This paper was presented at the Speech Communication Seminar (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm) in September, 1962; it will appear in the Proceedings of that seminar.

SPEECH SYNTHESIS BY RULES

Franklin S. Cooper, Alvin M. Liberman*, Leigh Lisker**, Jane H. Gaitenby
Haskins Laboratories, New York

"Synthesis by rule" provides a convenient way to designate a class of speech processing techniques that emphasizes the production of acceptable speech by methods that are somehow rather special. Certainly, more is implied than the mere conversion of control voltages into acoustic output as, for example, in the Vocoder; there is a clear implication that the procedure by which such control voltages are obtained from the available input is sophisticated enough to deserve our attention or is, at the very least, non-trivial. This excludes existing bandwidth compression systems, however ingenious the synthesizers that implement them, though it might well include a phoneme recognition system on the basis that the transmitted information would be coded in non-machine terms.

Bandwidth compression to the level of the phoneme is only one of the potential systems that might employ synthesis by rules; it has been mentioned to focus attention on a distinguishing characteristic of all such systems, namely, that the input in the form of discrete symbols is coded in linguistic terms -- not machine terms. Thus, synthesis by rule implies two processes, first a conversion of input symbols into machine control signals, and then the utilization of these signals by an appropriate synthesizer to generate speech.

There are, to be sure, processes other than true synthesis that will accept linguistic symbols and yield a speech waveform. In the simplest case, the output is assembled directly from prerecorded segments of speech, and rearranged into new sequences for each new message. Such a compilation method contrasts with synthesis by rule in significant ways; moreover, there are hybrid methods that avoid some of the difficulties inherent in either synthesis or compilation.

For these reasons, we shall depart from the strict confines of our title and consider the broader topic of how to proceed from an input consisting of discrete linguistic units to speech as an output signal, by whatever method or combination of methods.

Speech By Compilation.

Various attempts have been made to generate connected speech by assembling recorded fragments of spoken utterances; the principal differences have been in the size and nature of the segments. Briefly,

* Also, the University of Connecticut.
** Also, the University of Pennsylvania.

the method appears to be feasible only if these prerecorded segments are rather large (phrases, words, or possibly syllables) and correspondingly numerous.

Liberman et al.[1] and Sivertsen[2] have discussed, from rather different points of view, the difficulties and quantitative problems involved in employing short segments of various kinds. The use of letter names -- an attractive choice if the input symbols are themselves letters -- results in a kind of "spelling bee" that may well be of some use in a reading machine for the blind[3] but is hardly acceptable as speech. The attempted use of sound segments corresponding to the phonemes of a message encounters the very fundamental problem that, in the acoustic domain, these perceptually and linguistically discrete units are not strung together like beads on a string; rather, they are usually merged, or encoded, into units that are more than one phoneme in length. Hence, just as surgical separation of the sound segments is excluded, so too is synthesis by the assembly of such phonemic "building blocks". It might, of course, be possible to employ different recordings, or allophonic variants, of most of the phonemes for most of the combinations in which they occur,[4] but this would require a very large inventory of prerecorded segments. Indeed, one might prefer to deal with segment pairs, or dyads,[5] that contain "parts of two phones with their mutual influence in the middle of the segment". An analogous unit, the syllable dyad, is obtained by cutting the sound stream at the mid-points of successive syllables. Cuts at syllable boundaries would yield, in addition, the syllable and half-syllable as potential segments for a compilation process.

The problems that are encountered in using short segments of one type or another are of three kinds: matching the cut ends to give cursive speech, avoiding an excessive inventory of segments, and finding the storage address of a desired segment from the available input symbols. We have only hinted at the basic difficulties in cutting and matching within word or syllable patterns: not only must formant frequencies and intensities flow smoothly together, but the voice harmonics must agree if pitch jumps are to be avoided. The relationship between length of segment and size of inventory has been examined by Sivertsen;[2] she concludes that "the segment inventory becomes disproportionately large when the segments are not co-terminus with linguistic units". Finally, the task of locating the stored segments will be easiest if the input symbols -- usually letters and spaces -- can themselves serve as addresses. For all these reasons, it appears that the word is the smallest practical unit, and probably the optimal compromise between the problems involved in its use and the quality of the resulting speech.

