SPEECH SOUNDS AND SEQUENCES

MORRIS HALLE*

It is traditional to regard utterances as sequences of discrete entities: speech sounds, phonemes, allophones, or whatever else one chooses to term them. It is in this way that utterances are represented in alphabetic writing systems, and there are numerous facts of language that require this picture of speech. For example, the formation of the regular plural of English nouns, is usually described in a fashion much like the following:

/iz/ is added if the noun ends in /s/, /z/, /š/, /ž/, /č/, /ǰ/
/s/ is added if the noun ends in /p/, /t/, /k/, /θ/, /f/
/z/ is added in all other instances.

This formulation is obviously predicated on the assumption that speech is composed of phonemes. While alternative formulations in terms of dyads, syllables, words, or other units larger than the phoneme are not ruled out, it is, however, clear that these will have to be considerably more complex since in place of the final phonemes the rules would have to list the much more numerous larger units. The discrete picture of speech, moreover, is easily and naturally integrated into the widely accepted view of speech production, as of a process in which sequences of discrete entities are translated into gestures of the vocal tract and thence into sound. Since it is important for my further argument, I would like to spell out here in some detail how the process envisioned in this account of speech production might actually take place.

It is assumed that stored in the memory of the speaker there is a table of all the phonemes and their different actualizations. This table is basically a dictionary in which can be found the different vocal tract configurations or gestures that are associated with each phoneme, and the conditions under which each of the configurations or gestures is to be used. Associated with some phonemes there may be but a single configuration or gesture; with others the number of gestures may be large. The number of entries per phoneme corresponds, of course, to the number of allophones of the phoneme in question. It should be observed that a given configuration or gesture need not be associated exclusively with a single phoneme. "Overlapping

* This work was supported in part by the U.S. Army (Signal Corps), the U.S. Navy (Office of Naval Research) and the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and in part by the National Science Foundation.

allophones" present, therefore, no conceptual difficulty to the present model and need not, as far as the model is concerned, be ruled out. Parenthetically it may be noted that the model is not restricted to operating with phonemes, but can handle other types of symbols equally well. In particular, it would experience no difficulty if the phonemes and the instructions for their actualization were framed in terms of distinctive features.

In producing an utterance the speaker looks up in the table each phoneme in the utterance and then causes his vocal tract to assume in succession the configurations or gestures corresponding to the phonemes composing the utterance. The vocal tract behaviour in turn causes disturbances in the air which are transmitted to our ears as acoustical signals. Since the acoustical signals are completely determined by the vocal tract behavior that produces them, we shall talk here primarily in articulatory terms and only occasionally refer to the acoustical properties of utterances.

Since the vocal tract does not require the same amount of time for actualizing each phoneme, it must be assumed that stored in the speaker's memory there is also a special schedule that determines the time at which the vocal tract moves from one configuration to the next. The timing will evidently differ depending on the speed of utterance, it will be slower for slower speech and faster for faster speech. Because of the inertia of the vocal tract, however, it is conceivable – even highly probable when utterances are produced at high rates of speed, though by no means only then – that a given vocal tract configuration or gesture may not be reached in the time foreseen in the schedule, so that the vocal tract may be able only to approximate the required sequence of gestures. In extreme cases the vocal tract may omit altogether some of the gestures or configurations foreseen in the schedule. This is a very important fact, for as a consequence of it we must expect difficulties in trying to divide utterances into segments that stand in a one-to-one relationship with the discrete entities of the input; i.e., with the phonemes in the utterance.

In sum, the traditional view of the process of speech production assumes that the speaker possesses a set of instructions or rules which allow him to transform a sequence of discrete entities (i.e., phonemes) into quasi-continuous behavior of the vocal tract and thence into a quasi-continuous acoustical signal. We shall call this set of instructions the *generative rules.*

In addition to producing utterances, speakers also perceive utterances. We shall be interested here only in one aspect of speech perception, namely the ability of speakers to analyze an utterance into the phonemes which compose it. I do not believe that this ability is the result of learning to read and write in an alphabetic script, but rather that it is acquired at the same time as the ability to speak. As a bit of supporting evidence for this view I might cite the fact that in the speech of my illiterate sons the plural of *ox* is /aksiz/. Since no one in their surroundings uses this form I assume that they produce this form in accordance with the general rule for the formation of the regular plural of English nouns, which requires that the suffix /iz/ be added to nouns ending in /s/. But, as has previously been noted, in order to

use this rule the speaker must be able to analyze the noun into its component phonemes.

The model of the analysis process that has enjoyed the widest, if not exclusive acceptance among linguists postulates that the listener first segments the utterance and then identifies the segments as particular phonemes. In order to do this the hearer must have in his memory a list of the acoustical equivalents of the phonemes, which is essentially the reverse of the dictionary that was postulated for the speech production process. A desirable feature of this model is that it places a very modest burden on the memory, for even if we admit quite a number of allophones per phoneme, the dictionary would hardly exceed a few hundred items. In certain refinements of the model the segmentation and identification are performed after analyzing the utterance into a set of pertinent properties, e.g. distinctive features, which further reduces the requirements on the size of the memory.

