Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 31, No. 11, 1490–1499, November, 1959
Copyright, 1959 by the Acoustical Society of America.
Printed in U. S. A.

Minimal Rules for Synthesizing Speech*

A. M. LIBERMAN,† FRANCES INGEMANN,‡ LEIGH LISKER,§ PIERRE DELATTRE,‖ AND F. S. COOPER
Haskins Laboratories, New York, New York

(Received July 23, 1959)

It has been found to be extremely difficult to isolate phonemic elements from recorded utterances or to synthesize speech by assembling prerecorded phonemic segments. One reason for the difficulty lies in the fact that the perceptually discrete phonemes are typically combined, and in some cases encoded, into units of essentially syllabic dimensions. As a result, prerecorded elements must, in many cases, approximate syllables, and the synthesis of speech by this means will require a large inventory of recordings. By taking advantage of knowledge about the acoustic cues for speech perception, however, it is possible to write rules for synthesis in terms of phonemes (rather than syllables) and thus reduce considerably the number of separate rules or items needed. Indeed, one can reduce the number of rules still further by writing them at the level of subphonemic dimensions, *viz.*, place and manner of articulation. Several complicating factors make it impossible to achieve an ideal minimum. First, rules or rule modifiers must be added to take care of certain prosodic and positional variations. Failure to do so not only affects naturalness, but also impairs intelligibility, even at the level of segmental phonemes. Second, it is necessary in a few special cases to have different rules for a single consonant phoneme (or a dimension of that phoneme) before different vowels. This reflects the occasionally complex relation between phoneme and articulation on the one hand and sound output on the other; presumably, this complication would not affect the rules of synthesis for an articulatory model.

A system of rules for synthesis framed largely in terms of subphonemic dimensions is described with reference to an example. Words and sentences of rather high intelligibility have been synthesized by such rules.

INTRODUCTION

DURING the past ten years a series of studies has been carried out at Haskins Laboratories in an attempt to uncover the acoustic cues that underlie the perception of speech. Many different aspects of the problem have been investigated. Some of the results have been published in acoustical, linguistic, and psychological journals, and some have been quietly entombed in the files of the Laboratory. A few members of the staff have been close to all stages of the work, and so, with all the published and unpublished information quite literally at their fingertips, they have been able for some time to paint spectrographic patterns appropriate for the synthesis of almost any utterance. That is, they can paint to order, as it were, simple, schematized spectrograms which, when run through the Pattern Playback or the Voback,[1] produce speech at rather respectable levels of intelligibility. These spectrograms are prepared largely on the basis of research results and without looking at a real speech spectrogram of the utterance being synthesized. To that extent we have for a rather long time now been synthesizing speech by rule. At least in a sense. But not in the sense that the phrase "rules for synthesizing speech" is used in the title of this paper. In the ideal case, and that is what we want to talk about here, the rules would be together in one place, written down for all to see, and they would be perfectly explicit in all particulars, so that a person with no knowledge of speech or spectrograms could, by reference to the rules, synthesize speech as well as anyone else.

Recently, one of the authors of this paper, Frances Ingemann, undertook to prepare just such a set of rules. For this purpose she combed, winnowed, refined, and distilled the material in our files. Her first set of rules for synthesis was described to this Society at the Ann Arbor meeting in 1957,[2] and recorded samples of the results were played.

We do not propose in this paper to set forth all of the rules in detail or to consider the improvements that have been made since Dr. Ingemann's earlier report to the Society. Rather, we intend to talk about rules for synthesis in relation to some general aspects of the processes of speech production and perception. We shall try, then, to organize some relatively familiar data and concepts in terms of their relevance to a somewhat less familiar problem.

It may help in setting the problem to think in terms of a machine that will process a discrete phonemic input in such a way as to produce a speech output. We shall suppose that the information available at the input is in the form of a succession of phonemes such as would result from an analysis of a series of utterances by a competent linguist. Fortunately, we need not be concerned here with the precise nature of the phonemic system that was assumed in making this analysis. For our present purposes it is sufficient to know that these phonemes represent discrete elements of the kind

[*] This paper was read by invitation before the Fifty-Sixth Meeting of the Acoustical Society of America in November, 1958. The work described here was supported in part by the Carnegie Corporation of New York, in part by the Prosthetic and Sensory Aids Service of the Veterans Administration in connection with Contract No. V1005M1253, and in part by the Department of Defense in connection with Contract No. DA49-170-sc-2564.

[†] Also at the University of Connecticut.

[‡] Also at the University of Kansas.

[§] Also at the University of Pennsylvania.

[‖] Also at the University of Colorado.

[1] For accounts of these research tools, see: F. S. Cooper, J. Acoust. Soc. Am. 22, 761–762 (1950); Cooper, Liberman, and Borst, Proc. Natl. Acad. Sci. 37, 318–325 (1951); F. S. Cooper, Some instrumental aids to research on speech, in "Report of the fourth annual round table meeting on linguistics and language teaching," pp. 46–53 (Washington, D. C., Institute of Languages and Linguistics, Georgetown University, 1953); J. M. Borst and F. S. Cooper, J. Acoust. Soc. Am. 29, 777 (A) (1957).

[2] F. Ingemann, J. Acoust. Soc. Am. 29, 1255 (A) (1957).

everyone knows as consonants and vowels. So far as the output is concerned, we ask simply that it be easily intelligible at normal rates of production.

This exercise may be considered to have any one or all of several purposes. On the one hand it may be practical. One thinks, for example, of the synthesizer end of a speech-recognizer band width compression system or, perhaps, of a reading machine for the blind. On the other hand, the aim may be quite academic, and, in a rather specific sense, not too different from that which motivates the linguist. Given that we know something about the acoustic cues for the various phonemes, we should like to systematize the data by deriving from them an orderly set of rules for synthesis, and, ideally, we should like to produce rules that are few in number, simple in structure, and susceptible of mechanization.

