A Model for Synthesizing Speech by Rule

LAWRENCE R. RABINER, Member, IEEE
Bell Telephone Laboratories, Inc.
Murray Hill, N. J. 07974

Abstract

A general model for speech synthesis by rule is presented along with a discussion of one specific implementation of the model. The conversion from discrete input signals to continuous synthesizer control signals is performed by the synthesis strategy. The details of the synthesis strategy, including linguistic preprocessing of the input and separate but interdependent segmental and suprasegmental models, are described. An experimental evaluation of the specific model is included, along with specific recommendations as to areas of speech synthesis and speech production requiring further study.

Manuscript received October 9, 1968.

This paper was presented at the Speech Symposium, Kyoto, Japan, August 29, 1968.

Introduction

In recent years there has been a great deal of interest in techniques for the synthesis of speech. Besides the applications in voice answerback systems, reading machines for the blind, and automatic intercept systems, synthetic speech has served as a tool enabling the researcher to learn more about how speech is perceived. By giving the experimenter precise control over the parameters used to synthesize the speech, he is able to determine the effects of each of these parameters.

Speech synthesizers have been highly successful [1], [2], provided the experimenter is able to adjust parameters of the synthesizer after at least one trial run. This process, referred to as hand synthesis, embodies a continuous process of feedback and adjustment until the desired quality, or some approximation to it, is achieved. Of great practical importance are synthesis techniques which produce high quality speech on the first trial; i.e., the speech is synthesized entirely by rule without the need for subsequent feedback and adjustment.

Speech synthesis by rule can be thought of as a method of converting from a discrete representation of speech to a continuous acoustic waveform. Fig. 1 shows a pictorial representation of this transformation. The discrete set of input symbols usually consists of phonemes, sentence punctuation, and vowel stress marks. Other possible input information might be a parsing of the sentence, or information about its syntactic structure.

The synthesis strategy determines the transformation from the discrete symbols to a continuous set of control parameters which are capable of driving a speech synthesizer. The nature of this transformation will be examined in detail in a following section.

The speech synthesizer, the last link in the chain, is the device used to convert from a parametric description of speech to a continuous waveform. Several types of synthesizers have been developed. These include lumped-constant electrical network analogs of the human vocal mechanism (known as serial or parallel terminal analogs, depending on their structure) and acoustic pipe analogs of the vocal tract. Holmes [1] and others have demonstrated, by means of hand synthesis, that speech of high quality can be generated using a terminal analog synthesizer. Terminal analog synthesizers, however, are not well suited for reproducing stop consonants and other sounds involving rapid transitions. The concept of a formant is meaningless when there is rapid change in the vocal tract configuration, and difficulties are encountered in generating these sounds, even approximately, by means of time-varying resonant circuits. Terminal analog synthesizers are also not well suited for modeling unusual source modes such as vocal fry [3]. Present-day terminal analogs do not take into account the source–system interactions of speech, which are not always insignificant. These interactions are most noticeable when the first formant is at a low value, i.e., for consonants.

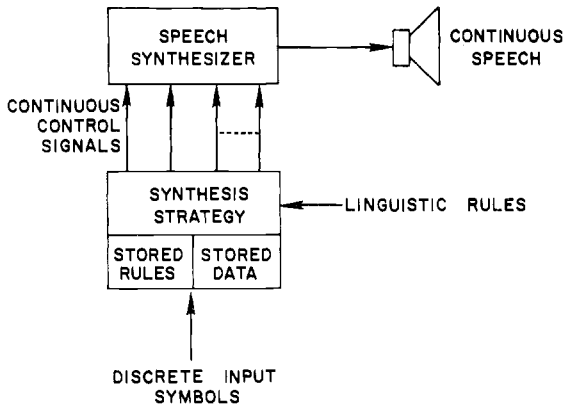


Fig. 1. Model of synthesis by rule.

Acoustic pipe analogs (vocal tract analogs) can be an improvement over terminal analogs in some situations, but they also have their inherent limitations. Comparatively little attention has been paid to this type of synthesizer, and it has yet to be demonstrated that high quality speech can be generated using the pipe analogs. Perhaps a hybrid synthesizer, combining the desirable features of the terminal analog with those of the acoustic pipe analog, will prove most useful in future synthesis applications.

