METHODS & DESIGNS

Some experiments with computer synthesized speech

PETER B. DENES, *BELL TELEPHONE LABORATORIES, INC., Murray Hill, New Jersey 07974*

The generation of satisfactory artificial speech involves a great variety of semantic, linguistic, and articulatory factors. Many of these complex elements are as yet unknown, and computers are used extensively for their investigation. An account of an on-line, interactive computer facility used for such research at Bell Laboratories is given. The use of this facility for exploring the duration and pitch correlates of stress in English sentences is described.

THE SPEECH SYNTHESIS PROBLEM

Research on generating artificial speech has been going on more or less continuously for several centuries (Dudley & Tarnoczy, 1950). More recently, a significant part of this effort has been devoted to "speech synthesis by rule," the process that converts sentences specified on a typewriter keyboard into readily understandable speech. Synthesizers of this kind are constructed to serve as models of the speech mechanism and are used for the study of the human speech process. They could also have practical applications, for example, for automatic voice read-out from computer-based information-retrieval systems and from teaching machines, and

for other ways of improving man-computer interaction. Unfortunately, nobody has yet been able to design a satisfactory method of speech synthesis by rule. This is not due to lack of effort, but is more an indication of some unexpected complexities of the speech process.

The basic speech-generating process is deceptively simple: Natural speech is produced by the vibration of our vocal cords, and this buzzy sound is transmitted through the tube formed by our throat and mouth. This tube, the vocal tract, has a number of resonances, called the "formants," and the frequencies of these acoustic resonances depend on the shape of the vocal tract. The quality of the vocal-cord buzz is modified by the formants when the sound is transmitted through the vocal tract. As speech sound after speech sound is pronounced, the movements of tongue and lips change the shape of the vocal tract; this will alter the frequency of the formants and thereby the quality of the sounds produced. Intensive research in the 1940s and 1950s resulted in satisfactory explanations of what determines the acoustic resonances of nonuniform tubes of shapes like those of the vocal tract (Fant, 1960; Flanagan, 1965). The understanding of the relevant acoustics, however, is insufficient for an explanation of the speech process as a whole. Experiments carried out at Haskins

Laboratories in New York (Liberman et al, 1952, 1967), by Ohman (1966) and Lindblom (1963), and by others indicate that the acoustic features typical for a phoneme when pronounced in isolation are modified extensively when this same phoneme is pronounced in a connected sentence. The changes are complex and are a function of a variety of contextual circumstances. Essentially similar differences are observed between words pronounced in isolation and in sentence context. The spectrogram of the sentence shown in Fig. 1 provides a number of examples of this. The syllable, "comp," occurs twice in this sentence, the first time in a stressed position and the second time unstressed. The difference in the corresponding acoustic features is obviously very large. Again, the phoneme /*l*/ occurs in a word initial position and again in a final position and shows significant acoustic differences; similarly for the phoneme /*p*/ when followed by different vowels.

Broadly speaking, there are two classes of contextual influences that are significant. One is a result of such linguistic factors as stress, inflection, grammar, and certain relations between speech sounds that vary from language to language. The other class is the result of articulatory context: Our vocal organs are constituted in specific physical ways, their muscular

Fig. 1. Sound spectrograms plot time along the horizontal axis and formant frequency along the vertical axis and indicate sound intensity by the darkness of lines. The broad, dark, horizontal bars indicate the formants. This spectrogram of the sentence, "It's a compact little computer," illustrates how phonemes vary according to context. Note the differences between the two "l" sounds in "little" and between the "comp" sounds in "compact" and "computer." These differences are influenced by many factors, which include sentence stress and inflection.

control for continuous action is coordinated at some high physiological level, their musculature allows only certain movements, and the acceleration and velocity of their masses obeys the usual laws of dynamics.

FORMANT SYNTHESIS

For many years, our principal method for investigating natural speech was by examination of speech spectrograms. All effects, acoustic as well as articulatory and linguistic, were therefore evaluated mainly in terms of formant specifications. Naturally enough, most synthesis schemes are also controlled in terms of formants, and the corresponding devices are called "formant synthesizers." They simulate the vocal cords by pulses that repeat at the rate of the vocal-cord vibrations, and the frequency of these pulses determines the pitch of the synthesized speech sounds. The pulses are applied to a number of resonators, each representing one of the formants of the vocal tract. The vocal-cord generator and the resonators can be simulated by computation or realized by conventional electronic circuits. In either case, the frequency and amplitude of the vocal-cord pulses and the frequencies and bandwidths of the resonators have to be readjusted at frequent intervals, typically once every 1/100 sec, in order to produce the varying sound qualities appropriate for a spoken sentence.