SYNTHESIS FROM PRERECORDED ELEMENTS

For the purposes of this paper it will be helpful to begin by assuming that we know nothing about the acoustic patterns that underlie language, and that we are going to try nevertheless to convert a phonemic input to speech. In that case we are likely to consider, as the simplest solution, a system in which an inventory of prerecorded sounds is assigned in one-to-one fashion to the phonemic signals at the input end. In this system the incoming phonemes simply key the prerecorded sounds. If we instrument such an arrangement, we will almost surely find it quite unsatisfactory. Of all the various difficulties that one will ultimately experience with this system, the most immediately obvious will be a noticeable bumpiness and roughness in the output. One thinks, then, of setting up various smoothing operations, and, indeed, it is surely possible to improve the output by such means. But no amount of smoothing will solve what is here a very fundamental, and, by now, familiar problem. One has only to look at spectrograms to see that speech tends to vary more or less continuously over stretches of greater than phonemic length. The patterns rarely break at what might be considered to be phoneme boundaries, and those who have tried to find the acoustic limits of the phoneme have come to know this as the problem of segmentation.

Now none of this should be taken to deny the existence of the phoneme, either as a convenient linguistic abstraction or as a perceptual unit. It indicates merely that the perceptually and linguistically discrete phonemes are often combined and, indeed, in some cases encoded, into units that are more than one phoneme in length. They are not strung together like beads on a string. It is for this reason that one encounters difficulties when he tries to snip phonemes out of a magnetic tape recording, or when, conversely, he tries to synthesize speech from prerecorded phonemic elements.[3]

If one insists, nevertheless, on trying to produce speech from prerecorded phonemes, he is likely to be forced into one of two undesirable courses. One possibility is to employ different recordings, or allophonic variants, of most of the phonemes for most of the combinations in which they occur. This obviously requires a formidable inventory of prerecorded elements. The number of elements can be reduced by creating classes of variants, each class being represented by a single typical form. But this reduction in the number of items is only to be had by severely compromising the quality of the output; in short, the rougher the approximation to proper junctions, the rougher and less intelligible the speech.

An alternative is to try to record, or recover, the speech sounds in very brief form as, for example, in a rapid recitation of the alphabet (plus a baker's dozen of additional sounds). The difficulty, of course, is that the phonemes have now become syllables and the intended synthetic speech has become a kind of "spelling bee." Nor is this difficulty avoidable: a shift from spelling to phonetic pronunciation only shortens and centralizes the vowel that clings to almost every consonant; indeed, it is difficult to imagine how a voiced stop, for example, could possibly be produced or heard without some vowel-like sound preceding or following it. Thus, we see that this alternative does violence to the speech process; moreover, it has but limited practical utility, since the spelling-bee output will not be so readily or rapidly comprehended as ordinary speech with its phonemes in syllabic combination.

What has been said so far does not mean that one cannot work from prerecorded elements. Rather, it suggests that if one wants by this technique to produce speech rather than spelling, and if he prefers not to deal with allophonic variants, then he has got to include among the prerecorded elements a number of units which exceed one phoneme in length. An inevitable result is that the inventory must be very large (as it was found, above, to be for the method of allophonic variants when high-quality speech was desired). Thus, in a recent attempt to synthesize speech from discrete segments, Peterson, Wang, and Sivertsen[4] have used what they call "dyads," a dyad being a segment which contains "parts of two phones with their mutual influence in the middle of the segment." To produce one idiolect by this technique Peterson, Wang, and Sivertsen estimate that some 8000 dyads are necessary. (It should be noted that this number includes provision for three levels of intonation for many of the dyads.)

[3] For the purposes of this discussion it does not matter greatly whether the elements are pronounced and recorded in isolation or, alternatively, cut out of recordings of connected speech and then reassembled into new combinations. Some of the difficulties that arise in connection with the latter procedure are illustrated in Harris' account of his attempt to isolate the "building blocks of speech." See C. M. Harris, J. Acoust. Soc. Am. 25, 962–969 (1953).

[4] Peterson, Wang, and Sivertsen, J. Acoust. Soc. Am. 30, 739–742 (1958).

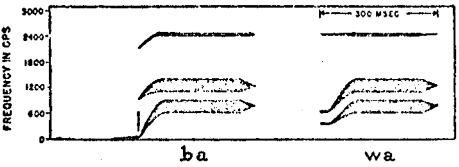

FIG. 1. Hand-drawn spectrographic patterns illustrating some of the acoustic cues for the stop consonant /b/ and the semivowel /w/.

For some purposes such a system may well represent a practical solution. It is not the only solution, however, and from one standpoint not the most interesting. We have in mind here that one may quite properly regard a set of rules for synthesis as a description of the acoustic basis for the perception of language.[5] If so, it must be concluded that discrete segments provide an uneconomical description, since, as has been seen, the number of segments or entries in the system is extremely large.[6]

SYNTHESIS BY PHONEMIC RULES

Although it is very difficult to produce speech from prerecorded phonemic segments, it is nevertheless possible to generate speech from discrete phonemic instructions. That is, rules for synthesis *can* be written which make it possible to go from phonemic units to speech, and thus reduce by a very large factor the total number of rules needed.[7] This can be done by taking advantage of what is known about the cues for speech perception.

The patterns of Fig. 1 illustrate some of these cues and also point to one of the reasons why it is so very difficult to cut and re-assemble phonemic segments. When converted into sound by the Pattern Playback, the hand-drawn spectrograms seen in the figure produce reasonably close approximations to the consonant-vowel syllables indicated. All that we can say about these particular spectrograms that is relevant to the present discussion has been said at other times in talks before this Society and in published papers.[8] Therefore, we ask the indulgence of the reader and, in return, promise to be brief.