There have been many attempts at speech synthesis by rule in the last ten years. This paper provides a review and discussion of one of these schemes [5]. The synthesizer, the set of input signals, and the synthesis strategy are described in the following sections.

Synthesizer

A computer-simulated serial terminal analog synthesizer was used in this scheme. Fig. 2 shows a block diagram of the synthesizer. The sampling rate at the output is 20 kHz and the control signals are supplied at a 100 Hz rate. There are two excitation sources, a pitch pulse generator for voiced speech, and a frication generator for voiceless speech. The signal processing paths are used to produce the various speech sounds. A detailed description of the synthesizer is available elsewhere [6].

Input Symbols

The set of allowable input signals for this scheme includes phonemes, vowel stress marks, word and sentence markers, pauses, and punctuation. No information as to parts of speech of the words is included.

A typical input string will best illustrate the input formant. The sentence "We saw the cat," with emphasis on the word cat, appears as

W-IY-STR3-SPACE-S-OW-STR2-SPACE-
THE-UH-SPACE-K-AE-STR1-T-END

or according, to IPA notation,

3 2 4 1
/Wi sɔ ðə kaet/.

The phonetic transcription is a bookish one, being derived solely on the basis of individual words. Hence the word an is phonetically /aen/ although it often becomes /ən/ when imbedded in a sentence. A discussion of the conversion from a book transcription to a usage transcription is included in the next section.

Synthesis Strategy

The major purpose of the synthesis strategy is to convert a string of input symbols, such as those described above, to the appropriate set of synthesizer control signals. The input string contains only the information-bearing elements of speech and its rate is on the order of 50 to 100 bits per second [7]. Good quality output speech has an information rate on the order of 30 000 bits per second. Hence most of the redundancy of the output information, and a good deal of nonessential information about the speaker, is not present at the input to the synthesis strategy. Therefore, one goal of the synthesis strategy is to insert by rule the redundancy which was eliminated at the input.

The synthesis strategy can be thought of as a two-stage transformation, as seen in Fig. 3. The first stage involves preprocessing the discrete input sequence, and replacing it with another discrete sequence. This preprocessing occurs in many steps. The first step is to choose a quantum of speech for use in synthesizing segmental features. The phoneme is perhaps the simplest and most often used quantum, but the diphone, syllable, or word may also be useful units to work with.

The second step in the preprocessing makes use of information about English usage. Words in common usage, such as an, the, and, etc., are often pronounced differently in a sentence than in isolation. For example, the transcription /də/ is often used in place of /ðə/ (the); the final /d/ in /aend/ is often dropped; and the schwa vowel /ə/ is used in place of /ae/ in /aen/. The replacement of a book transcription by a usage transcription can often have a significant effect on the output speech. At the present time, the only way the replacement rules can be applied is by exhaustively listing all known cases where such substitutions occur. There are undoubtedly many general cases of these substitutions but they have not been investigated in sufficient detail to uncover them.

A third step in the preprocessing is to apply linguistic rules for consonant and/or vowel substitutions. In selected environments, certain phonemes are predictably replaced by other phonemes. For instance, intervocalic /t/ becomes /d/ as in writer. A voiced consonant, in the region of one or more unvoiced consonants, will often become unvoiced as the /b/ in /abscond/ (/ae b s k a n d/). There are many such rules