This analysis procedure depends crucially on the hearer's ability to perform segmentation. But if speech is produced in the manner that has been described above, then it is in principle not possible that the hearer will be able to segment all utterances. As we have seen, there will be utterances or parts of utterances that will not be segmentable.

Once this is granted, however, we must inevitably face the question that was raised by Ladefoged at the Teddington Symposium on the Mechanization of Thought Processes: "why, and in what sense, speech is a sequence of discrete entities?" Since we cannot hope ever to propose a fool-proof, perfectly general segmentation procedure, there is only one path open to us: we must show that the speech signal can be analyzed into a sequence of discrete entities by a procedure that does not depend crucially on segmentation. We must describe a process that recovers the discrete phonemes from the continuous speech signal even when the signal cannot be segmented.

I shall examine now three devices which are capable of performing this type of analysis. The first two devices will be rejected on the grounds that they fail to satisfy certain obvious requirements that have to be imposed on analogs of speech perception. The third device satisfies these requirements, and I shall, therefore, suggest it both as a possible model of speech perception, and as an adequate reply to the question in what sense utterances are sequences of discrete phonemes.

Perhaps the crudest device capable of analyzing a continuous acoustical signal into a sequence of discrete entities without prior recourse to segmentation is a dictionary in which utterances are entered as acoustical signals – or in some convenient transformation of these, such as Visible Speech sonagrams – and each entry is provided with its phonemic representation. The operation of such a device is epitomized in Fig. 1. The utterance under analysis is placed in the component labeled "comparator", where it is compared with the first item from the "dictionary". The result of the comparison is communicated to the "control", which performs several functions: 1) it decides on the next dictionary item to be sent to the "comparator"

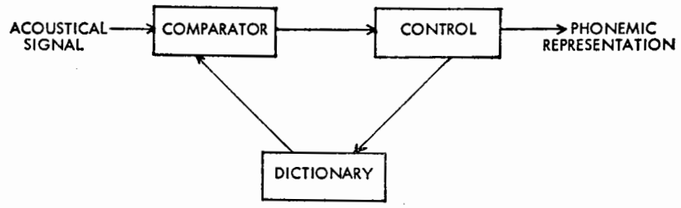

Fig. 1.

(in the present instance "control" would have to insure only that each item in the dictionary is compared no more than once and that no item is inadvertently omitted); 2) it remembers the dictionary entry that resulted in the best fit with the utterance under analysis; 3) it prints out the phonemic representation of that entry once the dictionary has been exhausted.[1] The number of utterances that can be identified by this device is directly proportional to the size of the dictionary. Hence, if a device of this type were to approach even remotely the capabilities of a normal speaker, it would have to contain a dictionary so large as to rule it out as a plausible model of speech perception.

The need for a large dictionary can be overcome, however, if the principles or rules of construction of the dictionary entries are known. It is then possible to store in the memory only the rules which, in the case of speech, would be identical with the "generative rules" mentioned in our discussion of the production process. A model of this type is shown in Fig. 2. It differs from that shown in Fig. 1 only in two respects. First, the "dictionary" is replaced by the "generative rules"; secondly, the "control" component will now determine the item to be sent to the "comparator" by supplying the "generative rules" with all possible phoneme sequences, systematically exhausting all one-phoneme sequences, two-phoneme sequences, etc. Incidentally, we may now include in the "control" information which would allow it to reject or appropriately modify inadmissible phoneme sequences.

While the model in Fig. 2 does not place excessive demands on the size of the memory, it suffers from the crucial flaw of requiring a very long time to achieve positive identification, since every possible phoneme sequence has to be sent to the "comparator". This undesirable feature can be eliminated if one is able to suggest a preliminary analysis which would exclude from consideration all but a very small

[1] Several investigators have seen the process of speech perception in precisely this light. In their important paper "Some Experiments on the Perception of Synthetic Speech Sounds", F. S. Cooper and his collaborators write: "The problem of speech perception is then to describe the decoding process either in terms of the decoding mechanism or – as we are trying to do – by compiling the code book, one in which there is one column for acoustic entries and another column for message units, whether these be phonemes, syllables, words or whatever." (*Journal of the Acoustical Society of America*, 24, 605, 1952). This model of the decoding process is utilized also in the well-known digit recognizer Audrey; cf. K. H. Davis, R. Biddulph and S. Balashek, "Automatic Recognition of Speech," *loc. cit.*, 637–642.