Research with patterns such as those shown in Fig. 1 has shown that a primary cue for the perception of these and certain other consonants is the relatively rapid shift in the formant frequencies seen at the left of each pattern. These shifts have been named "transitions," which is unfortunate because this designation implies that they are mere incidents in the process of going from phoneme to phoneme. Far from being incidental links between phonemes, these transitions are themselves among the most important cues to the perception of many of the consonants. It cannot be too strongly emphasized that the perceptual function of the transitions is not to avoid clicks and thumps, but rather to provide important and sometimes essential information for phoneme identification. This is to say that the essential perceptual cue is sometimes given by information concerning the change from one frequency position to another. For the consonant phonemes of Fig. 1, and for others too, it is unqualifiedly true that there is no position in the pattern that will be perceived as the intended consonant, or, indeed, as any consonant, when it is in steady state. Sounding the initial steady-state portion of /w/ will cause the listener to hear the vowel /u/. Every point on the transition leading into the steady-state vowel will, if prolonged, produce a vowel-like sound.[9,10] The listener will perceive /w/ only if he is given information about where the formant begins, where it ends, and how long it takes to move from the one frequency to the other. Normally, this information is conveyed continuously by the transitions. It is always possible, of course, to degrade the patterns to some degree, as for example by erasing parts of the transitions, without utterly destroying the phoneme as perceived. Indeed, in the case of /w/ one can synthesize it reasonably well by moving from the initial steady state to the steady state of the vowel without actually sounding the transition at all, provided the normal time relationships are preserved.[11] This is a rather extreme case—one cannot remove nearly so much of the transition for the /b/ of Fig. 1 or, indeed, for any of the stop or nasal consonants—and even so it is clear that some indication of the /w/ transitions, as given by the abrupt shift from the initial steady state

[5] We are here concerned only with those aspects of the acoustic pattern that carry the linguistic information.

[6] A description of the acoustic basis of language in these terms is, of course, also incomplete unless the patterns present in each segment are fully described in acoustic terms.

[7] As used in this paper a "rule" will refer to all the statements that must be made in order to specify whatever unit (e.g., phoneme, subphonemic feature, syllable) of the language is being used as a basis for synthesis.

[8] Cooper, Delattre, Liberman, Borst, and Gerstman, J. Acoust. Soc. Am. 24, 597–606 (1952); Liberman, Delattre, Cooper, and Gerstman, Psychol. Monogr. 68, No. 8, 1–13 (1954); Liberman, Delattre, Gerstman, and Cooper, J. Exptl. Psychol. 52, 127–137 (1956); O'Connor, Gerstman, Liberman, Delattre, and Cooper, Word 13, 24–43 (1957).

[9] It was found in an earlier study (see O'Connor *et al.*, reference 8) that in the case of /w/ in initial position a brief steady-state segment at the onset helps to avoid a stop consonant effect, but it is not really essential. One must be careful, however, not to have the steady-state segment exceed about 30 msec, because at longer durations the listener hears a vowel preceding the /w/. It is not always clear in spectrograms of real speech whether or not there is an initial steady-state segment, and, if so, how long the segment is.

[10] With the Pattern Playback it is possible to stop the pattern at any point and determine what that part of the pattern sounds like in steady state.

[11] For the purposes of producing speech by recombining prerecorded phonemic segments, one might take advantage of this possibility with /w/ by isolating something approximating the initial steady state of /w/ which, when spliced in the proper temporal relationship to any of several vowels would, perhaps, produce a fair impression of /w/ plus vowel. This technique would almost certainly not work nearly so well with other consonants, and it will in any case probably be harder to do with real speech than with the idealized, schematized, hand-drawn patterns described in the text. In general, we should expect the application of this technique to be somewhat limited and to produce something less than ideal results, for at best it represents a way to force speech into a wholly unnatural mold.

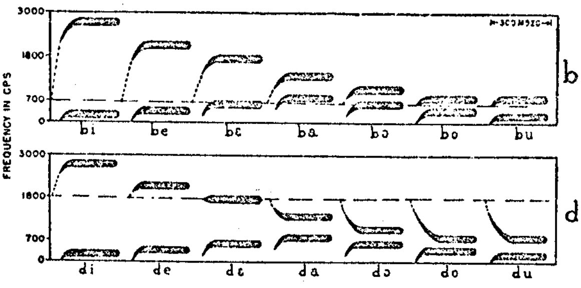

FIG. 2. Second-formant transitions appropriate for /b/ and /d/ before various vowels.

to the vowel, is a necessary condition for the perception of the /w/ phoneme. These considerations lead us to disagree with an assumption that Peterson, Wang, and Sivertsen took as basic to their segmentation technique, namely, that "the intelligibility of speech is carried by the more sustained or target positions of the vowels, consonants, and other phonetic features."[4] We would rather say that for many of the consonants an important and sometimes necessary condition for intelligibility is that the listener be provided with information concerning the direction, extent, and duration of formant "movement." When we consider that this information is normally present in formant transitions, and that it cannot really be dispensed with, we see one of the reasons why it is so difficult, starting with recorded utterances, to isolate and recombine phonemic segments.

Now to arrive at "phonemic" rules for the generation of syllables like those of Fig. 1, we begin by taking into account that all the transitions for a given consonant have a common feature. This is illustrated in Fig. 2, where we see in the bottom row that, although the extent and direction of the transitions are different for /d/ before different vowels, it is nevertheless clear that the transitions have originated from approximately the same place. This common origin has been called the locus,[12] and it has been possible to define characteristic loci for essentially all the consonants.