A variety of computer programs have been tried for converting a sentence, specified as a sequence of phonemes or letters, into appropriate values of frequencies and amplitude for controlling formant synthesizers.

Sentences synthesized by such rules usually have less than acceptable quality. At first, it was thought that this was caused by inadequacies of formant-type synthesis. John Holmes (Holmes et al, 1964), at the Joint Speech Research Unit in England, tested this assumption by a fundamental experiment. He synthesized the same sentence in two ways. First, he controlled a synthesizer by copying, as closely as he could, the spectrogram of a human utterance of the sentence. The quality of the synthesized utterance was almost the same as that of the original human utterance. He then generated the same sentence by rule, using the same formant synthesizer and by applying the most sophisticated rules available. The resulting speech quality left much to be desired. This experiment established the fact that the inadequacies of synthesis by rule are due not to deficiencies in the synthesizing circuits but are caused by less than satisfactory understanding of the rules for combining successive phonemes into a continuous sentence.

Many different sentence-synthesis rules have been tried during the last few years. Certain of these rules are concerned with transitions of parameter values from one phoneme to the other: Some compute simple linear transitions, others compute two-part linear transitions (Holmes et al, 1964), and others, again, use exponential transitions (Rabiner, 1968). These rules almost never take into account the influence of more than three successive phonemes and are, therefore, essentially concerned with short-term dependencies. Other synthesis-by-rule programs also accept stress and punctuation marks as part of the typed input (Rabiner, 1968; Mattingly, 1966). These additional symbols are used by the program to control the formant values and duration of individual elements and the pitch contours throughout the sentence. These grammatical factors effect successive elements over entire sentences rather than just two or three phonemes. Such a synthesis scheme is described in the latter part of this paper. There has even been a relatively successful effort to calculate the appropriate duration and pitch of the sounds from a grammatical analysis of the input sentence (Umeda, Matsui, Suzuki, & Omura, 1968). Whatever the method, as will be seen from the experiments described below, the factors to be considered greatly influence each other: Only by computer techniques can we hope to implement these complex long-term dependencies with reasonable ease.

TUBE ANALOGS OF THE VOCAL TRACT

A substantially different method for generating artificial speech simulates the action of our speech-producing mechanism by modeling the vocal tract directly as a nonuniform tube. Cross-sectional areas along the length of this tube are adjusted to correspond with variations of these areas between vocal cord and lips in the real vocal tract. The tube is simulated as a sequence of short, uniform cylindrical tubes where the cross-sectional area of each section is adjusted to represent the area of vocal tract at different points along its length. In earlier tube analogs (Dunn, 1950), the cylindrical sections were simulated by electrical networks. Others (Kelly & Lochbaum, 1962; Mermelstein, 1969) calculated the progression of the sound pulses from section to section all the way from vocal cords to lips, taking into account reflections and transmissions at the area discontinuities at section boundaries. So far, only computer models of this kind are available and they require anywhere from 30 to 100 times real time for their calculations. Efforts are being

made to construct special digital processors to obtain faster operation.

The best way for determining the variations in time of tube shape that are appropriate for particular utterances is as yet not well understood. One reason for this is that relatively little is known in detail about the activity of the human vocal tract during speech. Available x-ray and other methods for observation are not fully adequate; methods for measuring the vocal-tract shape from its acoustic output are only just emerging (Mermelstein, 1967; Atal, in press).

A powerful way for investigating appropriate vocal-tract shapes for given utterances is to model the individual movements of tongue and lips; several of such schemes have been tried (Henke, 1966). Another interesting attempt to achieve such articulatory synthesis is a computer program (Fujimura & Coker, 1966; Coker, 1968), in which the body of the tongue was modeled by a circle, the tip of the tongue by an attached wedge, and the lips by a short cylinder whose length and area were variable. Movement of these articulators during speech was defined by specifying five parameters: the X and Y coordinates of the center of the circle, the height of the tongue tip, the length and area of the lips. These five parameters were adjusted by rule, and the resulting vocal tract shapes were computed. Although the rules and constraints utilized with this model were not sufficiently appropriate to produce utterances that sounded better than the output of some formant synthesis by rule schemes, this represents a significant move towards using articulatory models for achieving more satisfactory synthesis.

