The Use of Spectrograms for Speech Analysis and Synthesis*

JOHN M. BORST

Haskins Laboratories, New York, N. Y.

The use of spectrograms and a pattern playback for research on speech has the unique advantage that it permits the study of isolated acoustic cues for speech perception. This method has shown that the consecutive sounds of the language are usually so intimately connected that they cannot be separated and recombined in a different order without serious loss in intelligibility. Formant transitions, which often characterize the consonants, are found to fall into well-defined patterns and groups. The knowledge gained by studies with synthetic speech has guided the construction of a speech synthesizer which copies the sounds generated by the human speech organs and which can be operated from the information found in spectrograms.

INTRODUCTION

A STUDY of the perception of speech poses a number of questions: How does one speech sound differ from another insofar as the ear is concerned? What serves to specify acoustically any given speech sound? What are the minimum requirements for speech to be intelligible? What are the minimum units or "building blocks" of speech, or are there any?

It is most helpful, in such an investigation, to have the speech in a permanent form so that one can examine its parts —take one part away or alter another and observe the effect. This is possible if the speech can be transformed into visual forms. So far, there are two different visual transformations available, the oscillogram and the spectrogram. Of these two, the spectrogram has proved to be the more useful. It is, as may be seen in Fig. 1, very difficult to recognize and classify the fine distinctions between similar sounds in an oscillogram, whereas in a spectrogram one can recognize distinctive patterns. The pattern for a given word is the same regardless of who spoke it, and its shape is as distinctive as the sound to which it corresponds.

One would, of course, like to know the significance of individual features—for instance, of the broad dark bands on the spectrogram: What would happen if the pattern contained only these bands and nothing else? Or what would happen if one of the bands were omitted or its curves altered? A direct answer is possible if one can modify or alter these patterns—or even make entirely artificial ones—and then turn them back into sound again to be evaluated by ear. For such studies, two translating devices are needed: one to make a picture from the speech—the spectrograph—and the other to turn a spectrogram into sound—the pattern playback.

The spectrograph is already well known. The machine which we have used was made at our own laboratory. It differs from the Bell Laboratories spectrograph in construction, though not in principle of operation. Our instrument records the spectrogram on film rather than on paper, and thus permits a greater dynamic range (36 db) without compression. Also, we use a longer sample of speech (12 seconds).

THE PATTERN PLAYBACK

The other translating machine, the playback, is illustrated in Fig. 2. A light source provides an intense thin line of light which is focused radially on a rotating disc. This "tone wheel" is a film negative which carries 50 concentric variable-density sound tracks. The inner one has 4 sine waves on it, the next one 2 × 4 sine waves, the next one, 3 × 4 sine waves, etc. Thus, by rotating the disc at 1800 rpm (30 rps), we modulate the first 1/10th in. of the line at 120 cps, the next 1/10 in. at 240 cps, and the fiftieth segment at 6000 cps. The line becomes, in a sense, the optical equivalent of a piano keyboard, though each of its keys plays only a single sine wave.

The modulated line of light is focused on the moving spectrogram. If this is an original spectrogram, the transmission system is used, and the phototube is placed below it, so that whatever light passes through will be collected. For an artificial or hand-painted spectrogram, the reflection system is used, and the light collector is then placed above the spectrogram. Wherever there is any paint on the transparent tape, light will be diffusely reflected into the collector. This light is modulated at a frequency determined by its

* Presented at the Seventh Annual Convention of the Audio Engineering Society, New York, October 12–15, 1955.