The production of running speech even from prerecorded words encounters certain difficulties: most important is the "language problem" of how to provide recordings of individual words that can be shuffled about like moveable type and yet fall into the accustomed patterns of connected speech; there is, in addition, the machine problem of how best to store and retrieve a very large number of such recordings. The number of items to be stored will depend, of course, on the application, but if the machine is to deal with a wide range of printed material -- as it must in a reading machine for the blind, as one example -- then it seems unlikely that less than ten to twenty thousand stored words will be required. Various forms of storage might be employed: analog or digital versions of spoken or synthetic words, or the control parameters (in analog or digital form) from which synthesis could be effected. A real-time machine would, by present-day techniques, require a pair of very large disc or drum memory units; on the other hand, a more limited device, intended to explore and evaluate the word-by-word method as applied to reading machines for the blind, will soon be in operation.[6]

The stored segments that will be used in the prototype Word Reading Device are being recorded by voice under carefully controlled conditions. Indeed, the development of suitable instructions to the speaker has been one of the important, and most difficult, parts of the problem. A workable way has been found to instruct a trained speaker so that he will record words that can be combined into a speech output that is highly intelligible at slow-to-average reading rates, and approximates the phrasal patterns of normal connected speech. The general nature of the instructions is indicated in Fig. 1. Each word is first assigned to a grammatical class in accordance with its most frequent usage. Each class entails a desired pattern of duration, intensity, and pitch change; the instructions to the speaker are then phrased in terms that he can apply in monitoring his own productions.[7]

The task of obtaining acceptable speech has been complicated, in the particular case described above, by the requirement that each printed word shall have only one recorded version to represent it, regardless of the many ways that the word is normally used and spoken. An obvious improvement would be to allow several recorded versions per printed word, and to provide linguistic rules and logical circuitry to select a recording appropriate to the actual grammatical function of the word in its specific context. There would, no doubt be a gain in performance; there would certainly be an increase in complexity and cost.

Synthesis Based On Spectrum Data.

Most of the speech synthesizers that have been developed thus far for experimental use or for application in bandwidth compression systems operate on the basis of information about the acoustic spectrum. The signals that control the synthesizer can be in the form of a spectrographic pattern or parameters derived from it. Many factors are involved in the optimal choice of a synthesizer[8] but, for our present purpose, we can neglect such considerations as flexibility of use and even quality of speech output; instead, we are concerned with how to arrive at spectrum data when we are given only a sequence of phonemes, or even of letters. The spectrum data can be readily adapted to the input requirements of the particular synthesizer and so we can, without loss of generality, deal with spectrographic patterns alone.

The basis for a set of rules for the conversion of phoneme sequences to spectrographic patterns exists in the results of extensive experiments on the acoustic cues for the perception of speech[9] In the particular formulation given by Liberman, et al.,[1] the objective was to formulate rules that would be few in number, simple in structure, and capable of converting a phoneme string into speech of reasonable intelligibility at normal speaking rates; these are referred to as "minimal rules" to emphasize both the desire for economy and the realization that a price would be paid in the naturalness of the synthetic speech.

The structure of these minimal rules reflects certain broad generalizations that can be made about the cues themselves, namely, that they are independent, additive, and fit naturally into a relational framework that parallels the familiar articulatory frame. The rules take full advantage of this latter characteristic and are written in terms of the subphonemic dimensions of place, manner, and voicing. A rule, then, includes all the statements that must be made in order to specify the acoustic cues that pertain to one of these dimensions for a stated class of phonemes. Each phoneme in the message is represented by a set of subphonemic rules which, as will be obvious, must contain provisions for their own simultaneous participation in determining the spectrum pattern.