THE USE OF COMPUTERS FOR SPEECH SYNTHESIS

The foregoing discussion will have shown that speech synthesis is a complex process that requires the use of computers. Computers have, in fact, been used extensively for both formant and tube analog type of speech synthesis over the last 10 years (Schroeder, 1969). Human evaluation by listening is an essential part of the experimental procedure, and immediate modification of the synthesis rules by the listener is often necessary to make rapid progress. Inexpensive on-line machines that are affordable for hours or days of experiments are needed for these studies. They must have efficient input and output facilities for providing effective acoustic and visual displays and for convenient program control by the listener; they must also have high arithmetic and processing capacity—far greater than that available on time-shared consoles.

Such a facility was assembled at Bell

Fig. 2. Schematic diagram of computer-controlled formant speech synthesizer.

Laboratories in the mid-1960s, around a medium-sized (24-bit, 12,000-word, 5-microsecond speed) computer (Denes, 1965). The work load on this machine was so great that in 1967 it was supplemented by a larger and faster machine (16,000 words, 2 microseconds). The two machines are used in a similar fashion, for similar problems. They are essentially program-compatible and the arrangement of their peripheral devices is similar.

Input equipment includes a typewriter, a Rand Tablet, and an analog-to-digital converter that is shared by a number of control knobs and a microphone; output equipment utilizes a variety of digital-to-analog converters to control a cathode-ray tube, analog sound-generating equipment, and a loudspeaker.

Both machines have disks that allow quick and convenient storage of data and programs. The disk is also essential for making the simple, but versatile, operating system really convenient to use. The operating system provides only a simple symbolic assembler, but it does have a standard Fortran IV compiler. There are simple Fortran callable subroutines for controlling the numerous peripheral devices that provide visual and acoustic input and output. This Fortran has the additional advantage of being source-program compatible with the Fortran of a large batch processing machine at Bell Laboratories. Programs can therefore be debugged with the aid of the diagnostic capabilities of the large computer's Fortran and finally recompiled and loaded on the on-line computer.

The on-line machines have standard programs for tube analog synthesis, for the control of an electronic formant synthesizer, and, alternatively, for computer simulation of the formant synthesizer.

A typical arrangement for using the electronic formant synthesizer is shown in the block diagram of Fig. 2. The computer controls the synthesizer (Coker & Cummiskey, 1965) by outputting a set of 12 numbers every 10 msec, which set the 12 control parameters of the formant synthesizer (pitch, amplitudes, formant frequencies, and bandwidths, etc.). The E not only hears the synthesized speech in real time but he can also see any or all control parameters displayed as a function of time on the screen of the computer-controlled cathode-ray tube. Programs are available for controlling each of the 12 parameters in one of two ways:

(1) The E can specify parameters as a

function of time on the typewriter or by drawing the appropriate curve on the cathode-ray-tube screen; he can instantly listen to the corresponding change in the synthesized speech. This program is used extensively by phoneticians with no programming knowledge whatsoever.

(2) A sentence to be synthesized is typed into the computer, phoneme by phoneme. In addition to the phonetic symbols (see Table 1), a special symbol has to be typed to mark the stressed syllables, and a period or question mark has to be typed at the end of the sentence to indicate whether the sentence is a statement or a question. A computer program then calculates the variations in time of each parameter by using values stored for each phoneme and by applying programmed algorithms for determining the phoneme-to-phoneme transitions of these parameters. The computed parameter curves are outputted to the synthesizer and to the display scope. The program also allows modification in real time of the computed parameter values or durations

either by using the Rand Tablet and the display oscilloscope or by typewritten instructions.

This synthesis-by-rule program will now be described in somewhat more detail because it was the program that was used to prepare the test material for the stress and inflection experiments to be discussed in the last part of this paper.

In the synthesis-by-rule program, each phoneme was typically specified by a central steady-state segment, an initial transition, and a final transition. Stored values were available for the steady-state segments of each of the phonemes for each of the 12 parameter values and also durations for each of the three segments and a "weight" for the transitions. Phoneme-to-phoneme transitions were computed as shown in the examples of Fig. 3. The three examples show the influence of the relative "weights" of adjacent phonemes.