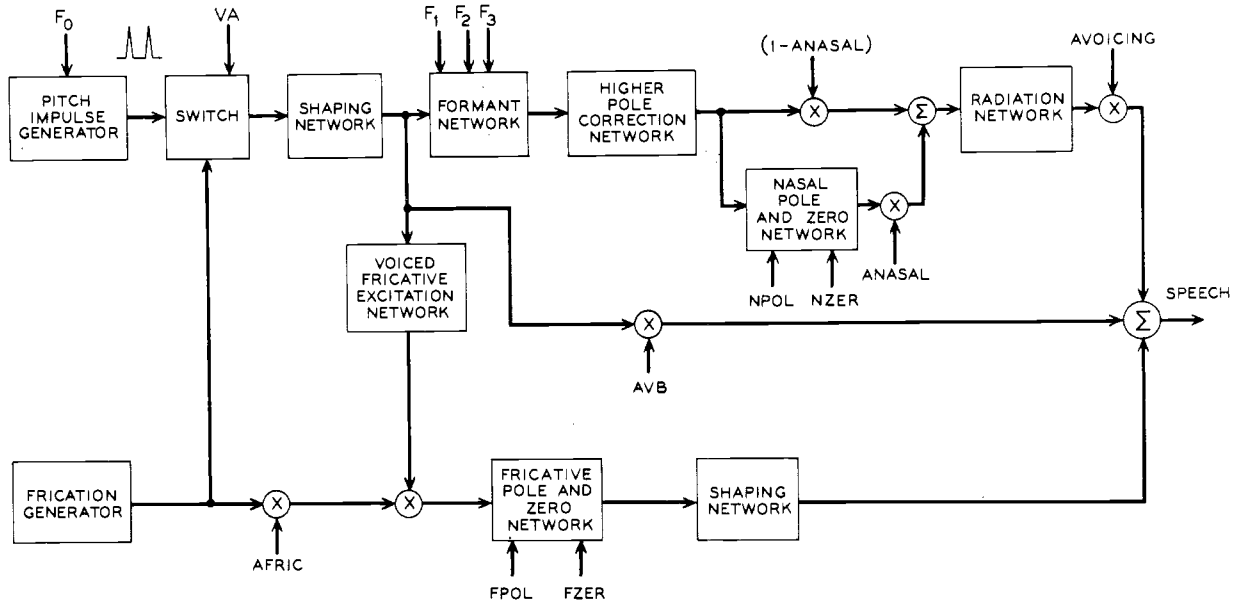
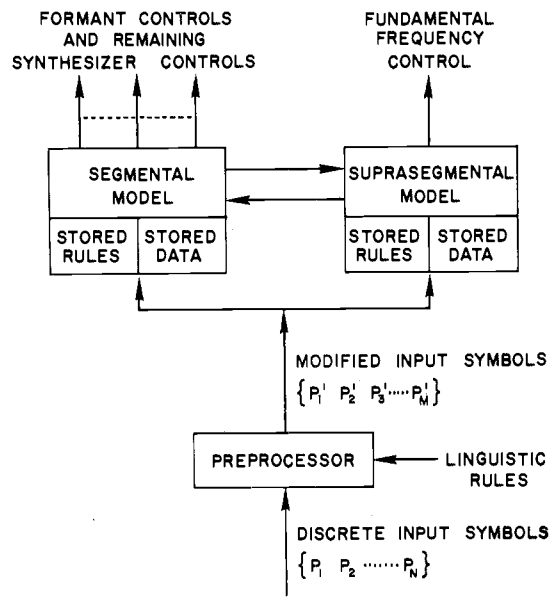


Fig. 2. Block diagram of speech synthesizer.

Fig. 3. Expanded view of synthesis strategy.



for English and their applications for synthesis by rule are important.

The final step in the preprocessing is to make use of information about the surface structure and deep structure of the sentence. Just how one would use this information is unclear with our present knowledge but it does seem that it should provide useful information about large units of the utterance. For instance, a given word plays different roles depending on its relation to the structure of the sentence and one would expect the acoustic characteristics of the word to reflect this result. This step of the preprocessing would provide modifiers tagged to each phoneme, or group of phonemes, indicating the acoustic effects of the word. With an increased understanding of the relationships between syntactic structure of a sentence and the sound generated, this stage of the preprocessing will prove invaluable in improving the timing of events in synthesis by rule. At the present time, however, there are no rules of this type in use.

The preprocessed sequence serves as the input for two separate but interdependent models for control of the segmental and suprasegmental features of the utterance, as seen in Fig. 3. The prime purposes of the segmental model are to generate the formant data and to control the timing of events during the utterances. The suprasegmental model is responsible for generating fundamental frequency control data. Since the mechanisms which account for these data in the human appear to be independent, at least to a first approximation, there are separate models for each in the synthesis by rule strategy.