Knowing the first-, second-, and third-formant loci for all the consonants is the key that unlocks the syllable and makes it feasible to write rules at the phoneme level. For example, we may say of /d/ that its second formant should start at about 1800 cps and proceed then at a certain rate to the steady-state level appropriate for the second formant of the following vowel. If, alternatively, we want to synthesize a syllable consisting of /b/ plus vowel, we see from the patterns in the top row of the figure that we should start the second formant at about 700 cps and proceed to the vowel level from there. In fact, the situation is somewhat more complicated than this in several ways. For example, the stops must not actually start at the loci—rather, they should only "point" to them. In the patterns of the figure the dashed lines represent nonexplicit portions of the complete transition specified by the locus hypothesis. This characteristic of the locus is one of the class markers for the stops,[13] as it is also for the nasals. For these and other classes of consonants it is, of course, necessary to add other acoustic cues, such as the noises that occur with stops, affricates, and fricatives,[14] and the relatively brief steady-state resonances that mark the nasals, liquids, and semivowels.[15]

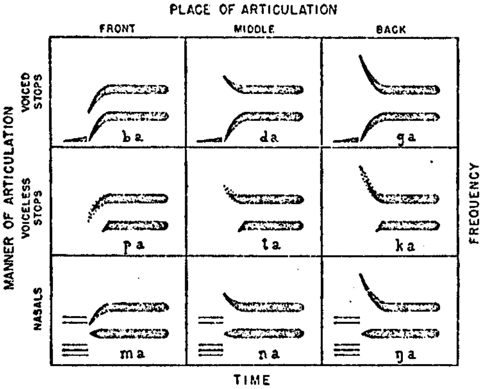

FIG. 3. Patterns illustrating some of the acoustic cues for the stop and nasal consonants.

At a different level of complication it is, as we have already implied, necessary that the application of a phoneme rule be made in relation to the phonemes on either side. Thus, in the example used, the second-formant transition for /d/ led to the second-formant level appropriate for the next vowel, wherever that might have been. This means that contextual information must be used in *applying* the rules for successive phonemes, but only to the extent that one must know—as he must in any case—the appropriate formant levels for the next phoneme so that the transitions may be properly connected. Given that the situation is even approximately this simple, we can see how, in principle, the number of rules can approximate the number of phonemes.

SYNTHESIS BY SUBPHONEMIC RULES

But if economy in terms of number of rules is our aim—and it would appear to be a reasonable one—we can go further by setting up the rules in terms of subphonemic dimensions. Figure 3 contains hand-drawn

[12] For a detailed treatment of the "locus," see Delattre, Liberman, and Cooper, J. Acoust. Soc. Am. 27, 769–773 (1955); Harris, Hoffman, Liberman, Delattre, and Cooper, J. Acoust. Soc. Am. 30, 122–126 (1958). A rationalization in terms of articulatory-acoustic considerations is contained in a paper by Stevens and House [see K. N. Stevens and A. S. House, J. Acoust. Soc. Am. 28, 578–585 (1956)]. In certain ways the locus is similar to the "hub" [see Potter, Kopp, and Green, *Visible Speech* (D. Van Nostrand Company, Inc., New York, 1947)].

[13] See Delattre *et al.*, reference 13; O'Connor *et al.*, reference 8.

[14] Liberman, Delattre, and Cooper, Am. J. Psychol. 65, 497–516 (1952); C. Schatz, Language 30, 47–56 (1954); C. W. Hughes and M. Halle, J. Acoust. Soc. Am. 28, 303–310 (1956); Halle, Hughes, and Radley, J. Acoust. Soc. Am. 29, 107–116 (1957); K. S. Harris, Language and Speech 1, 1–7 (1958).

[15] Liberman *et al.*, reference 8; O'Connor *et al.*, reference 8.

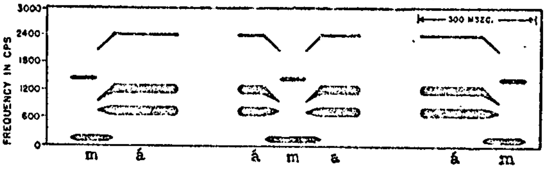

FIG. 4. Patterns illustrating some of the cues for /m/ in different positions.

spectrographic patterns that illustrate how this can be done. Here we see hand-painted spectrograms that will produce reasonable approximations to the syllables /ba, da, ga, pa, ta, ka, ma, na, ŋa/. All the sounds having the same place of articulation—that is, all the sounds in a given column—have the same second-formant transition. Similarly, all the sounds having the same manner of articulation—that is, those in a given row—have the same first-formant transition, and, in some cases, additional markers, as for the nasality of /m, n, ŋ/. Thus, it is possible to set up a rule for a front place of articulation, a middle place of articulation, and a back place of articulation. Similarly, there is a rule for the class of voiced stops, one for the voiceless stops, and one for nasality. In this way we obtain nine phonemes with six rules.

It should be noted here that when the rules are written at a subphonemic level, arrangements must be made for simultaneous (as well as sequential) combination. Thus, for the consonant phoneme of a syllable, for example, we must put together, at the very least, the appropriate rule for place of articulation and the appropriate rule for manner; these, in turn, must be "meshed" with the rules for the vowel or other consonants of the syllable.

As we have seen, the number of rules is considerably reduced by operating at a subphonemic level. In the ideal case we would, of course, have only as many rules as there are subphonemic features, and this would be in the neighborhood of ten. However, for reasons which will be given below, it is not possible to achieve this ideal.

ADDITIONAL RULES FOR POSITION

One complication at either the phonemic or subphonemic level is that we must sometimes make special provision for positional variations. The few simple examples so far have been of consonants in initial position. Now in most cases it is possible to produce patterns suitable for other positions from the same basic rules.[16] That is, it is usually possible to frame a basic rule for a phoneme or a subphonemic dimension and then derive the particular patterns for each of several positions. As an example, let us take the patterns for the nasal labial consonant /m/ in initial, intervocalic, and final positions, as shown in Fig. 4. The basic rules for /m/ require that there be steady-state formants of specified duration, intensities, and frequencies. Furthermore, they require that any adjacent formants have transitions of a specified duration which are discontinuous with the nasal formants and which point to certain locus frequencies. As we see from Fig. 4, the differences among the initial, intervocalic, and final patterns for /m/ involve only the presence or absence of transitions on either side of the nasal formants. Whether or not a transition is to be drawn depends on whether adjacent formants are specified, and that depends, of course, on the rules appropriate for the immediate neighbors of /m/ in the sequence of input phonemes. In other words, before we can have a transition we must have, at the input, two contiguous phonemes both of whose rules call for this acoustic feature.