The stress marks automatically doubled the stored duration of the steady-state segment of the stressed vowels and halved

Table 1
List of Phonemes Accepted by Synthesis Program*

Phonetic Symbol Used	Key Word	Phonetic Symbol Used	Key Word	Phonetic Symbol Used	Key Word
P	Pan	M	Moon	EE	Heat
T	Tan	N	Noon	I	Hit
K	Can	NG	Sing	E	Head
B	Ban	L	Limb	AE	Had
D	Dan	R	Rim	AH	Hot
G	Gun	Y	Yes	AW	Ball
F	Fan	W	Win	U	Put
TH	Thin	DH	Then	OO	Boot
S	Sin	Z	Zebra	UH	But
SH	Shin	ZH	Garage	ER	Sir
H	He	V	Victor		

* Diphthongs were considered sequences of two vowels.

Fig. 3. Examples of transitions between phonemes for three different "weights."

the stored duration of the unstressed vowels.

The pitch was originally computed on a monotone. The pitch of each stressed syllable was made to decrease; the start of this pitch fall was above the monotone level.

The pitch change of the final stressed syllable in each sentence decreased or increased for statement or question, respectively. The intelligibility of the sentences generated by this program varied greatly for different sentences.

DURATION AND PITCH CORRELATES OF STRESS

Most people would consider that the principal acoustic feature of stress was the increased amplitude of the stressed syllable. In actual fact, the amplitude does not seem to affect our stress perception greatly. There is evidence, however, that the pitch and duration of sounds do have a strong influence on stress perception, but not enough is known about the correct values of these acoustic cues to perform satisfactory synthesis by rule (Fry, 1958).

The synthesis program just described was used to generate speech sequences with a variety of duration and pitch values to evaluate quantitatively their effect on stress perception (Oster, 1968).

The sentence, "They are flying planes," was chosen for one of the tests. As is well known, this is a grammatically ambiguous sentence. In one interpretation, "are flying" is the verb and "planes," the

object; in another way, "are" is the verb and "flying planes," the noun phrase. The first version's 2413 grammatical stress pattern (1 is the greatest stress) becomes a 3521 if semantic stress is added to "planes"; the second interpretation's 2341 stress pattern becomes 3412 if semantic stress is added to "flying." The clear difference in meaning due to the stress oppositions of the last two words of the sentences with semantic stress was explained carefully to the Ss. This insured that their responses were in terms of the differences in meaning caused by the position of stress in the sentence and avoided the need to explain the concept of "stress" to the Ss.

The duration and pitch of the steady-state elements of the leading vowels of the diphthongs in "flying" and "planes" were varied systematically. The pitch of the sentence was kept constant at 130 Hz, except as required in the test vowels. The results are shown in Figs. 4 and 5. They indicate that increased duration and increased pitch of a vowel increase the likelihood of hearing the syllable that includes that vowel as stressed. However, stress perception was influenced not just by the duration and pitch of the vowel concerned but also by the duration or pitch of other vowels in the sentence. The results also show a trade-off between duration and pitch: If stress is perceived on Syllable A because of its higher pitch level, then stress can be shifted to Syllable B by changing either the duration or the pitch of its vowel. Another important effect is that "flying" required considerably less pitch or

duration increase to be heard stressed than did "planes." This means that different syllables have different requirements of relative pitch or duration in order to be heard stressed.

This may have a number of explanations. It may simply be due to the different vowels in the two syllables or to their different phonetic environments. It could also have been due to the sentence-final position of "planes," or due to various semantic influences. Whatever the reason, it shows an important effect that was also observed in other synthesized sentences: Equal duration or equal pitch of all the vowels of a sentence do not produce equal stress; under these conditions of equality, the stress falls on the syllable that, for grammatical or semantic reasons, should have the least stress. Yet another factor is that the definition of vowel duration must necessarily be arbitrary; the particular ad hoc definition adopted could lead to difficulties when the same vowel is used in different phonetic environments.

In summary, the strong influence of vowel duration and pitch on stress perception can be demonstrated easily. The quantitative relationship, however, is variable and affected strongly by a great variety of phonetic, grammatical, and semantic conditions.