The interconnections between the models reflect, to an extent, the interactions between source and system in speech production. The suprasegmental model relies on the *timing* information of the segmental model to raise and lower fundamental frequency for voiceless and voiced consonants and stressed vowels. The segmental model relies on durational information of the suprasegmental model to modify timing of events for stressed vowels, pauses, and sentence modifiers and punctuation as introduced by the preprocessor. It also obtains information from the suprasegmental model on vowel reduction and uses it appropriately. A form of source–system interaction observed in real speech, but not used in this model, is modification of the source waveshape depending on the

value of the first formant. In the synthesizer, there is no dynamic control over the source, only the pulse rate.

Details of the Segmental Model

The segmental model uses primarily the segmental phonemes and consists of an algorithm for generating formant contours. A primary goal is to bridge the discrete phoneme sequence by allowing the influence of each phoneme to be felt for as many phonemes as possible. In the current model the phoneme influence lasts for two phonemes.

The segmental model is basically as follows. Each segmental phoneme in the input string is characterized by a set of formant target positions and frequency regions surrounding the targets. The target positions contain quasi-static information about each phoneme, i.e., the steady-state formant values of that sound. For noncontinuant phonemes these targets represent virtual steady-state values. In these cases they serve more as an artifice of the model than as a physical entity. The frequency regions around the target contain information relevant to the dynamics of the formant transitions, and are used to a large extent as the basic mechanism for the control of timing. They are also used to account for such phenomena as vowel reduction, coarticulation, and hysteresis of formant contours.

Each formant contour is described by the solution to a second-degree differential equation of the form

$$\frac{d^2 f_N(t)}{dt^2} + \frac{2}{\tau_{AB}{}^N} \frac{df_N(t)}{dt} + \frac{1}{(\tau_{AB}{}^N)^2} f_N(t) = \frac{P_N(t)}{(\tau_{AB}{}^N)^2}$$

where N is the formant number; ($N=1$, 2, or 3); $P_N(t)$ is a series of step functions describing the Nth formant target values for the string of phonemes; $f_N(t)$ is the response to $P_N(t)$ and corresponds to the contour for the Nth formant; and $\tau_{AB}{}^N$ is the time constant of motion between phonemes A and B for the Nth formant.

The formant contour, in response to a single-step input beginning at $t=0$, is a smooth, continuous exponential of the form

$$f_N(t) = A_f + (A_i - A_f)(1 + t/\tau) \exp(-t/\tau);\ t \geq 0$$

for

$$P_N(t) = \begin{cases} A_i & t < 0 \\ A_f & t \geq 0, \end{cases}$$

and where τ is the appropriate time constant.

In general, motion between target positions does not proceed from a steady state; that is, the formant velocity is nonzero at the time the target changes. Motion to a target whose formant value is A_f from an initial position A_i with an initial formant velocity

$$V_i = \left.\frac{df_N}{dt}\right|_{0-}$$

is of the form

$$f_N(t) = A_f + (A_i - A_f)\exp(-t/\tau) + \left[V_i + \frac{(A_i - A_f)}{\tau}\right] t \exp(-t/\tau);\quad t \geq 0.$$

At times when the input to the differential equation is changed discretely, both the output value *and* slope are continuous. Thus the concept of smooth, continuous formant transitions is realized in all cases.

There are several reasons for using critically damped exponential solutions to describe the formant contours. They provide excellent fits to observed data on formant transitions from real speech. There is experimental evidence that exponential extrapolations of formant data provide better predictions during formant transitions into and out of consonants than linear extrapolation, or simply holding the formant fixed [8]. Investigators have also found that replacing linear transitions by nonlinear transitions have produced improvements in the speech.

The technique for the control of timing of phoneme changes is the most crucial step in the segmental model, and remains the most difficult aspect of synthesis by rule. It is unreasonable to specify for each phoneme a fixed duration, as this cannot possibly account for vowel stress, vowel reduction, and consonant lengthening and shortening as observed in speech.