Since phoneme sequences must lead to smoothly flowing spectrum patterns, there is need also for sequential accomodation between successive sets of rules; these statements are also contained within the subphonemic rules. It may be noted that the characterization of acoustic cues for the consonants in terms of their loci[10] provides the practical basis for these sequential statements.

An example of synthesis by minimal rules is shown in Fig. 2 (See also Ref. 1). The general comments about the relationships between rules apply, in the case of the phoneme /b/ for example, as follows: the set of rules for /b/ contains separate rules for the stop consonants /pbtdkg/, the labials /pbfvm/, and the voiced consonants /bdg/; certain statements in these rules must be combined simultaneously to specify, for example, the "silence"; and certain other statements must be combined sequentially with the rules for /æ/ and /z/, in particular those that generate the formant transitions.

It will be evident that such a collection of rules, organized subphonemically, achieves both parsimony and simplicity of structure by reducing the spectrum specification to only those acoustic elements that serve as important perceptual cues. One must expect that individual characteristics and naturalness will suffer accordingly; nevertheless, these rules do provide a workable and completely explicit procedure for converting a phoneme transcription into control signals for a speech synthesizer.

In practice, however, the rules are not quite so tidy as the discussion thus far would suggest. One complication is that we must sometimes make special provision for positional effects; that is, there are some few combinations of phonemes for which an appropriate pattern is produced only if the basic rules are qualified by the addition of a position modifier. In addition, positional rules of the type illustrated in Fig. 2 for /æ/ are needed to signal stress. These qualifications do not seriously compromise the system; for example, the set of minimal rules described by Liberman, et al., consists of nine rules for place of consonant articulation, five for manner of consonant articulation, three rules for voicing and, for the vowels, two manner and twelve place rules. This is a total of thirty-one rules (without provision for intonation) as the basic set. The addition of twelve position modifiers and one stress modifier, for a total of forty-four rules, takes care of the special cases.

In brief, then, at least one solution exists to the problem of converting linguistic symbols into machine control signals for the synthesis of speech. In practice, synthetic speech has been produced in this way. At the Haskins Laboratories and at the University of Edinburgh, various workers have applied the rules with the aid of a paint brush and the Pattern Playback and PAT[11]; at the Bell Telephone Laboratories, rules based on those described above have been used to program a large digital computer, which then served both to generate the spectrum information and to produce the speech waveforms.[12]

Synthesis Based On Articulatory Configurations.

The conversion by rule from phoneme sequence to speech waveform need not proceed by way of the acoustic spectrum; it should be possible, and indeed desirable, to control the synthesis in terms of the changing configurations of an equivalent vocal tract. The rules for synthesis, in this case, will convert from phonemes to electrical signals that control the "shape" of a dynamic vocal tract analog such as DAVO/DANA.[13] In this particular synthesizer, the control signals determine the (equivalent of) cross-sectional areas of segments of the vocal tract, the coupling to a nasal resonator, and the excitation.

In the ideal case a synthesizer of the vocal analog type would generate waveforms subject to built-in constraints like those that limit the possible range of speech sounds produced by the human vocal mechanism.[14] This virtue is only partially realized in existing devices since the equivalent configurations are not firmly tied to physiological possibilities.

The principal problem in synthesis by configurational rules is how to determine the shape at every part of the tract and at each successive instant of time from symbols that specify only gross adjustments at particular points in the tract, and these only at successive epochs. The most direct source of information for the solution of this problem comes from X-ray motion pictures; the technique is, however, cumbersome and the results not very precise. A second approach is to use a "best guess" configuration to generate an acoustic output and then to compare this output with speech at the corresponding moment in articulation; a match between the outputs does not, however, guarantee that the particular configuration corresponds to the actual one, but only that it is one of a set of shapes that would suffice acoustically. A third type of information comes from conventional articulatory phonetics. Although its descriptions are woefully incomplete (for machine control purposes), they do serve to stake out important landmarks to which the dynamic course of the articulation must needs conform; they do not, however, comprise an adequate basis for interpolation.

