufficiently large empirical store of data relating speech waves and message compositions have been collected.

However, even though our knowledge of the speech perception processes is rather limited it pays to formalize our concepts of various stages at the listener end of the communication into working models for practical recognition schemes. One such model is that of Bondarko, Zagoruiko, Kozhovnikov, Molchanov, and Chistovich (1968), see also
Kozhevnikov and Chistovich (1965), which recognize three basic stages, an acoustic data sampling stage, a phoneme recognition stage, and a word recognition stage. A somewhat more elaborated model is proposed here in Fig. III-A-1. It contains five successive stages of processing labelled

1. Parameter extraction
2. Microsegment detection
3. Phonetic transcript
4. Word identification
5. Semantic interpretation

The main path towards higher levels of recognition is paralleled by a downstream translation path for enabling a testing at lower level of tentative identifications at higher levels. A set of comparators marked \( c_1, c_2, c_3, c_4, c_5 \) are introduced for this purpose. Each stage has its storage \( a_1, a_2, a_3, a_4, a_5 \) of a priori information concerning inventories and constraints of the message representation at the particular level which provides reference data for processing and identification. Upstream and downstream by-pass pathways are also indicated in the block diagram. These shortcuts would, for instance, provide evidence for identification of a speech sound or of a word directly from the lowest level. The downgoing by-pass could enable a conditioning of lower centers to a state appropriate for testing a certain state predicted at a higher level. Multi-pathways in parallel and shortcuts appear to be a topological characteristic of the human brain also. Otherwise I do not claim that the block diagram of Fig. III-A-1 is representative of human perception. Its main purpose is to enable a discussion of recognition strategies but even for this purpose it is very incomplete, e.g. with respect to memory functions of shorter or longer span which is crucial for the recognition. The outgoing line would have some correspondence to speech production. In this sense the comparators would fit in a motor theory of speech perception although this is not strictly implied by the model. Reference patterns involved in a comparison should first of all be conceived of as belonging to a perceptual sphere, Fant (1964).

The successive stages of a maximally ambitious recognition system could be given the following functions.
SPEECH RECOGNITION

parameter extraction | microsegment features | phonetic transcript | word identification | semantic interpretation

stores of a priori knowledge

\[ a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5 \]

by pass

\[ \text{In} \quad \text{Out} \]

Successive stages of identification

comparators

Acoustics | Phonetics | Linguistics

Fig. III-A-1. Block diagram illustrating successive levels and functions in a generalized speech recognition scheme.
(1) Parameter extraction

Sampling intervals should be of the order of 5-20 msec and preferably pitch synchronous in voiced intervals. Many different sets of parameters can be selected such as:

a) Short-time average amplitude in several frequency bands.
b) Short-time average amplitude and zero-crossing density in a few frequency bands or as a single overall measure.
d) Main spectral maximum locations and spectral slopes.
e) Spectrum described by an expansion into moments or other functions.

In addition the voice fundamental frequency $F_0$ should be tracked. Formant and pitch tracking has proved to be rather difficult to execute without errors. This is the main reason why formant vocoders have not been very successful yet. The apriori knowledge store a1 could contain vocal transfer constraints such as the normal relation between spectral energy distributions and formant frequency patterns to avoid erroneous formant number labeling in c).

The short-time memory in stage s1 should span over at least three successive sample points. Time differentials of parameter values should be included to enable microsegmentation in s2.

(2) Microsegment extraction

Rapid changes in the speech spectrum and intensity induced by changes in the active source or sources or in the main characteristics of the vocal transfer function give rise to discontinuities that define boundaries of "microsegments", see Fant (1962, 1968). These minimal production units are not identical to phonemes but can constitute parts of phonemes. The "source", "manner" and "place" categorizations of such segments are essentially the same as for feature analysis in general whereas the typical "place" correlate is a spectral feature that often varies continuously across segment boundaries, e.g. from

* Frequency analysis along a $\sigma+j\omega$ line in the $s$-plane in order to sharpen formant peaks.
a stop burst to a following vowel. The situation is similar to when there is a shift from noise to voice source in the course of a normally voiced consonant following an unvoiced consonant. However, not all manner correlates give rise to sharp boundaries. An example is velum lowering well ahead of a nasal consonant in which case the appearance of the nasalization is gradual. Also, there is not always a continuity of place cues, e.g. the spectral energy is going from a fricative such as [s] to a vowel, e.g. [a].

The output of the microsegmenter stage s2 is a sequence of segments classified as to "source", "manner", "duration", and "place of articulation" or if the segment place cannot be identified the cues are to be reconsidered at the next stage, s3. The memory span required is of the order of 4-7 segments. The a priori information contains vocal-tract physiological and acoustic constraints, including rules for speaker normalization. Spectrum "place" parameters should be selected according to established perceptual models.