The technique for control of timing depends on defining a metric for the approximation of target positions of a phoneme. This metric is in terms of the frequency regions of the phoneme. In order to initiate motion to a new phoneme, in general, all formants must be within the frequency regions of the targets. If the current phoneme is not a stressed vowel, as soon as the condition on the formants is met, a transition to the next phoneme is initiated. If a stressed vowel is being generated, the supra-segmental scheme is called upon to determine the amount of vowel lengthening.

This technique is capable of accounting for contextual differences in duration, vowel reduction, coarticulation, and hysteresis of formant contours. The frequency regions for a given phoneme can be modified by the preprocessor, thus modifying timing, and accounting for much of the versatility of these regions.

To illustrate the generation of formant contours, a simplified example (for a two-formant system) is shown in Fig. 4. The frequency regions are indicated by Δ's and the initiation times of phoneme changes by $t1$, $t2$, and $t3$. The transitions to new phonemes begin only when both formants are within their respective frequency regions. Note also that the time constants of the formant transitions are specified independently of each other. Fig. 5(A)

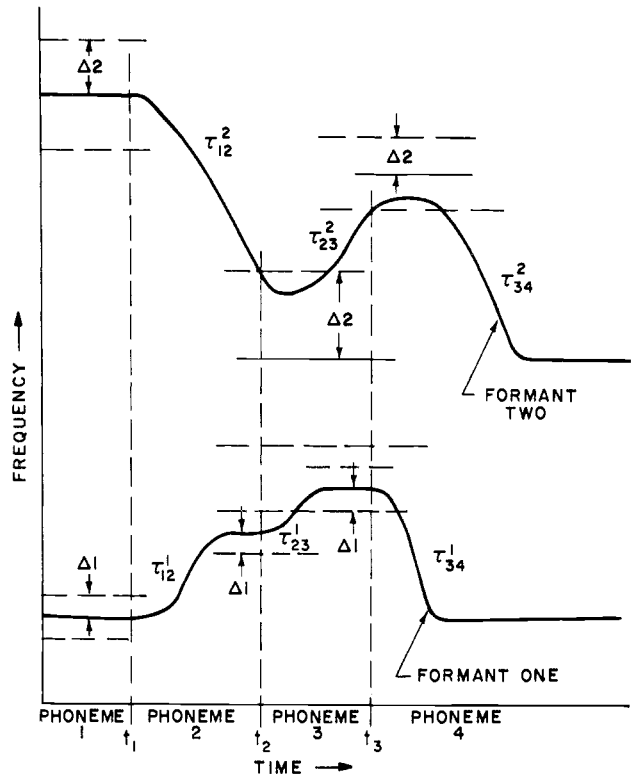
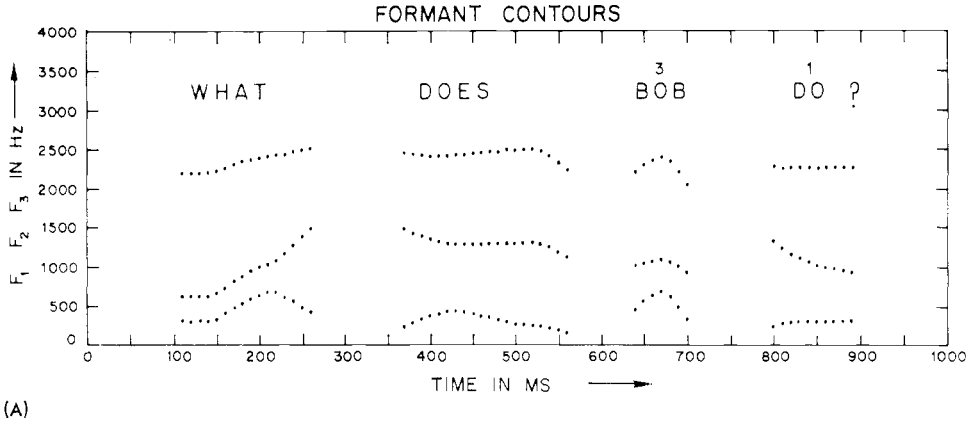