The preceding example illustrates the most common type of positional variation that must be accommodated by our rules of synthesis. As we have elected to handle them, such positional variations follow from the different ways in which rules for adjacent phonemes "mesh" to specify the transitional portions of our patterns; therefore, additional "connection" rules are not necessary.

In certain cases, however, it is not possible to derive a desired pattern entirely from the basic rules for the constituent phonemes, although we are never forced to the extreme of having to write an entirely new rule for such a case. Rather, we find that an appropriate pattern can be produced simply by applying a qualification or "position modifier" to the basic rule. An example of this is the pattern for the syllable /glu/ shown in Fig. 5. The basic rule for /g/ calls for an interval that is silent except for a voice bar, followed by a burst, and it further stipulates that adjacent formants have transitions which point to particular locus frequencies. The rule for /l/ calls for steady-state formants of a certain duration and specified intensities and frequencies, and it further requires that these /l/ formants be continuous with transitions to any adjacent formants. The rule for the vowel /u/ specifies the duration, intensity, and frequency of each of three formants which are steady state, except as rules for neighboring phonemes prescribe transitions. Now a rigid application of the basic rules for the phonemes constituting the syllable /glu/ yields an ultimate acoustic output of less than tolerable intelligibility. A marked improvement is achieved if the basic rule for each phoneme is modified as follows: /g/ before /l/ requires only a burst of specified frequency; /l/ before /u/ has the frequency of its second formant lowered somewhat; /u/ following /l/ has a second formant which first rises from the second-formant frequency of /l/, and then, after a specified duration, shifts at a given rate to the normal steady-state frequency for /u/. At this point it should be remarked that these position modifiers operate on classes of phonemes; thus, the modification for /g/ applies also to the other stops, the modifier for /l/

[16] L. Lisker, Word 13, 256–267 (1957); L. Lisker, Language 33, 42–49 (1957).

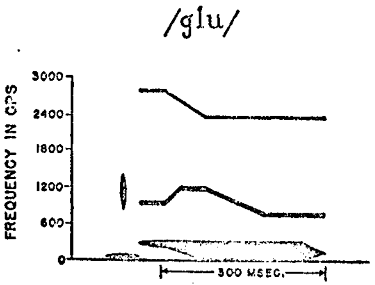

FIG. 5. Pattern appropriate for the syllable /glu/.

applies also to /w, r, j/, and the modifier for /u/ applies also to /o/ and /ɔ/. In other words, the kind of economy gained by going from phoneme rules to subphonemic rules extends to the position modifiers as well.

Similar problems occur and similarly general solutions are found for other positional variations, as, for example, in the neighborhood of juncture.

LINGUISTIC DIGRESSION

We should like to digress here to discuss briefly the implications of what has been said for the problem of how to define the phoneme. As you remember, we began by referring to a machine that would process a discrete phonemic input so as to produce a speech-like output. The phonemic input would be furnished by linguistic analysis. We might soon discover on consulting several equally competent linguists that they were of divided opinion on two subjects at least. First, they might have different ideas on the best way to define the phoneme; and second, they would not agree entirely on what the phonemes of a particular language are. Now the first point may be dismissed as a bit of academic quibbling, for we observe that two linguists with conflicting definitions of the phoneme can come out with phonemic analyses that are remarkably alike. It is of interest, nevertheless, that at least one linguist, Zellig Harris,[17] has proposed an operational definition of the phoneme which would require for its application that we synthesize speech from prerecorded utterances cut up into segments of phoneme length. For example, if the question were whether two sounds in different environments were or were not the same phoneme, one would interchange the appropriate snippets of tape, play back, and listen to determine whether the resulting utterance, as perceived, was reasonably satisfactory. Now we know that in many cases this operation cannot really be performed satisfactorily, and therefore has very little utility as a tool for phonemic analysis. However, the linguist may be able to do a roughly equivalent thing in terms of our rules and their modifiers. For the case of the sounds in different environments, the question would be whether one could satisfactorily synthesize them by using the same rule in both cases, provided only he applied the appropriate positional modifier.

The second point of dispute among the linguists is more important to us, since it actually affects what is to go into the input of the synthesizer. For example, one linguist will transcribe the vocalic part of the word *cake* with a single symbol where another will write it as a sequence of two. Then again, they may have differences of opinion about where to put the phonemes that sometimes mark boundaries between words. Instead of waiting for the linguists to resolve these conflicts among themselves, we might try each of the alternative analyses they provide, and then select that one which yields the most intelligible and natural-sounding speech output. Of course, if two alternatives yield the same kind of results by this test, then we may conclude that the problem is phonetically irrelevant and hand it back to the linguists.

ADDITIONAL RULES FOR STRESS AND SYLLABIC ENCODING

Before this digression into linguistics, we were considering the necessity of adding rules beyond the ideal minimum, and had discussed the matter of positional variations.

There remain two other types of complication that deserve mention. The first of these arises in connection with prosodic features, particularly stress.[18] We might have supposed that the basic rules, derived as they are largely from experiments with isolated syllables, would, if anything, yield connected speech that is "over-intelligible" to the point of sounding stilted. Now the speech we get certainly sounds stilted if differences in stress are not provided for, but it is also often markedly less intelligible than would be predicted from the levels of intelligibility achieved for its constituent vowels and consonants when these are tested in nonsense syllables. The quality of the synthetic speech is significantly improved, both in intelligibility and in naturalness, if at least two degrees of stress are provided for in the rules. The stress differences can be specified by one or more acoustic features, such as fundamental frequency, intensity, and duration. (Fundamental frequency is also, of course, the basis for variations in intonation, but no attempt has yet been made to include this feature in the rules.) At the present time only duration is actually being used in the rules for stress.