(3) Phonetic transcript

The phonetic transcript stage, s3, converts the microsegment data to a representation in terms of a string of phonemes and superimposed prosodemes. Alternatively, the output of stage s3 is an equivalent distinctive feature matrix, Jakobson, Fant, and Halle (1952) or Chomsky and Halle (1968). The particular choice of feature is not very important as long as the phoneme is uniquely defined by the features. These need not be binary.

It should be noted that a microsegment description process is non-unique, being arbitrarily dependent on the choice of parameters and thresholds. There is also a speaker dependent variability to take into account. These variabilities are compensated for by the inherent redundancies. The algorithm for identification of a phoneme or of one of its distinctive features, i.e. "nominal" phonetic category should take into account such free variations.

The a priori knowledge storage a3 contains rules for the phonetic correlates of the distinctive features and includes the sequential constraints in terms of possible phoneme or feature sequences. The location of vowels within a string of phonemes and prosodic categories, such
as stressed and unstressed syllables and tonal accents, should be identified first. Rules for coarticulation, Öhman (1967 A, B) and reduction, Lindblom (1967) and more detailed rules of segmental durations and structuring related to underlying sequences of phonatory and articulatory commands also enter the a3 storage. A memory span of about seven syllables is needed. An intermediate representation in terms of syllables might be advantageous. However, coarticulation effects are not limited to syllabic units and the initial identification of syllabic nuclei is accordingly more important than exact identification of syllabic boundaries.

(4) Word identification

The segmentation and identification of a tentative string of phonemes and superimposed prosodemes as a sequence of words has to proceed according to linguistic rules. A primitive system was utilized in the study of Reddy and Robinson (1968), see also Zagruiko (1970). The concepts of modern generative grammar could be exploited in more sophisticated systems.

The a pricri storage a4 contains the equivalent of lexical, syntactical, and phonological rules of the language. Probabilistic information should be added.* The output of stage s4 is a string of words properly identified and transcribed according to the particular lexical and syntactical conventions adopted. A print-out in ordinary spelling is one possible form of the output of s4. "Parts of speech" of the order of sentences might be bounded by reference to overall prosodic patterns, Öhman (1967 B).

(5) Semantic interpretation

The semantic interpretation acts as a support for proper selections at the s4 stage by the requirement that the decisions shall conform with the semantic sphere, e.g. the topic as judged from the prior history of the message.

* The overall gain is limited by incorrect identifications of very infrequent words. This risk may be minimized if reliability measures are computed already in stage s3, e.g. if probabilities of alternative phoneme strings have been determined.
The phonetic versus the engineering approach

When developing an overall synthesis strategy there are two extreme philosophies to choose between. One is the "phonetic" approach in which case one or both of the stages s2 and s3 of Fig. III-A-1 is incorporated. The other, by-passing s2 and s3 to directly identifying words on the basis of parametrical speech wave data could be called an "engineering approach". Which should be chosen? In favor of the common engineering approach we have a general feature of simplicity. Phonetic decisions are avoided and one can directly operate on overall correlation functions which preserve distinctive elements from the initial learning process or from subsequent adaptive learning sessions. This argumentation is deceptive and has determined the approach in most research groups.*

In the extreme phonetic approach on the other hand phonemes are decoded on the basis of their distinctive features. This is graphically represented by a branching tree with nodes for each feature. It can be rightly critizized that an error in one feature will invalidate the recognition whereas the engineering approach with overall correlation functions would emphasize the most likely word in spite of local variations. However, the choice is hypothetical only, since a strict serial feature recognition system without parallel branchings to one and the same phoneme is of academical interest only. Practical phonetic schemes should employ non-exclusive algorithms, Reddy (1967 B), i.e. a many to one input-output relation.

Most adheres to the engineering approach fall for the temptation to use a simplified parametric representation which causes an initial information loss. Clipped speech is intelligible to a human observer thanks to the capacity of the human brain but is not equally intelligible to a computer. The amplitude distribution in a few filter bands has a limited capacity to define the speech message as experienced from the low performance of vocoders with too few filters. Of course this is a matter of cost also and very limited vocabularies can be identified with a sampling of the overall speech intensity contour only. However, the performance could be much improved by a more sophisticated sampling system reducing the initial information loss.

* We have some limited experience of such systems.
One draw-back of the "engineering" approach is that speaker-dependent nonphonemic variations of parametric patterns might dominate the correlation functions. On the other hand, this mere fact could be made use of for a retrospect analysis of the causes of errors in the systems and according to the engineer, such an approach could contribute to the developing of knowledge concerning distinctive elements in the patterns. This possibility is interesting and could be exploited for the benefit of the phonetic approach where a collection of knowledge for separating speaker-dependent variations from distinctive cues is a basic requirement.