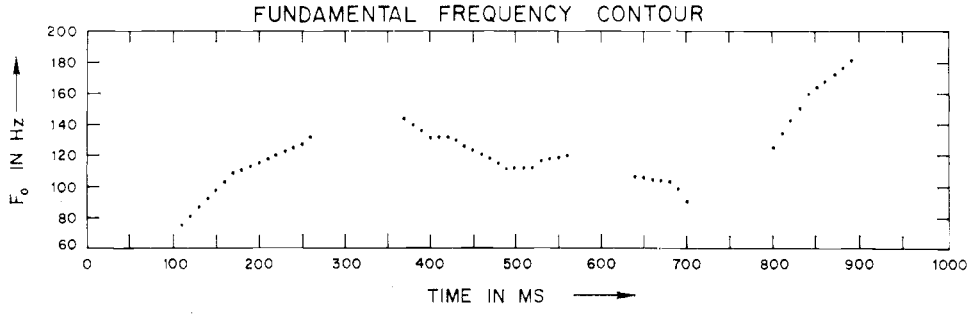
Fig. 4. Example of formant contours.

Fig. 5. Formant and fundamental frequency contours for synthetic utterance.



(A)

(B)



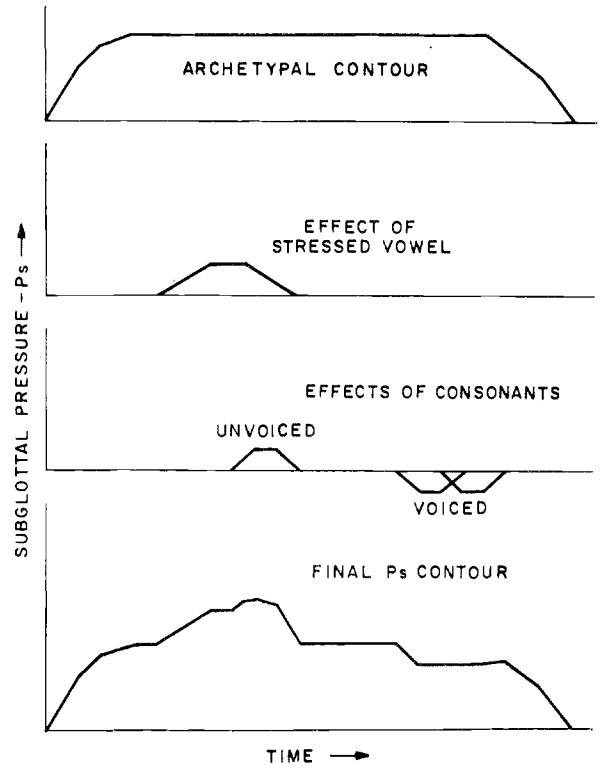


Fig. 6. P_s contour composition.

shows the formant contours generated by this technique for a synthetic utterance.

The additional control signals generated by the segmental model are highly dependent on the formant contours. A discussion as to how they are generated, given the formant contours, is available elsewhere [5].

Details of the Suprasegmental Model

The suprasegmental model is a physiological model using stress level data as input, and generating a fundamental frequency (F_o) contour as output [9]. The timing information from the segmental model is also used in generating the F_o contour. The variables of the model are subglottal pressure (P_s) and laryngeal tension (LT). Except for yes–no questions, LT is held constant for the utterance. In these cases fundamental frequency is linearly proportional to P_s.

The P_s contour is composed of three principle components: an archetypal contour for an entire utterance, a local perturbation due to unvoiced or voiced consonants, and a perturbation due to vowel stress. Fig. 6 shows a typical contour and its composition from the component parts. The height of the F_o perturbation due to vowel stress is logarithmically proportional to the stress level, attaining a maximum of 32 Hz for a vowel marked stress 1. (Stress also lengthens the vowel depending on both context and the stress level.) The height of the F_o perturbation for consonants is ± 20 Hz, depending on whether they are voiced or voiceless. The parameters of this model were, for the most part, determined experimentally. Fig. 5(B) shows the F_o contour for the synthetic utterance of Fig. 5(A).