Stevens and House[15] have provided an important interpolation device in their three-parameter description of the vocal tract, and they have shown its practical capabilities. It may well be that the methods of analysis-by-synthesis,[16] when applied to the problem of tracking the articulatory shape and aided by such interpolation devices, will lead to the early formulation of a set of rules for synthesis comparable to those already developed for generating spectrographic patterns. Such an accomplishment would have exciting consequences, both practical and theoretical. From the latter point

of view, it will be extremely interesting to compare the configurational constraints (or the procedures for interpolating between successive events) with the sequential accommodations between sets of rules for spectrum synthesis.

Synthesis Based On Motor Commands.

We have seen that a salient feature of rules for synthesis via either spectrum data or articulatory configuration is the need for mutual accommodation between successive elements. This requirement follows quite simply from the nature of the speech process, in which we assume that the phonemes of the intended message are transformed into a set of motor commands which are then encoded into the changing configurations of the tract, and that these, in turn, are further encoded to yield the acoustic signal. If one adds the hypothesis[17] based largely on perceptual data, that the signals at the level of motor commands provide a simple and direct representation of the phonemes, then one is tempted to undertake synthesis by rules that operate directly in motor command terms.

The rules for converting linguistic units (phonemes) into machine control signals (motor commands) would be simple indeed; in fact, they might amount to no more than a look-up operation in a table that would contain about as few entries as there are phonemes. But what has happened to the complexities encountered in other synthesis procedures? These complications cannot simply have been charmed away, since they are inherent in the speech process; rather, they have all been transformed into constraints on the synthesizer.

The requirements on such a synthesizer are not, in principle, so very awesome. Perhaps the major difficulty is that they call for hardware and techniques that lie outside the main stream of current technology. One can visualize -- though he might hesitate to begin building -- an analog machine containing actuators embedded in flexible, constant-volume casings, and variously anchored to each other and to a rigid structure with a hinged attachment. This would be the direct way to incorporate the real-life constraints.

A more attractive approach may be the indirect one of using a computer to convert (from stored tabulations of empirical data) the desired sequence of motor command patterns into tract configurations, and then to turn these into acoustic sequences with the aid of a vocal tract analog.

It may appear from this rather hypothetical discussion that nothing has been gained by concentrating the complexities within the synthesizer. This may prove to be true, but let us note that this is where Nature put the difficulties. Moreover, this approach to the problem does serve to focus attention on the nature of the constraints that will be most helpful in programming a dynamic vocal tract analog; specifically, the constraints should be organized, not in terms of shape per se, but rather in terms of the configurational relations that follow from localized muscular contractions in the flesh-and-blood transducer. There are, in addition, theoretical consequences that flow from a motor command approach to synthesis by rule, and these alone would justify some part of the foregoing speculation.[18]

Rules Or Dictionaries?

We have taken, as our overall topic, the conversion of linguistic symbols into speech waveforms. Two basically different procedures have emerged, but they tend not to remain distinct in practice. These two, plus some of the possible hybrid procedures, are shown in Fig. 3. The simplest of the processes we have considered (the bottom line in the figure) is surely that of compiling speech from a dictionary of recordings, even though the number of stored items may prove to be very large; the more interesting processes, in our view, are those (on the top line of the figure) that exploit what is known about speech and speech perception in order to arrive at rules for synthesis. The two boxes representing synthesis by rule (at the middle and right of the top line) might be implemented in any of the three ways we have described; that is, the rules may serve to convert the incoming phoneme sequence into control signals based on acoustic spectra, vocal tract configurations, or motor commands; the appropriate synthesizer then uses these signals to generate speech. (The phoneme sequence may itself require a generative procedure employing either rules that relate spelling to phonemic (or phonetic) transcription or a "pronouncing" dictionary. These are indicated by boxes at the upper left in the figure.)

Some of the procedures that employ a combination of dictionary look-up and synthesis-by-rule may well prove to have important practical advantages; given a specific set of requirements, the highest quality in speech output for the lowest cost in instrumental complexity is more likely to be met by a hybrid system than by one limited to either compilation or synthesis alone.