In order to achieve the greatest gain from adjusting vowel durations for two degrees of stress it is necessary to reduce the durations of some vowels, specifically those in medial unstressed syllables, to such an extent that no steady-state remains. By the rules for stressed syllables, a simple consonant-vowel-consonant pattern consists at the very least of an initial transition, a

[17] Z. S. Harris, *Methods in Structural Linguistics* (University of Chicago Press, Chicago, Illinois, 1951).

[18] D. B. Fry, J. Acoust. Soc. Am. 27, 765–768 (1955); D. L. Bolinger and L. J. Gerstman, Word 13, 246–255 (1957); D. L. Bolinger, Lingua 7, 175–182 (1958); D. B. Fry, Language and Speech 1, 126–152 (1958); D. L. Bolinger, Word 14, 109–149 (1958); L. Lisker, J. Acoust. Soc. Am. 30, 682 (A) (1958).

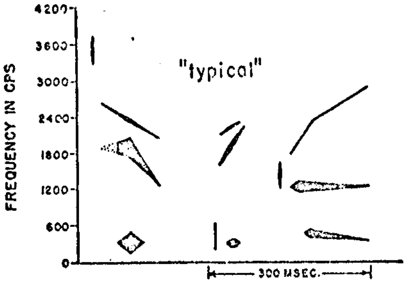

FIG. 6. Pattern appropriate for the word "typical."

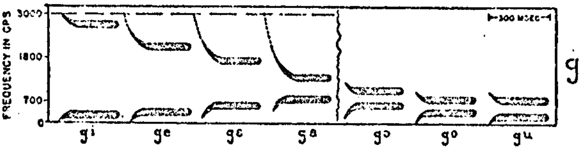

FIG. 7. Patterns illustrating second-formant transitions appropriate for /g/ before various vowels.

steady-state segment, and a final transition. The steady-state segment has formant frequencies characteristic of the vowel alone; the transitions have durations and end points fixed according to the place and manner rules for the consonants. To convert such a syllable into a form appropriate for the unstressed condition, we must effectively omit the steady-state segment, as pointed out above. This means that the second and third formants are in fact drawn as straight lines connecting the end-point frequencies given by the place rules for the adjacent consonants. (It is necessary that the second formant pass through the 1000–2000 cps region. Where the straight line rule would violate this restriction—as, for example, in the case of a vowel between two labials—the second formant must be curved to bring it up or down into the required frequency range.) The configuration of the first formant will depend on whether the adjacent consonants are voiced or voiceless: if voiced, the first formant will move from its initial frequency to 500 cps and then to its final frequency; if voiceless, it will remain at a steady-state frequency of 500 cps. In Fig. 6 the pattern for the word "typical" shows an unstressed vowel between voiceless stops drawn according to this rule.

The other kind of complication is infrequent enough to be of no great practical consequence, perhaps, but is of some interest nevertheless. This difficulty arises because there is occasionally a rather complex relation between the phoneme as a perceptual unit and the sound that elicits it. An example is given in Fig. 7. Here we see a single locus for /g/ before the vowels /i, e, ɛ, a/, but between /a/ and /ɔ/ there is a large and sudden jump to a new locus. It is of more than passing interest that there is no corresponding break in the articulation of the consonant. Between /a/ and /ɔ/ there is a change from unrounded to rounded in the articulation of the vowel, and we may suppose that some rounding of the vowel concomitant with consonant articulation produces the sudden shift in consonant locus. It remains true, however, that so far as the consonant articulation itself is concerned, there is no discontinuity. This is to say, then, that the relation between articulation and phoneme is more nearly one-to-one than that between phoneme and sound. We have in other papers discussed the reason for and the possible significance of this fact.[19] Here we will simply note, first, that this requires an addition to the acoustic rules; second, that it must occasionally wreak havoc with attempts to work from prerecorded phonemic segments; and third, that this complication would not affect the rules of synthesis for an articulatory model. We should also stress the point, so clearly evident in this instance, that very often phonemes are literally *encoded* into syllables at the acoustic level; in such cases the syllable becomes, in a very real sense, the irreducible acoustic unit.

COMPLEXITY *VS* NUMBER OF RULES; RESULTANT INTELLIGIBILITY

We have so far talked about the number of rules required to do the job as if the matter of number were the only significant dimension of this problem. It is not. Obviously, we must consider not only the number of rules but also their simplicity. A rule, as we have been using the term, includes all the statements that must be made in order to specify a given unit of the system. Thus, at the subphonemic level, a rule includes all the specifications for a bilabial place of production, for example, or a stop consonant manner. A given rule may require many specifications or only a few. Simplicity or complexity is largely independent of the number of rules, and becomes, therefore, a separate consideration.

In the discussion so far we have also, perhaps, given the impression that there is a single, ideal, and final set of rules. It is, we suppose, obvious that beyond a certain point reduction in the number of rules, or an increase in their simplicity, will be accomplished only at the cost of naturalness and intelligibility. It remains to be determined just how and within what limits intelligibility and naturalness will vary as a function of the number of rules. In this paper we have more or less implicity assumed some particular and reasonable level of intelligibility, and have considered the *minimum* set of rules for that level. At present there are nine rules for place of consonant articulation, five for manner of consonant articulation, and three rules for voicing. For the vowels we have two manner and twelve place rules. In addition, we have one stress modifier and about twelve position modifiers. We should emphasize that there is nothing hard and fast about the numbers cited; they serve only to indicate roughly how large an inventory we are currently dealing with.

[19] See Liberman, Delattre, and Cooper, reference 14; A. M. Liberman, J. Acoust. Soc. Am. 29, 117–123 (1957).