Evaluation of Model

An experimental evaluation of the segmental model was undertaken and produced average intelligibility scores of about 80 percent in vowel-consonant-vowel tests, and about 90 percent for simple declarative sentences. With more difficult test material (test sentences from Beranek's [10] list) the average scores were about 80 percent, but these were far more variable across subjects than in the other tests.

The suprasegmental model has been tested against similar models having different parameter values, as well as against an earlier model [11]. It has been found to be favored in most cases over the other versions. Comparisons of F_o contours generated by the model, with those of real speech have shown a good deal of similarity.

Conclusions

A technique for speech synthesis by rule has been discussed and one possible implementation has been shown. The quality of the output is far from perfect indicating the

need for further research. This section will discuss those areas in which more research and understanding of the basic processes of speech are necessary.

The major unsolved problem concerns the timing of events in the segmental data. Although in many cases the speech was intelligible, the timing resulted in highly unnatural speech. The implementation of the rules of the preprocessor may help a good deal here as these rules reflect the properties of quanta of speech much larger than a phoneme. In order to introduce rhythm into a sentence, some account of the syntactic structure, as well as the stress pattern of the vowels, must be made. Other modifications in the suprasegmental rules for controlling vowel duration, and in the timing rules of the segmental model may be necessary.

Further studies of the source for voiced sounds are well worthwhile. An analysis of F_0 contours from real speech may reveal what is perceptually significant in these contours, and will provide guidelines as to how closely these need to be modeled in synthetic speech.

Voiced synthetic speech has been termed very buzzy. Studies as to the origin of this buzziness will also be of value. Perhaps it is due to the lack of precise detail in both the F_0 contour and the formant contours. The use of synthetic source functions may also be a cause.

Although the synthetic speech produced by rule is far from ideal, it is still sufficiently good to justify further investigation. The applications foreseen for synthesis by rule justify the amount of work necessary to uncover the secrets of speech.

REFERENCES

[1] J. N. Holmes, "Notes on synthesis work," Speech Transmission Laboratory, Stockholm, Quart. Prog. Rept., 1961.
[2] W. J. Strong, "Machine-aided formant determination for speech synthesis," *J. Acoust. Soc. Am.*, vol. 41, pp. 1434–1442, 1967.
[3] H. Hollien and R. W. Wendahl, "Perceptual study of vocal fry," *J. Acoust. Soc. Am.*, vol. 43, pp. 506–509, 1968.
[4] J. L. Flanagan and D. I. S. Meinhart, "Source-system interaction in the vocal tract," *J. Acoust. Soc. Am.*, vol. 36, pp. 2001(a), 1964.
[5] L. R. Rabiner, "Speech synthesis by rule: an acoustic domain approach," *Bell Sys. Tech. J.*, vol. 47, pp. 17–37, 1968.
[6] ——, "Digital formant synthesizer for speech synthesis studies," *J. Acoust. Soc. Am.*, vol. 43, pp. 822–828, 1968.
[7] J. L. Flanagan, *Speech Analysis, Synthesis and Perception.* New York: Academic Press, 1965.
[8] J. M. Pickett and D. C. Coulter, "Statistics of F2 adjacent to consonants and prediction of F2 onsets," *J. Acoust. Soc. Am.*, vol. 39, pp. 940–960, 1966.
[9] P. Lieberman, *Intonation, Perception and Language.* Cambridge, Mass.: M.I.T. Press, 1967.
[10] L. L. Beranek, *Acoustic Measurements.* New York: Wiley, 1965.
[11] L. R. Rabiner, H. Levitt, and A. E. Rosenberg, "Investigation of stress patterns for speech synthesis by rule," *J. Acoust. Soc. Am.*, vol. 45, pp. 92–101, 1969.

Lawrence R. Rabiner (S'62–M'67) was born in Brooklyn, N. Y., on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in June, 1964, and the Ph.D. degree in June, 1967 in electrical engineering, all from the Massachusetts Institute of Technology, Cambridge.

From 1962 through 1964, he participated in the cooperative plan in electrical engineering at Bell Telephone Laboratories, Inc., in Whippany and Murray Hill, N. J. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques at Bell Telephone Laboratories, Murray Hill.

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the Acoustical Society of America.