Speech Synthesis by Rule as a Research Technique
Ignatius G. Mattingly*
Haskins Laboratories, New York City

"Speech synthesis by rule" means the generation of synthetic speech utterances from phonemic information -- including both segmental and prosodic phonemes. The input looks like this:

A 'B ER D IN DH A 'H AA N D I Z W ER TH 'T UU IN DH A 'B OO SH.

Segmental phonemes are represented by letters or sequences of letters; prosodic phonemes, by stress marks and intonation contour marks. From such information a set of rules is used to calculate the control data for a speech synthesizer, that is, values for each of several parameters for successive time increments. Synthesis by rule thus differs from other types of speech synthesis, in which the control data may be derived in accordance with an experimental plan (as in psycholinguistic investigations); or by automatic analysis of natural speech (as in a vocoder); or by manual analysis of natural speech, as in "direct synthesis," the object of which is to produce as faithful an imitation of some natural utterance as the constraints of the synthesizer will permit.

The first investigator to attempt synthesis by rule was Homer Dudley of Bell Laboratories, who demonstrated a device called the Voder at the 1939 World's Fair. It consisted of the synthesizer of a Dudley Vocoder operated by control signals from keys and foot pedals. The rules were implicit in the training of the operator. In describing the Voder, Dudley et al. (1939)

† An Invited Paper given at the 73rd Meeting of the Acoustical Society of America, New York City, 19-22 April 1967.
* Also, the University of Connecticut.
made some comments which still hold true today. First, they marked out the best strategy for synthesis by rule. Rather than trying to reproduce particular instances of natural speech sounds in all their details, they said, "the most fruitful attack was to search for the desired sounds by manipulation guided by the ear." It is the perceptions rather than the productions of the native speaker which provide the surer guide. And Dudley pointed to the snare which ultimately catches anyone interested in speech synthesis: "...the listener becomes expert at interpreting badly formed words and ceases to be critical." No one will quarrel with this who has ever had the experience of demonstrating to an uncomprehending visitor what he has come to regard as perfectly intelligible speech.

Twenty years later Frances Ingemann of Haskins Laboratories brought together the information on the cues for speech perception which Haskins researchers had accumulated during the previous decade, and developed a set of rules for use with the pattern playback synthesizer (Liberman et al., 1959). Her rules consisted of instructions for painting simple patterns on the celluloid belt which drove the playback.

When digital computers became available to speech researchers, interest in synthesis by rule increased. Execution of the rules could now be automated, quite complex rules could be quantitatively formulated, variations on a rule could be readily tested and records of these tests could be conveniently preserved. The first investigators to use a computer for synthesis by rule were Kelly, Gerstman and Lochbaum of Bell Laboratories (Kelly & Gerstman, 1961; Kelly & Lochbaum, 1962). Many others have since entered the field: Dennis (Dennis, 1962) and Lee, both of MIT;
Rabiner (Rabiner, 1966) of MIT and Bell Laboratories; Denes of Bell Laboratories; Estes and his colleagues (Estes et al., 1964) of IBM; Nakata and Mitsuoka (Nakata and Mitsuoka, 1965) of the Radio Research Laboratories in Tokyo. I myself was introduced to the field by Holmes and Shearme (Holmes et al., 1964; Mattingly, 1966) of the Joint Speech Research Unit (JSRU) in England and am continuing this work at Haskins.

Using a computer, two approaches are possible. The first is to program a computer -- which need not be very powerful -- to generate control information for a hardware synthesizer. Preferably the synthesizer should be driven on line by the computer, but this is not essential; at JSRU a paper tape containing control data was produced by a very slow computer located several miles from the laboratory. A tape mover and decoder were then used to transmit the control data to the synthesizer. But of course, on-line control of a synthesizer from control data stored in the core of a laboratory computer like the DDP-224, such as we have at Haskins and Denes has at Bell Laboratories, permits much more flexible experimentation: given appropriate editing routines, a researcher at the computer console can listen to a sentence generated by rule and try various modifications of the control data with a view to improving his rules. The only difficulty in this approach is that the design of the synthesizer itself cannot be readily modified.

The alternative is to simulate the entire system including the synthesizer, on a large conventional computer. The only other hardware needed is some means of converting a digital output to sound, and the design of the synthesizer can be modified as readily as the rules. But unless some sort of time-sharing arrangement is in effect, on-line experimentation with the
well as for steady-states. I think, however, that such a procedure not only is less apt to produce acceptable speech than a set of transition rules, but also makes difficult the study of other aspects of synthesis by rule, unless one is willing to multiply diphones indefinitely. Kelly used simple linear transitions with reasonable success. In the JSRU system a transition was linear from the steady state of one phoneme to a boundary value, and from the boundary value to the steady state of the next phoneme, the boundary value being a function of information associated with the adjacent phonemes. Rabiner's transitions are calculated by solving a differential equation; Nakata used exponential transitions. Haggard uses an approximation to an exponential curve composed of two abutted straight lines. In my own work I propose to use exponential curves for formant frequencies and linear transitions for formant intensities. Though I think such curves will resemble actual formant transitions more closely than do the straight lines we used at JSRU, the difference in perception they will make is not yet known. Such variables as the duration and the initial and target frequencies of transitions may well be of far greater importance than their precise shape.