The greatest problem in the engineering approach is to achieve a proper normalization in the time domain. It is quite apparent that a contribution to the overall correlation between stimulus and reference will be highly negative at intervals of overlap between a vowel of a reference word and a fricative in the same word uttered at a different tempo during the test. In the phonetic approach, on the other hand, the microsegmental structure will provide a natural segmentation which in spite of being non-unique provides an inherent normalization. However, we need more experience on the output and efficiency of such segmental procedures.

Various schemes for time normalization have been suggested. Schroeder (1968) has suggested a continuous method which provides information on both the pathway selected in a multidimensional frame and the time spent within each part. Velichko and Zagoruiko (1969) have used linear time normalization for more simple cases and a non-linear system when the demands on the primary recognition is greater. The method involves the comparison of successive 14 msec samples within the reference and the stimulus starting from the beginning of each word. Successive pairs of stimulus sample and reference sample are selected by adding one time increment to either or to both, whatever gives the best match. The sums of the correlation functions along this pathway are compared for each pair of stimulus and possible reference words.

A proper identification of a sequence of microsegments labelled according to voiced/voiceless and other source and manner categories might suffice for a limited vocabulary recognition.
Distinctive feature theory

The distinctive feature concept is a very powerful tool in linguistic theory as well as in practical applications of speech research but the relations between the two levels are not always simple. The basic principle implies that all members within a group of phonemes, e.g. all stop sounds, or all voiced sounds, all nasal sounds, all front vowels, etc. have a feature in common in terms of the human production process and in terms of corresponding speech wave characteristics and auditory sensations. One and the same processing stage would accordingly suffice in an automatic recognition process for each of the distinctive features and the number of parallel processors is simply the number of features of the classificatory system. However, the variability induced by the context in a general sense can be considerable. A feature decoding must take into account what other features co-occur in the matrix column of the phoneme and which features occur in adjacent phoneme columns and furthermore the values of all prosodic descriptors. In extreme cases a direct identification of a phoneme is the simplest way of decoding its distinctive features and the latter are then of secondary interest only. This is the main reason why I have chosen to state that a representation in terms of distinctive features matrices is an alternative to phonemes at the output of stage s3 rather than an intermediate representation in an earlier stage. Instead, the microsegmental phonetic representation at the output of stage s2 serves as a set of cues for recognition of either phonemes or distinctive features in stage s3. However, phonetic and linguistic regularities, e.g. segmental constraints, are often simpler to state with reference to distinctive features than with respect to phonemes. In this sense it is the "classificatory" function of the feature defining a certain category of phonemes rather than its phonetic manifestation that is of importance.

The phonetic value of a distinctive feature can be regarded as a vector in a multidimensional signal space. The variability due to context shall be expressible by rules how the feature vector is changed when the conditioning elements are varied. A minimum requirement of phonetic reality of a feature would be that any two sounds in any context differing by one and the same feature only shall display a difference vector of the same sign along a common phonetic dimension.
The production correlate of the feature is a specific control channel of the speech mechanism. This idea of features pertaining to natural "biological" channels is of course strong as long as a clear cause-effect relation can be tested. The feature systems of Chomsky and Halle (1968) have not been followed up on the acoustic and perceptual level as rigidly as the Fant, Jakobson, and Halle (1952) system which emphasized the perceptual identity of features.

In a practical analysis scheme when several alternatives shall be tested for maximal likelihood it is convenient to contrast phonemes directly and not by the classificatory features only. Thus a choice between [l][r][d] can be optimized by defining differential features for testing each of the alternatives with respect to all other alternatives, Derkach (1970). In general, a feature by feature pathway for going from one phoneme to another phoneme in the complex signal space may display a very uneven and long sequence of lines whereas the direct resultant vector may be of a rather small magnitude, Fant (1966).

This is one consequence of the present trend in linguistics to define features as a set of convenient minimal sign for formulating rules of language regularities in the first place and as phonetic manifestations in the second place only. This situation reflects our great lack of systematic knowledge in general phonetics and the need to establish generative rules not only on the linguistic level but also on a "subphonological" level to derive all the conditioning rules and constraints of speech production, speech waves, and speech perception.

There does not exist any acceptable feature system worked out for applications in automatic speech recognition. The phonetic approach of Reddy (1967A, B) is interesting but rather "ad hoc". For limited inventory analysis one can choose a system that suits the particular lexical inventory. For more general purposes, as an alternative to the Chomsky and Halle (1968) or the Jakobson, Fant, and Halle (1952) systems, one could apply a conventional set of phonetic features ordered into four main groups.