We have made no comprehensive attempt as yet to measure the intelligibility of the speech produced by such rules. This is not because we are uninterested in intelligibility, but rather because the rules have been changing rapidly, and the data on a really long test are likely to be out of date by the time they have been tabulated. We think it is safe to say that the intelligibility is of a fairly high order. In a few short and rather informal tests, sentence intelligibility has ranged between 60% and 100% depending on the nature of the sentences and the extent to which the listeners are accustomed to hearing this kind of synthetic speech.

EXAMPLE OF THE RULES AND THEIR APPLICATION

To illustrate just how the various categories of rules are combined to specify a pattern let us derive the pattern for the word "labs" as shown in Fig. 8. This word is represented in the input language by the sequence of phonemes: /læbz/.

/l/: The first phoneme is a member of the class of resonant consonants, i.e., /w, r, l, j/. The *manner* rule for the resonants calls for three formants to be maintained (with specified intensities) at appropriate locus frequencies for 30 msec. The manner rule further specifies that adjacent formants shall have transitions of 75 msec drawn so as to be continuous with the locus formants. (This manner characteristic is referred to in the table of Fig. 8 as an "explicit locus.") The manner rule for the resonants also fixes the first-formant locus at 360 cps. Lastly, the resonant manner rule specifies a sound of the harmonic or "buzz" type. The *place* rule for /l/ specifies locus frequencies of 360, 1260, and 2880 cps.

/æ/: The next phoneme of the input is a member of the class of long vowels. The manner rule for this class calls for three formants of the buzz variety, having a duration of 150 msec. The place rule for /æ/ fixes formant frequencies at 750, 1650, and 2460 cps, and also specifies formant intensities.

/b/: The next phoneme shares its manner rule with all the other stops, /b, d, g, p, t, k/; this rule calls for an interval of "silence" (an interval devoid of acoustic energy at all frequencies above the fundamental of the buzz) followed by a burst, and further specifies that adjacent formants have 50-msec transitions pointing toward locus frequencies given by the place rule appropriate to the particular stop. ("Pointing to" means that the end point frequencies of the actual transitions are midway between the locus frequencies and the formant frequencies of the next phoneme. This characteristic of the stop consonant manner is referred to in the table of Fig. 8 as a "virtual locus.") The manner rule also fixes the locus of the first formant at the frequency of the voice bar. The labial place rule, which serves equally for /b, p, m, f, v/, specifies that adjacent second- and third-formant transitions point to frequencies of 720 and 2100 cps, respectively. The *voicing* rule for stops (applicable equally to /b, d, g/) requires that the duration of the "silent" interval be 70 msec and that this interval be filled by a "voice bar," that is, acoustic energy at the buzz fundamental frequency.

/z/: For the final phoneme, the manner rule is that appropriate for the fricatives, /f, v, θ, ð, s, z, ʃ, ʒ/, and it calls for an interval of band-limited noise (that is, a "hiss" rather than a buzz sound). The fricative manner rule also specifies that adjacent formants have 50-msec transitions pointing toward virtual loci given by the place rule for the particular fricative; further by the manner rule, the first-formant locus is at 240 cps. The alveolar place rule for either /z/ or /s/ specifies that the noise (required by the fricative manner rule) have a lower cutoff frequency of 3600 cps, and that adjacent second- and third-formant transitions point to frequencies of 1800 and 2700 cps, respectively. The voicing rule states that the noise should be of low intensity, have a duration of 100 msec and be accompanied by a voice bar.

Finally, we apply a *position modifier* for syllables immediately before silence or juncture which doubles the duration of the vowel, making the over-all duration of /æ/ 300 msec.

At this point we have completely specified a pattern that is directly convertible to an acoustic stimulus which the naive listener will readily identify as the word "labs."

RULES *VS* PRERECORDED ELEMENTS: FURTHER DISCUSSION

Instead of trying now merely to summarize what has already been said, we would rather try to bring into the open a few considerations that have been only implicit in the discussion so far. In particular we should like to call attention to the fact that we have casually mixed two rather different aspects of the problem. The first has to do with the size of the unit in terms of which the rules are written, the second with the difference between assembling prerecorded elements on the one hand and the real honest-to-goodness fabrication of speech on the other. When we introduced the matter of prerecorded elements earlier, it was primarily to make a point about speech. This was somewhat unfair. Although the use of prerecorded elements is synthesis only in the most sweeping sense of the word,[20] it is sufficiently interesting both in practice and in principle that we ought to deal with it in its own right. The practical advantages of such a system are obvious enough. The difficulty, as we tried to point out earlier, is largely in the matter of linkage. As we have seen, linkage presents great difficulties at the level of phonemes. Indeed, the difficulties are likely to be so great

[20] The speech sounds in the prerecorded elements are of course produced by human articulatory apparatus and recorded just as any other utterances can be. The synthetic aspect of this process consists only of entering these elements into combinations different from those in which they were originally recorded.

SYNTHESIS BY RULES: /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.		