Prosodic phonemes, such as intonation and stress, have in general received less attention so far than segmental phonemes. Investigators have either been content with very crude prosodic rules or else have supplied the prosodic control information together with the input of segmental phonemes. At JSRU, however, we did develop a prosodic synthesis-by-rule program which provided for increase of pitch and duration on stressed syllables, three intonation contours and prepausal lengthening. Rabiner has also
output of the system is prohibitively expensive.

The synthesizer, actual or simulated, can be either a vocal tract analog or a terminal analog. Dennis envisioned, and Kelly and Lochbaum as well as Nakata and Mitsuoka actually developed, synthesis-by-rule systems using vocal tract analogs, but most other researchers have used terminal analogs. There are several good reasons for this: first the construction (or even simulation) of a satisfactory vocal tract analog is no trivial task. Second, the collection of data on which the rules for a vocal tract analog system must be based requires X-ray photography and other fairly elaborate instrumentation, whereas the rules for terminal analog synthesis can be based for the most part on sound spectrograms. Finally, the rules are essentially linguistic statements which can be stated in acoustic as well as in articulatory terms. On the other hand it is likely that an articulatory statement of the rules would be simpler and more meaningful. For that matter, even greater simplification might be possible if the rules could be stated in motor command terms, as recent work in electromyography at Haskins suggests, so that the ideal synthesizer would be a motor command synthesizer. However, research on synthesis by rule is still a long way from exhausting the possibilities even of terminal analogs.

For most investigators the chief task of the rules has been the calculation of appropriate transitions for the segmental phonemes. Steady-state parameter values for each phoneme can be stored in a table, but the rules must somehow specify how to get from one set of steady-state values to the next. To be sure, Estes et al. of IBM sought to evade this problem by using a table look-up procedure for transitions, or "diphones," as
given considerable attention to the features of stress and intonation following a model proposed by Lieberman (1967).

In my view the art of synthesis by rule has reached the stage at which intelligible speech can be produced. I must immediately add that much depends on such factors as the phonemic composition of the utterance, the mental set of the listener, his previous exposure to synthetic speech and the familiarity of the utterance. A poorly synthesized phoneme which might otherwise pass unnoticed can render an entire sentence unintelligible if it occurs at a point where the listener has little predictive information. Some listeners are much more tolerant than others of synthetic speech. The importance of the familiarity of the utterance was brought home to me when demonstrating the synthesized line "...shoes and ships and sealing wax and cabbages and kings" to a procession of naive listeners at a scientific research exhibition in London. Most people, of course, knew the "Walrus and the Carpenter" and understood the line immediately; but for the minority unfamiliar with the poem, the line might as well have been part of a P-B word list.

But I think it best to let you judge the intelligibility for yourselves by demonstrating a few examples.

[Demonstration]

If one grants that synthesis by rule is possible, he must ask why one should go on doing it. Is it more than an interesting tour de force? I think it is. There are, of course, certain practical applications for synthesis by rule. A reading machine for the blind is one. This is the primary objective of Lee's work at MIT and very much on our minds at Haskins. Reading
instruction for children is another. Again many tasks now performed by stored speech -- stock market quotation services, for example -- could be performed using synthesis by rule with a great reduction in the amount of storage required. But I confess that for me the real interest of synthesis by rule is its potential as a research tool in linguistics. As Lisker et al. (1962) have argued, synthesis by rule is a way of making phonological statements which can be readily tested by listening to the synthetic speech. If the speech is not acceptable to a native speaker of the language being synthesized, the phonology must be wrong. The synthesis system thus plays the role of the student of a language who, having listened to the speech of an informant, himself produces tentative utterances for the informant to accept or reject. Just as the student hopes in this way eventually to master the entire phonology of the language, so in synthesis by rule we seek, ultimately, a set of rules in which the phonology of the language will be completely stated.

But if linguistic research is our objective, and if we wish to attach any theoretical significance to our results, we must be clear about what we are doing. First, let us consider more carefully the question of the input. The grammar of a language includes not only a phonological component but a lexical component, that is, a list of forms such as dog, cat, acoustics, and a syntactic component, that is, a set of rules for arranging and modifying these lexical forms. Once the syntactic component has operated on a group of lexical forms the phonological component must turn the resulting string of forms into sound. This string, the principal input to the synthesis by rule system, is phonemic: it marks only those

3.7
phonological structures whose occurrence is determined outside the phonology itself: segmental phonemes, syllables, junctures, stress, syntactic breaks requiring intonation contours. It contains no information about other structures which are subphonemic, that is, internal to the phonology and for which rules are also required: the basic breath group, the rhythmic occurrence of prominent syllables, coarticulation and transitions from phoneme to phoneme. This distinction between external and internal structure is not so obvious in practice as one might suppose. For example, in our JSRU prosodic rules we marked "stress" and the effect of this mark was uniformly to raise the pitch and lengthen the duration of the marked syllable. But this kind of stress is phonologically determinate. What we ought to have done, strictly speaking, was to mark (a) the syllables which, according to the lexical component, carry potential stress, and (b) the words which the syntactic component selects as prominent. The rules should then have actualized the potential stress in the prominent words.