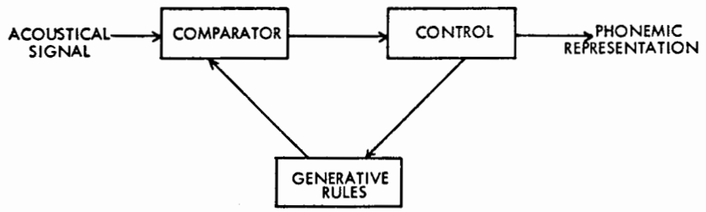

Fig. 2.

part of the potentially infinite number of items that can be generated by the "generative rules." A model of this type is shown in Fig. 3. The signal under analysis is first subjected to a "preliminary analysis" which includes various transformations that have been found useful in speech analysis, such as segmentation, identification of segments by special attributes, etc. The fact that all of these procedures can be only partially successful is no bar to their utilization here since the aim of the preliminary analysis is not positive identification. After being subjected to "preliminary analysis" the signal is sent to the "comparator", where it is processed as before. The results of the preliminary analysis are communicated also to the "control", which systematically supplies the "generative rules" with the items not excluded in the "preliminary analysis" and prints out the phoneme sequence resulting in the closest resemblance with the signal under analysis.

I should like to draw particular attention to the fact that for some signals several phoneme sequences may yield equally good matches in the "comparator". In such cases the "control" would print out more than one phoneme sequence, which is its way of indicating that the utterance is ambiguous. An important class of ambiguous utterances is due to "overlapping allophones"; e.g., *balm* and *bomb* in many American English dialects, or *reisst* "tears" and *reist* "travels" in many German dialects. In these cases the model will print out two possible representations, which is exactly

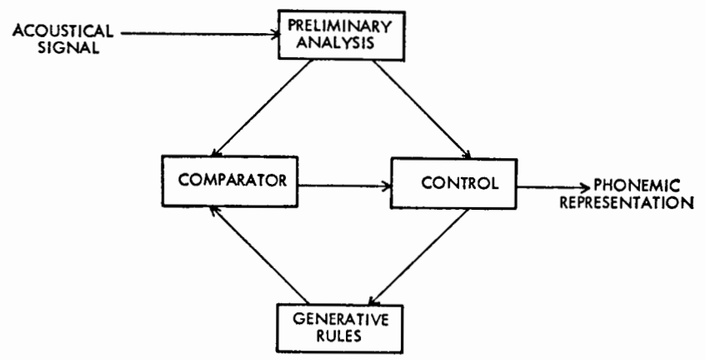

Fig. 3.

analogous to what the human listener would do in the same circumstances – i.e., where he cannot utilize grammatical or semantic information to resolve the ambiguity. This seems to me a much more reasonable procedure than the one implicit in conventional phonemics, where "overlapping allophones" are ruled out by definition and a segment possessing a particular set of phonetic properties is always assigned to the same phoneme.

It should be noted that in the "preliminary analysis" the signal may be radically transformed; as a result, all operations involved in the analysis are not necessarily performed on an acoustical signal, but rather on some fairly complicated transformation of the latter. This is very important since I do not mean to suggest that in order to perceive speech it is necessary actually to make sounds or move the tongue, lips, etc., any more than it is necessary to generate smells in order to perceive odors.[2]

The "generative rules", which are the heart of the proposed model of perception, also constitute the core of the process of production. The dual processes of production and perception are viewed, therefore, as separate utilizations of a comon core of rules rather than as distinct processes each with its own body of rules. This seems eminently reasonable to me, for I find it difficult to believe that a natural phenomenon like language should be designed in so uneconomical a fashion as to require two totally distinct processes. I can see, therefore, little utility in the recent efforts to develop special "grammars for the hearer" to supplement the traditional linguistic descriptions, which, it is claimed, have almost universally been framed from the viewpoint of the speaker. If perception involves the type of process that has been described here, there is no need for such duplication. A single set of "generative rules" adequately covers everything that is relevant.

To sum up, I have tried to bolster the traditional view of speech as a sequence of phonemes by presenting a model of speech production and perception in which phonemes play the central role. The proposed model overcomes in a natural manner the problems raised by "overlapping allophones" as well as those resulting from the impossibility to achieve complete segmentation of all utterances. The model itself presupposes only the ability to receive and emit acoustical signals and to perform logical operations; i.e., abilities that all human beings possess. Since the logical operations involved are of a very high order of complexity, it may be objected that these exceed human capabilities. This objection overlooks, however, the crucial fact that only a very small part of these operations is under conscious control, and it is well known that man excels even the most advanced electronic computer in the execution of complex logical tasks, as long as these require no conscious effort on his part.

The model of speech perception that has been discussed here is essentially identical with the one presented by M. Halle and K. N. Stevens. Cf. their "Analysis by Synthesis", W. Wathen-Dunn and L. E. Wood (edd.), *Proceedings of the*

[2] This comparison is borrowed from D. M. MacKay's paper, "Mindlike Behaviour in Artefacts," *British Journal for the Philosophy of Science*, II, 105–121 (1951).

Seminar on Speech Compression and Processing (Air Force Cambridge Research Center Technical Report 59–198; Bedford, Mass., Dec. 1959), and "Speech Recognition: A Model and a Program for Research," *IRE Transactions on Information Theory*, Vol. IT–8, 155–9 (1962). A very similar model of perception was proposed earlier by D. M. MacKay in "Mindlike Behaviour in Artefacts," *loc. cit.*

Department of Modern Languages and
Research Laboratory of Electronics
Massachusetts Institute of Technology