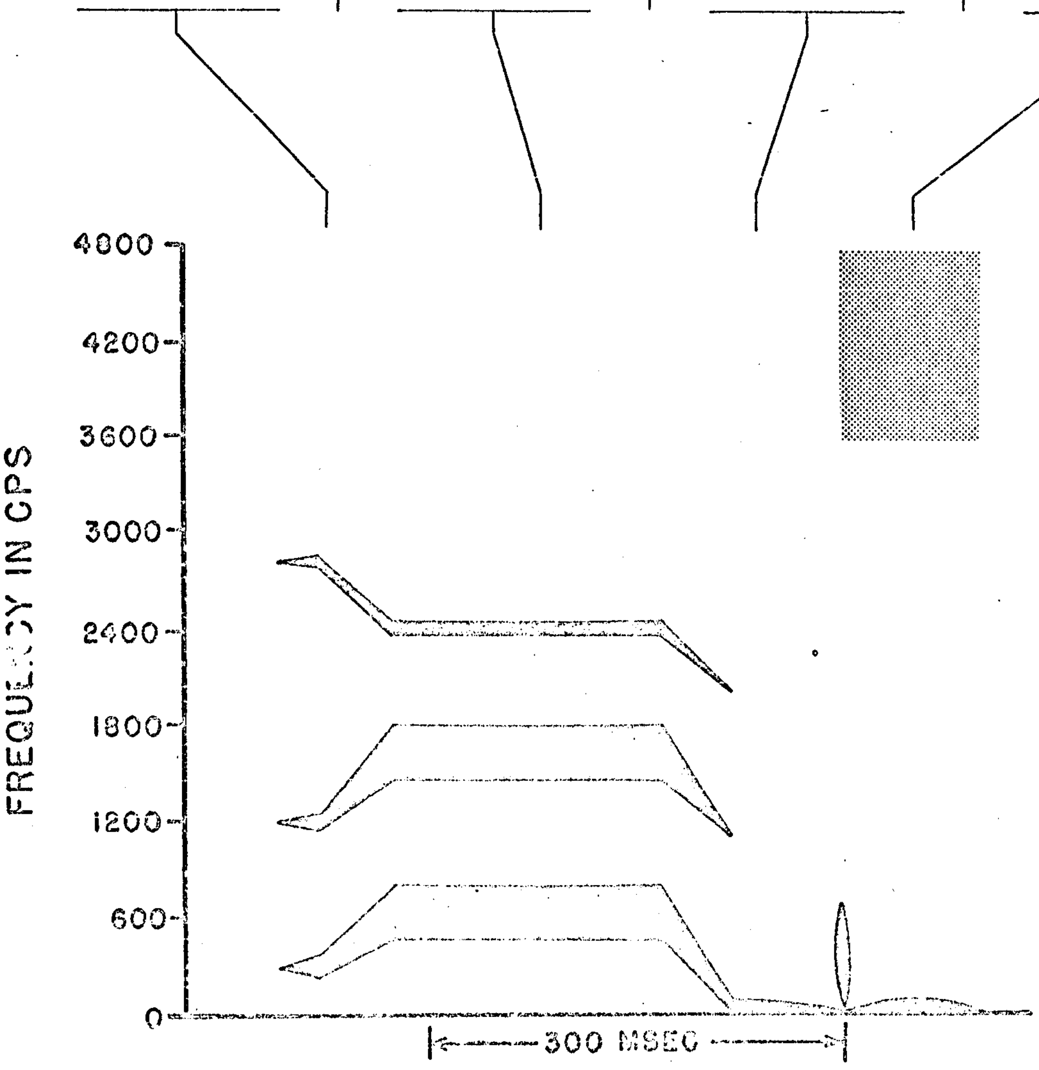

FIG. 8. Table illustrating the rules for synthesizing the word /læbz/, and the pattern derived therefrom.

that one will be driven to use elements which are of essentially syllabic dimensions, and even here problems will occur in the matter of joining the syllables. With units the size of words one will, of course, have much less difficulty about linkages, though he may not, even so, be completely out of the woods.

If we have seemed earlier in this discussion to be unenthusiastic about synthesis from prerecorded elements, we should say now that we have been sufficiently interested ourselves to have begun to explore the possibilities of such a system, at least at the level of words.[21] We have been particularly interested in trying to find the minimum number of versions of each word which will produce appropriate stresses and intonations when the words are arranged in various combinations. This has proved to be a challenging and interesting problem in the sense that its solution will either depend on the application of already known linguistic principles, or, alternatively, will provide information basic to the formulation of such principles.

To return to the point about linkages, the obvious generalization is that the problem grows less severe as the size of the prerecorded unit increases. There is, presumably, a function relating intelligibility or maximum speed of communication to size of the prerecorded units, and this function must certainly rise, though at an ever decreasing rate, as we go from smaller to larger units. At present, we know that with prerecorded phoneme units we are way down on the intelligibility or speed scale, if, indeed, we are on it at all. With prerecorded words we may be within shouting distance of the asymptote. We strongly suspect, without benefit of evidence, that syllables will be marginally useful if we want to communicate at normal speech rates.

It may well be that, for some purposes, prerecorded elements will turn out to be the method of choice. In principle, the system is interesting because when the units are of phonemic dimensions the difficulties one encounters illustrate some important truths about speech, and when the units are the size of words, we encounter some partially soluble and therefore challenging problems of stress and intonation—as well as of instrumentation.

We have already dealt at length with true synthesis as opposed to the use of prerecorded elements. With true synthesis the linkage problem is soluble at all levels, and has to a large extent been solved. The rules for synthesis can be written at various levels. Indeed, this system is inherently flexible in all respects. Its limits are set primarily by the limits of our knowledge about speech and, from a practical standpoint, by the difficulties of instrumenting true synthesis rather than random access.

We saw that with prerecorded elements the total inventory of segments (or rules) will likely approximate the number of syllables at the very least. By using true synthesis we can considerably reduce the number of rules by writing them either at the phoneme or subphoneme level. In either case some rules must be added to take care of positional variations, essential prosodic features, and the special cases in which the acoustic encoding of the phonemes into syllables makes it impossible to get along with only one rule for a single phoneme or subphonemic feature. When the rules are written at the phoneme level there must, of course, be provision for connecting the formant transitions or formants of successive phonemes, and at the subphonemic level additional arrangements must be made for simultaneous combination of the rules pertaining to the several features that constitute the phoneme.

Exactly how and where one might wish to make practical use of the rules for true synthesis depends on a large number of considerations that lie far outside the scope of this paper. In this account, we have been interested in such rules primarily because they constitute a description of the acoustic basis of speech perception. The kind of information contained in that description will, we think, prove useful for a variety of practical and theoretical purposes.

[21] "Summary of fourth technical session on reading machines for the blind," Veterans Administration Washington, D. C., August 23–24, 1956; prepared by the Prosthetic and Sensory Aids Service, Veterans Administration, 252 Seventh Avenue, New York 1, New York.