Reprinted from the JOURNAL OF THE AUDIO ENGINEERING SOCIETY

1956, v4, 14-23

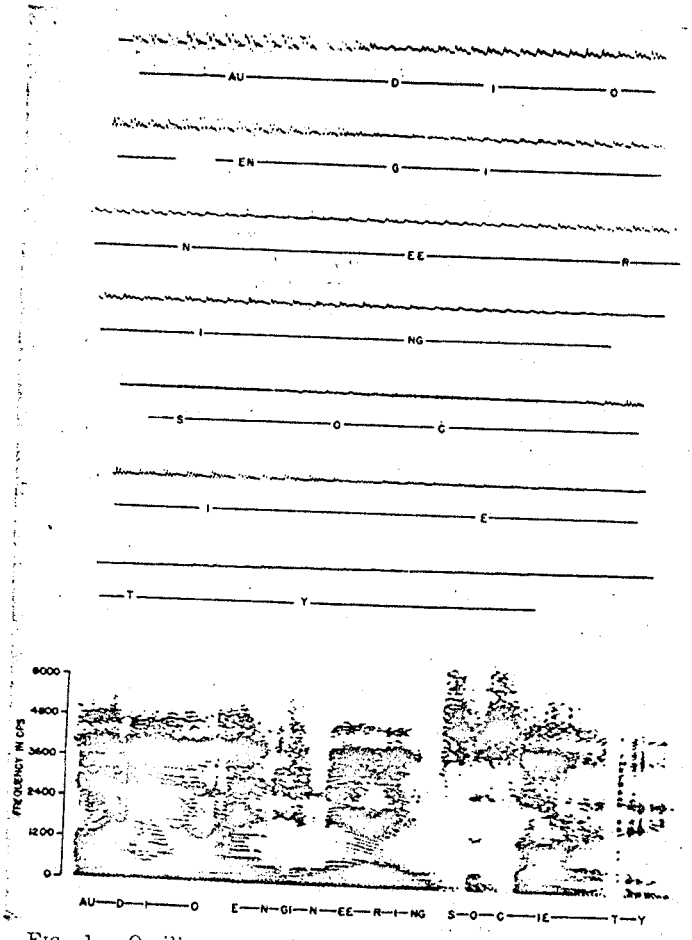

FIG. 1. Oscillogram and spectrogram of the spoken words "Audio Engineering Society." Note that the spectrogram gives a clearer "picture" of the speech.

position along the transverse, or frequency, axis of the spectrogram.

The machine is limited to a fixed pitch and one might expect that it could not make "hissy" sounds. However, it was found that noise can be generated by painting patches of fine dots, and reasonably satisfactory s-like sounds can be made.

A spectrogram of real speech will, when run through the machine, produce synthetic speech which is highly intelligible. So also will a hand-painted spectrogram (Fig. 3) derived from an actual one by copying only the main features. Some trial and error was necessary to find which features were important and how much the spectrographic patterns could be streamlined without impairing the intelligibility. Some of the dark bands in the upper part of the real spectrogram are a peculiarity of the particular speaker and are present whenever there is any voicing, but are not necessary for intelligibility. Not every speaker has these resonances, and some have them at other frequencies.

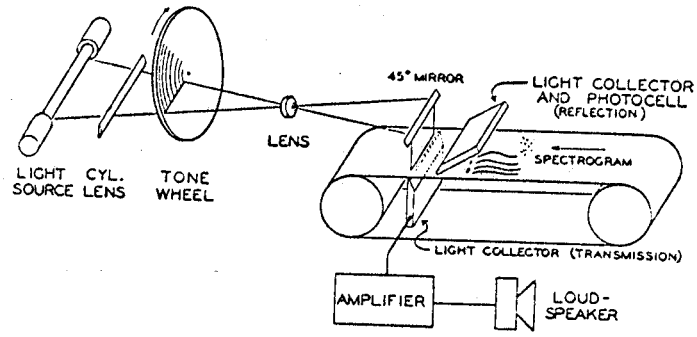

FIG. 2. Simplified diagram of the pattern playback. (Reproduced by courtesy of the *American Journal of Psychology*.)

With a spectrogram which has been reduced to its barest essentials, autosuggestion may play a part in determining what the listener will hear. If the experimenter knows what the text is supposed to be, he will very often hear it no matter how bad the sound may be. This effect was demonstrated with the spectrogram shown in Fig. 4. The lower line has been reduced to but a single "formant" and yet, when played through the machine, the sentence is still intelligible—to those who know the text!

FORMANTS

One may observe in Fig. 3, for example, that a hand-painted spectrogram consists mainly of broad lines running sinuously from left to right. These broad lines are called formants. They represent a concentration of energy within a narrow frequency band—i.e., a buzz sound passing through an acoustic filter. Or one may consider that a resonant cavity in the vocal tract is being shock-excited by a puff of air every time the vocal cords open and close. Each time this

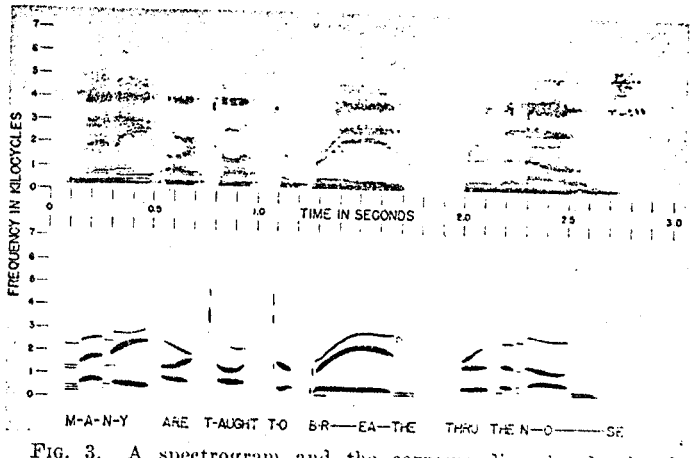

FIG. 3. A spectrogram and the corresponding hand-painted, simplified version.