So far as our meager knowledge permits, we can further divide the subphonemic information into what is peculiar to a given language and what is true for all languages. Thus the basic form of the breath-group, as Lieberman (1967) has taught us, is a language universal. But the pretonic pitch curve of British speech slopes much more steeply than that of American speech. The slope of this curve is thus a language particular. We should reflect the distinction between language particulars and language universals by embodying known universals in our basic program which is good for any language, while storing particulars for a given language in a table. Such a table
becomes a secondary input to our synthesis by rule system.

Finally, we should give some thought to what the output of our system is supposed to represent. Are we constructing a speaker who provides only the minimum information necessary for intelligibility, or a hyperorthodox speaker who puts in all cues consistently, as no human speaker ever does, or a speaker whose style is in some sense natural or typical? My own limited experience suggests that it may be difficult to reconcile the claims of naturalness with those of phonology, that the hyper-orthodox synthetic speaker will seem less natural and perhaps even less intelligible than a speaker who has actually given the listener less information, but has a more natural style. This paradox requires further attention.

When we compare the store of available phonological knowledge with what has so far been embodied in rules, we see that this store is far from being fully exploited. On the other hand, when we compare speech synthesized by rule with directly synthesized speech, we see that there are many deficiencies in our phonological knowledge which further work in synthesis by rule might remedy. From either viewpoint, much remains to be done. Prosodic features require much further study. Of these features duration is by far the most challenging and elusive, the hardest to get right. The reason for this is that duration is determined by so many different phonological structures: both external and internal. The segmental phonemes, the syllable, lexical stress, juncture, prominence, pausing, rhythm and tempo are all involved. Until we understand all of these structures much better than we do now, our durational rules will often be very obviously wrong.

Allophonic variation -- except for the coarticulatory variation
already provided for in most synthesis by rule systems -- also requires attention. From the point of view of the conventional linguist, allophones exist only to complicate his search for phonemes, and in phonological descriptions allophonic variants are apt to be treated as a ragbag of unorganized and disparate phenomena. How are we to incorporate all these phenomena systematically in a synthesis by rule system? We must begin by understanding the linguistic motivation for allophones. I suggest that all allophones except those in free variation have some structural purpose: usually, they serve as a marker of some higher structure: the syllable, the word or the breath group.

Let us consider, in particular, the allophones associated with the syllable. The interesting point about these allophones is that while the phonetic effects are diverse, the number of different conditioning environments is quite restricted. This is not an accident but is owing to the nature of the syllable itself. Though there are many suggestions in the literature (mostly traceable to Stetson, 1951) that the syllable as such has some articulatory basis, doubt has been cast on this notion by several investigators. The syllable is a phonological entity.

The reason for its existence goes back to the fact that the phonetic manifestations of successive phonemes usually overlap. This overlap means not only that information about two or more phonemes is simultaneously transmitted but also that the ensemble of phonemes is larger than would be possible in a non-overlapping system, since some phonemes must use others as carriers. Thus, as Cooper (1966) has pointed out, a high rate of information transmission is achieved using fairly slow-moving
machinery. But an overlapping system would be impossible to decode without some kind of higher level organization. This organization is provided by the syllable, a very ingenious device for maximizing intelligibility on one hand and information rate on the other. The studies of consonant clusters by Bloomfield (1933), Trim and O'Connor (1953) and Jones (1956) permit us to regard the syllable as a finite state machine for getting from one vowel to the next via final and initial clusters, the possibilities for the following phoneme being different at each state. To help the listener keep track of each change of state, the phonological rules of a language provide for various allophones that announce that a vowel has just occurred or is about to occur, that a final voiced stop is about to occur, that an initial /s/ has just occurred, and so on. But since the number of syllable states is quite limited, the number of such conditioning environments is limited in just the same way.

A synthesis by rule system can exploit these properties of the syllable in the following fashion. An analytic routine traces through each syllable and assigns the proper syllable state to each phoneme. This can be done quite simply just because the number of states is small. Included in the rules is a list of states and the modifications each state requires in the specifications for synthesis of a particular phoneme or class of phonemes. Before synthesizing the current phoneme, the program runs through the list, comparing each entry with the state assigned to the phoneme, and making the prescribed modifications if it finds a match. In this way statements about syllabic allophones can readily be incorporated in a synthesis by rule system in a way that is both practical and linguistically significant.

3.11
I hope I have been able to convince you that synthesis by rule can be something more than an interesting trick. At the same time a caution is in order. While this technique provides a powerful means of testing phonological statements, it cannot pretend to simulate the phonological machinery of the brain or the motor system any more than a terminal analog synthesizer simulates the actual behavior of the vocal tract. Nevertheless, I believe that synthesis by rule offers essentially unduplicated possibilities for linguistic research.

References