References

The work reported here was supported in large part by contracts with the Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration.

1. Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P., Cooper, F.S. Minimal rules for synthesizing speech. J. Acoust. Soc. Amer., 1959, 31, 1490-1499.

2. Sivertsen, E. Segment inventories for speech synthesis. Language and Speech, 1961, 4, 27-61.

3. Freiberger, H., Murphy, E.F. Reading devices for the blind. In Human Factors in Modern Technology, McGraw-Hill (in press)

4. Harris, C. Study of the building blocks of speech. J. Acoust. Soc. Amer., 1953, 25, 962-969.

5. Peterson, G., Wang, W.S-Y., Sivertsen, E. Segmentation techniques in speech synthesis. J. Acoust. Soc. Amer., 1958, 30, 739-742.

6. Cooper, F.S. Toward a high performance reading machine for the blind. In Human Factors in Modern Technology, McGraw-Hill (in press).

7. Gaitenby, J. Word reading device: experiments on the transposability of spoken words. J. Acoust. Soc. Amer., 1961 33, 1664 (A).

8. Cooper, F.S. Speech synthesizers. Proceedings of the Fourth International Congress of Phonetic Sciences, Helsinki, 1961. (To be pub.)

9. Liberman, A.M. Some results of research on speech perception J. Acoust. Soc. Amer., 1957, 29, 117-123.
Delattre, P. Les indices acoustiques de la parole: premier rapport. Phonetica, 1958, 2, 108-118, 226-251.

10. Delattre, P., Liberman, A.M., Cooper, F.S. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Amer., 1955, 27, 769-773.

11. Ingemann, F. Speech synthesis by rule. J. Acoust. Soc. Amer., 1957, 29, 1255 (A);
Ingemann, F. Eight-parameter speech synthesis. In progress report from the Phonetics Dept., Univ. of Edinburgh, Sept.-Dec., 1960.

12. Kelly, J.L., Gerstman, L.J. An artificial talker driven from a phonetic input. J. Acoust. Soc. Amer., 1961, 33, 835 (A).
Gerstman, L.J. Vowel duration in an artificial talker driven from a phonetic input. J. Acoust. Soc. Amer., 1962, 34, 743 (A).

13. Rosen, G. A dynamic analog speech synthesizer. J. Acoust. Soc. Amer., 1958, 30, 201-209.
Hecker, M.H.L. Studies of nasal consonants with an articulatory speech synthesizer. J. Acoust. Soc. Amer., 1962, 34, 179-187.

14. Fant, G. Acoustic theory of speech production. 's Gravenhage, 1960.

15. Stevens, K.N., House, A.S. Development of a quantitative description of vowel articulation. J. Acoust. Soc. Amer., 1955, 27, 484-493.
Stevens, K.N., House, A.S. Studies of formant transitions using a vocal tract analog. J. Acoust. Soc. Amer., 1956, 28, 578-585.

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

17. Liberman, A.M., Cooper, F.S., Harris, K.S., MacNeilage, P.F. Motor theory of speech perception. Preprint for Speech Communication Seminar, Stockholm, 1962.