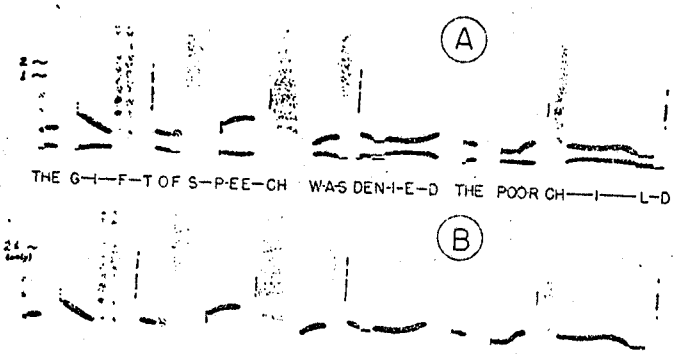

FIG. 4. Two versions of a painted spectrogram differing in the degree of simplification. The text of the first version (A) is readily understood; the second (B) is intelligible to those who know the text, but not to others.

happens, a damped wave results. We can imitate this behavior electrically and generate a single formant by applying a short pulse to a resonating circuit, and by then repeating the pulse a hundred times a second or so. This produces a formant like those we see in spectrograms.

The top line of Fig. 5 shows an oscillogram, a spectrogram, and a conventional frequency analysis—or section—of such a formant. The repetition rate of the damped wave trains represents the pitch. On the spectrogram, and in the section, the distance between the harmonics corresponds to the pitch. The frequency of the damped oscillation is the formant frequency. On the spectrogram it is the place (on the frequency scale) of greatest density, and on the section it is the peak of the envelope. This envelope, of course, is the resonance curve of the tuned circuit. It may be seen from the figure that there need not be a harmonic precisely at the peak; this would happen only if the formant frequency were an integral multiple of the pitch.

On the second line of Fig. 5 is another formant with the same formant frequency but with a higher *pitch*. The harmonics have moved farther apart but the resonance curve is still the same. On the bottom line is a third formant with a higher *formant frequency* but at the same pitch as in the top line. The spectrogram now has the narrower spacing between harmonics, but its center of density has moved to a higher frequency, and the peak of the resonance lies farther out on the frequency scale.

There is still a third independent variable, the rate of damping or the Q of the circuit, but it is of less importance.