I. **Manner features** (vowel, glide fricative, stop, nasal, lateral, r-sound)
II. **Source and secondary manner features** (voiced aspirated, tense)
III. **Consonantal place features** (labial, interdental, dental, alveolar, palatal, velar, uvular, pharyngeal, retroflex)
IV. **Tongue body features** (three back/front and four low/high levels labeled front, central, back; high, midhigh, midlow, low).
This set of features also applies to parts of the microsegmental description, Fant (1968). For recent discussions on feature theory, see Fant (1969A) and a forthcoming article by Fant. One advantage of a conventional phonetic feature system is that the terminology is easier to comprehend and that a natural articulatory category need not be defined by combinations of other more abstract features. The greater redundancy of this conventional feature set could be an advantage in automatic recognition.

Only + signs need to be used in the matrix. The empty spaces are either fully predictable from the selection of plus signs and usually negative, [+ stop] implies [-vocalic]. If not predictable the empty spaces are negative in an "unmarked" sense (vowels are usually [-nasal], stops and fricatives [-voice] if not otherwise noted. Here follows an example of feature encoding of an English sentence, "The cat saw the bird".

<table>
<thead>
<tr>
<th>Feature</th>
<th>( \hat{e} )</th>
<th>( \partial )</th>
<th>( \varepsilon )</th>
<th>( \varepsilon )</th>
<th>( \varepsilon )</th>
<th>( \partial )</th>
<th>( \delta )</th>
<th>( \partial )</th>
<th>( \partial )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glide</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stop</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-sound</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voiced</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspirated</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tense</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labial</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labiodental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interdental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alveolar</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palatal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usular</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharyngeal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retroflex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Front</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Central</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Midhigh</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Midlow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>
Note that the consonants are here specified by a single feature only in each of the categories I, II, and III and become coarticulated with the category IV feature of adjacent vowels. Coarticulation in the typical sense is thus the time-space combination of consonantal place features III and tongue body features IV which within certain limits appear to represent independent control channels.

The above proposed classification of features into four groups, pertaining to manner source, consonantal place, and main tongue body place I find more natural than that of Jakobson, Fant, and Halle (1952) or Chomsky and Halle (1968).**

Recognition strategies. Final words

It is not within the scope of this rather philosophical article to recommend any detailed system for speech recognition. Several principles have been taken up and the general flow diagram of Fig. III-A-1 could embody an infinite variety of solutions. There are some points I would like to stress. The variability of the acoustic correlates of phonemes and distinctive feature with the immediate and grosser frame could imply that the problem is unsolvable since the recognition of any

---

In group I [+vowel] can combine [+nasal] or with [+r-sound]. The latter "r-coloring" can also be transcribed [+retroflex]. In group II two features may be needed, e.g. for the Korean stops. In category III "Retroflex" can combine with alveolars and palatals. The vowel system IV requires a maximum choice of three + arranged according to classical dimensions.

** I do not claim that the feature system proposed here represents a best choice. The novelty lies in the four-group classification of features. A feature system based on a muscular control model of the articulatory organs would have inherent advantages but is not sufficiently far developed yet. An approach in this direction is the vowel model of Lindblom and Sundberg (1969). The relation of features derived from this model and other vowel feature systems is discussed in a forthcoming article by the author in STL-QPSR 2-3/1970. In my view there is a great need for investigations of the acoustic and perceptual correlates of articulatory features and to consider feature systems defined initially on the acoustic level and given production correlates in the second place only. A recent study of Swedish stop sounds, Fant (1969 B) contributes to the understanding of cues-feature relations.
feature would require some knowledge of all other features in a certain area of the matrix. However, a pretty good guess can be made initially in stage 3 of some distinctive features and phonemes, e.g. vowels and fricatives. The number of phonemes within a certain frame is not easily determined in an early stage. This uncertainty is largely a matter of difficulty of consonant segmentation.*

There thus develop initially enough of fixpoints to set up the conditioning factors for recognition of remaining phonemes and features. The stress and intonation pattern identification proceeds in parallel with the first identifications and constitutes an important subset of factors conditioning the inherent features. Sequential and combinatorial constraints rule out some of the alternatives. The final output of stage 3 can be thought of as a specification of alternative phoneme sequences, each evaluated with respect to its probability. Grammatical and semantic constraints add further evidence for identification of words and whole sentences. Generative rules can then enable the check on a lower level of alternative sequences and constructs predicted at higher levels, see Fant (1968).

References and literature


* The number of vowels and their durations should be determined at an early stage of analysis. However, unstressed vowels can be very short, one or two voice pitch periods in extreme cases.