18. Lisker, L., Cooper, F.S., Liberman, A.M. Uses of experiment in language description. Word (in press).

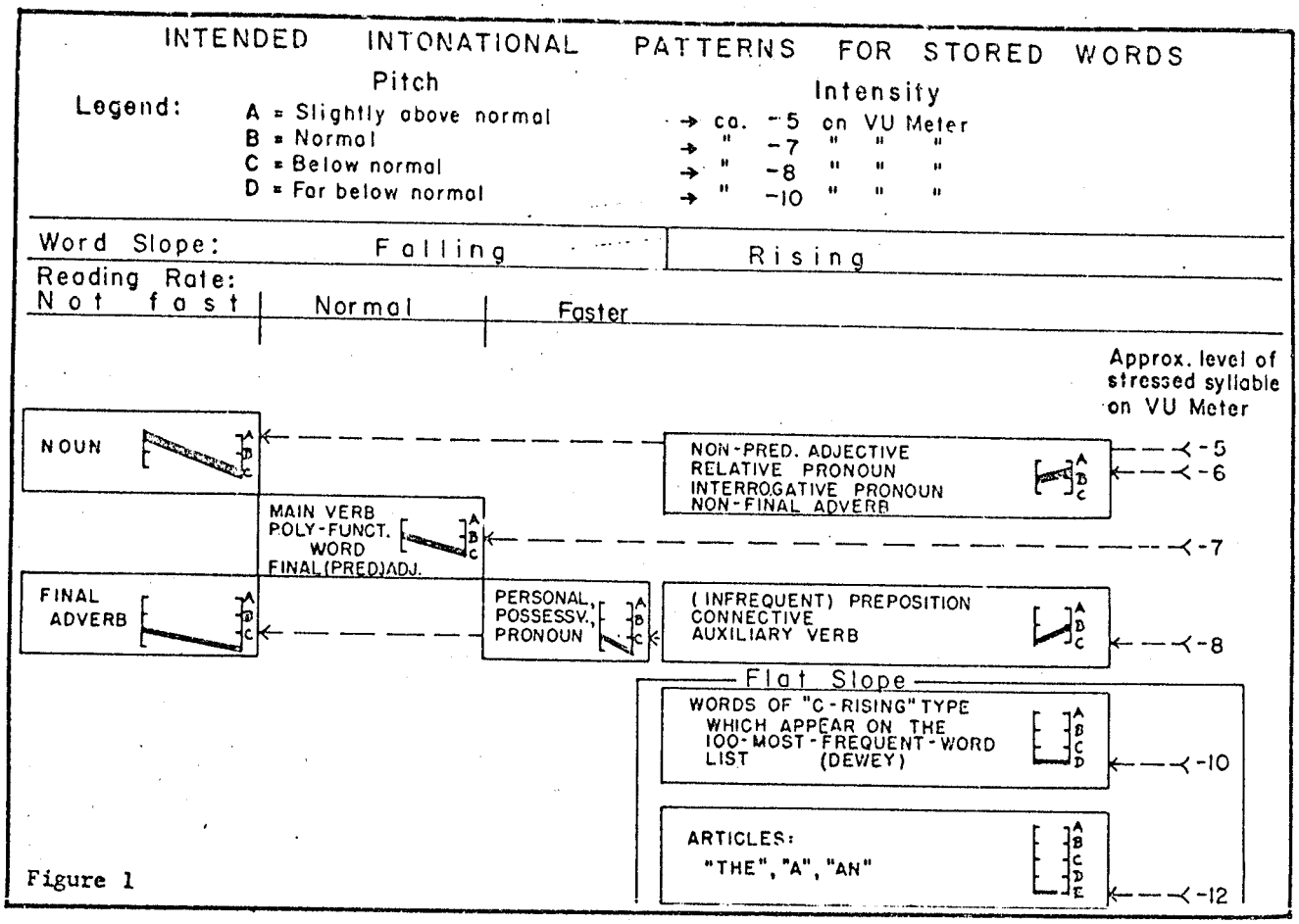

Figure 1

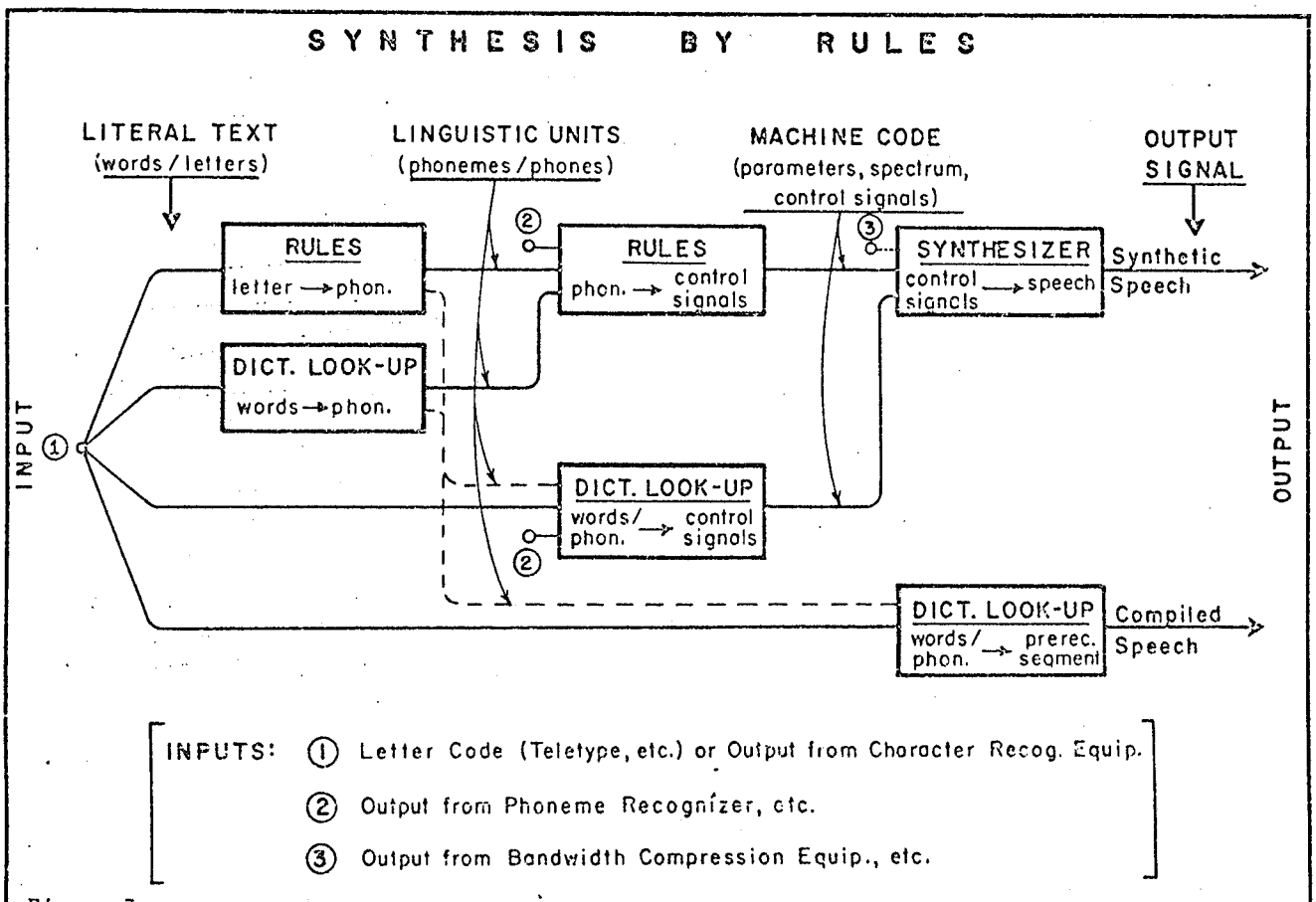

Figure 3

SYNTHESIS BY RULE: /læbz/

Manner	*Resonants* /wrly/: Periodic sound (buzz); formant intensities and durations are specified. F1 locus is high. Formants have explicit loci.	*Long Vowels* /ieɛæaɔo/: Periodic sound (buzz); formant intensities and durations are specified.	*Stops* /pbtdkg/: No sound at formant frequencies; i.e., "silence." Burst of specified frequency and band width follows "silence." F1 locus is low. F2 and F3 have virtual loci.	