The outlook for good-quality computer-generated speech is not quite as black as these results might imply. Phonetic, grammatical, and semantic constraints can be restricted by allowing only sentences made up from a small number of words and with a restricted

Fig. 4. Stress judgments as a function of vowel duration.

Fig. 5. Trade-off between duration and pitch of vowels for stress judgements. The symbols used in Fig. 4 also apply to this figure.

grammar. The resulting ensemble of sentences could still be significant for such applications as voice read-out of computer-stored information and at the same time exhibit more manageable long-range effects of stress on the duration, pitch, and formants of speech sounds throughout the sentence. This is particularly so today because inexpensive digital disks are now becoming available that allow the storage of a considerable number of prerecorded words; such words or phrases can then be assembled into acceptable sequences by simpler contextual rules. Work on such schemes is in progress now, and computers are indispensable to handle the exploration of the remaining complexities, which are still formidable. The lessons learned from such simpler synthesis schemes may well give useful cues for more ambitious schemes later.

REFERENCES

- ATAL, B. S. Determination of the vocal tract shape directly from the speech wave. *Journal of the Acoustical Society of America*, in press.
- COKER, C. H. Speech synthesis with a parametric articulatory model. Paper A-4, Speech Symposium, Kyoto, Japan, 1968.
- COKER, C. H., & CUMMISKEY, P. On-line computer control of a formant synthesizer. *Journal of the Acoustical Society of America*, 1965, 38, 940(A).
- DENES, P. B. On-line computing in speech research. *Proceedings of the 5th International Congress on Acoustics, Liege, 1965*.
- DUDLEY, H., & TARNOCZY, T. H. Speaking machine of Wolfgang von Kempelen. *Journal of the Acoustical Society of America*, 1950, 22, 151-166.
- DUNN, H. K. Calculation of vowel resonances and an electrical vocal tract. *Journal of the Acoustical Society of America*, 1950, 22, 740-755.
- FANT, C. G. M. *Acoustic theory of speech production*. The Hague: Mouton, 1960.
- FLANAGAN, J. L. *Speech analysis, synthesis and perception*. New York: Springer, 1965.
- FRY, D. B. Experiments in the perception of stress. *Language & Speech*, 1958, 1, 126-152.
- FUJIMURA, O., & COKER, C. H. Model for the specification of the vocal tract area function. *Journal of the Acoustical Society of America*, 1966, 40, 1271(A).
- HENKE, W. Dynamic articulatory model of speech production. *Journal of the Acoustical Society of America*, 1966, 40, 1271(A).
- HOLMES, J. N., MATTINGLY, I. G., & SHEARME, J. N. Speech synthesis by rule. *Language & Speech*, 1964, 7, 127-143.
- KELLY, J. L., & LOCHBAUM, C. Speech synthesis. *Proceedings of the Stockholm Speech Communications Seminar, RIT, Stockholm, Sweden, September 1962*.
- LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, 74, 431-461.
- LIBERMAN, A. M., DELATTRE, P., & COOPER, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, 65, 497-516.
- LINDBLOM, B. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 1963, 35, 1773-1781.
- MATTINGLY, I. G. Synthesis by rule of prosodic features. *Language & Speech*, 1966, 9, 1-13.
- MERMELSTEIN, P. Determination of the vocal-tract shape from measured formant frequencies. *Journal of the Acoustical Society of America*, 1967, 41, 1283-1294.
- MERMELSTEIN, P. Computer simulation of articulatory activity in speech production. In D. E. Walker and L. M. Norton (Eds.), *Proceedings of the Joint Conference on Artificial Intelligence*. Washington, D. C.: Gordon & Breach Co., 1969.
- OHMAN, S. E. G. Coarticulation of VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 1966, 39, 151-168.
- OSTER, R. The effect of syllable duration and pitch on stress perception. *Bell Laboratories*, 1968. Unpublished.
- RABINER, L. R. Speech synthesis by rule: An acoustic domain approach. *Bell System Technical Journal*, 1968, 47, 17-37.
- SCHROEDER, M. R. Computers in acoustics: Symbiosis of an old science and a new tool. *Journal of the Acoustical Society of America*, 1969, 45, 1077-1088.
- UMEDA, N., MATSUI, E., SUZUKI, T., & OMURA, H. Synthesis of fairy tales using an analog vocal tract. *Reports of the 6th International Congress on Acoustics, 1968, Vol. II, B 159-162*.