*Fricatives* /fvθðszʃʒ/: Aperiodic sound (hiss); intensity and band width are specified. F1 locus is intermediate. F2 and F3 have virtual loci.
Place	/l/: F2 and F3 loci are specified.	/æ/: Formants frequencies specified.	*Labials* /pbfvm/: F2 and F3 loci are specified. Frequencies of buzz and hiss are specified	*Alveolars* /tdsz/: F2 and F3 loci are specified. Frequencies of buzz and hiss are specified
Voicing	(The voicing rules are only applied to those phonemes for which the condition of voicing has differential value. For the resonants and vowels, which are invariably voiced, the acoustic features correlated with voicing are specified under Manner.)		*Voiced* /bdg/: Voice bar. Duration of "silence" is specified. F1 onset is not delayed	*Voiced* /vðzʒ/: Voice bar. Duration of hiss is specified. F1 onset is not delayed
Position		Vowels in final syllable: Duration is double that specified under Manner.		

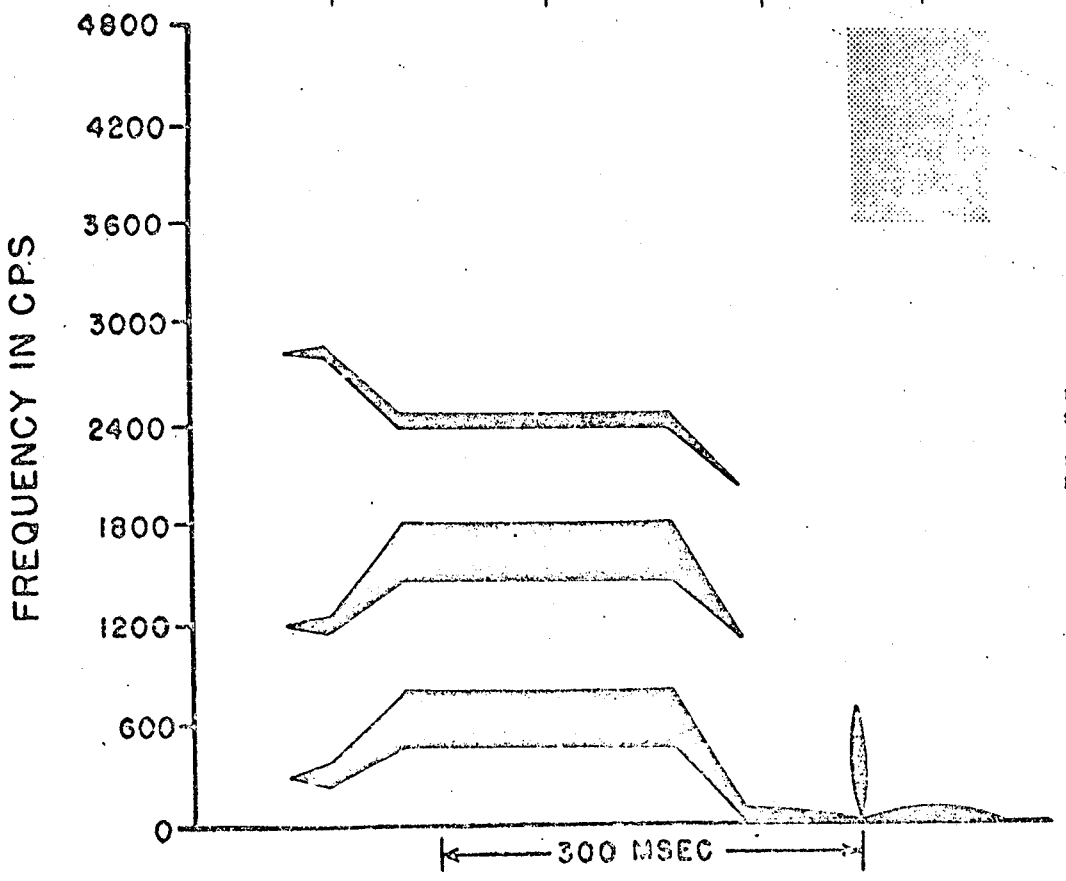

FIG. 2. Table illustrating the rules for synthesizing the word /læbz/, and the pattern derived therefrom.