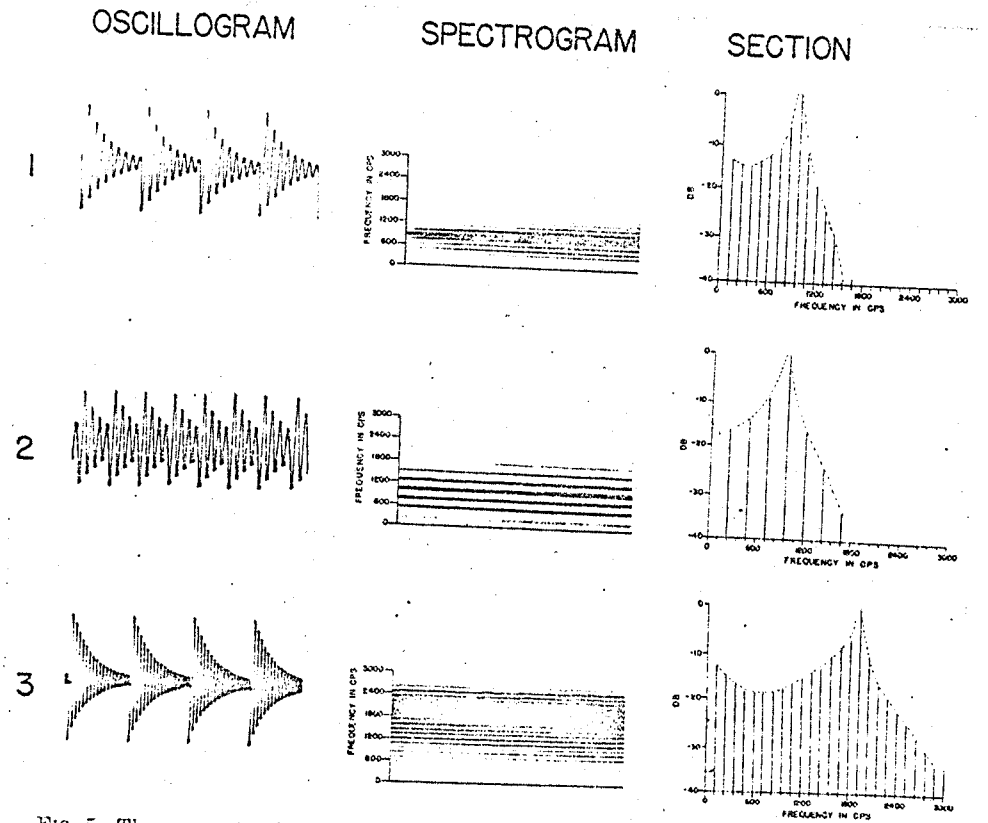

FIG. 5. Three aspects of a formant. The effects on the waveform, spectrogram and section of various combinations of pitch and formant frequency are shown for (1) low pitch and low formant frequency, (2) high pitch and a low formant frequency, and (3) low pitch and a high formant frequency.

The Q of vocal cavities is rather low—somewhere between 2 and 15. It increases with frequency, though not proportionally, and it varies with the individual.

A diagram of the circuit used to generate the formants described in the preceding paragraphs is shown in Fig. 6.

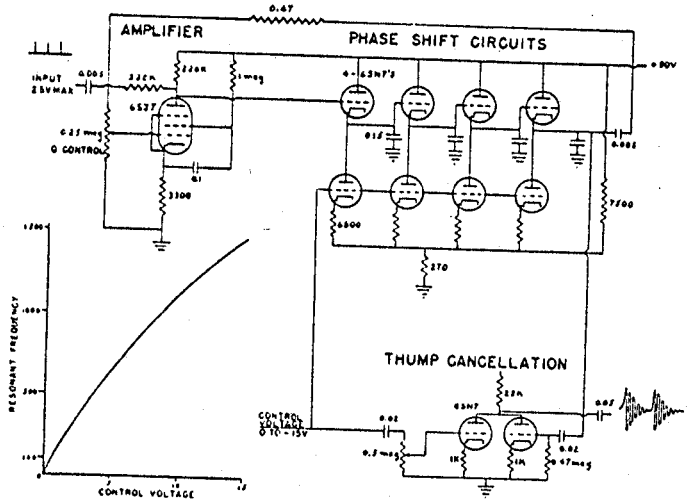

FIG. 6. Schematic of a formant generator consisting of a four-mesh phase-shift oscillator with insufficient gain to oscillate. Tubes replace the resistances in the phase-shifting mesh.

It consists essentially of a tuned circuit which can be adjusted electrically, i.e., by changing a voltage.[1] Since an inductance would become quite large at these low frequencies, an RC circuit is used instead. The circuit is a phase-shift oscillator with the gain set too low for oscillation. The resistances in the phase-shifting network are then replaced by tubes, so that the circuit can be tuned by adjusting the grid bias on the different tubes. The oscillator is excited by toothpick pulses—differentiated sawtooth waves, and the amplitude is controlled by a biased diode which passes more or less of the exciting pulses. By adding the signals from two or three such generators, one can make a vowel sound.

It has long been known that vowels are identified by the frequencies of the two or three lowest formants. These formants occur at virtually the same frequencies regardless of who speaks the vowels, although there are systematic variations among men, women, and children, just as there are very evident variations in average pitch: Thus, the formants for women's voices are about 10% higher than for men, although the pitch may be twice as high. (For a recent study of the formant frequencies of spoken vowels see an article by Peterson and Barney.[2])

CONSONANTS

A striking characteristic of the spectrographic patterns of Figs. 1, 3, and 4 is that there is almost no place where the formant frequencies remain the same very long. Either they glide, as slowly but constantly changing vowels, or they move abruptly, as *transitions* which are no longer perceived as vowels but rather as consonants. Indeed, much of the information in speech resides in these transitions, and the demonstration of their importance was one of the significant contributions of the synthetic method. Some consonants have, in addition to transitions, a burst of noise (examples *p, t, k*); or a noise portion of rather longer duration (example *s, sh*) or a weak steady-state formant (examples *l, r, m, n*). The transition which is present in addition to the noise may or may not be important to the listener. In the spectrogram of *ga* (Fig. 7), the noise burst and the transitions at the beginning of the formants, both play a part in the recognition of the consonant, but the part played by the transitions is by far the more significant. An

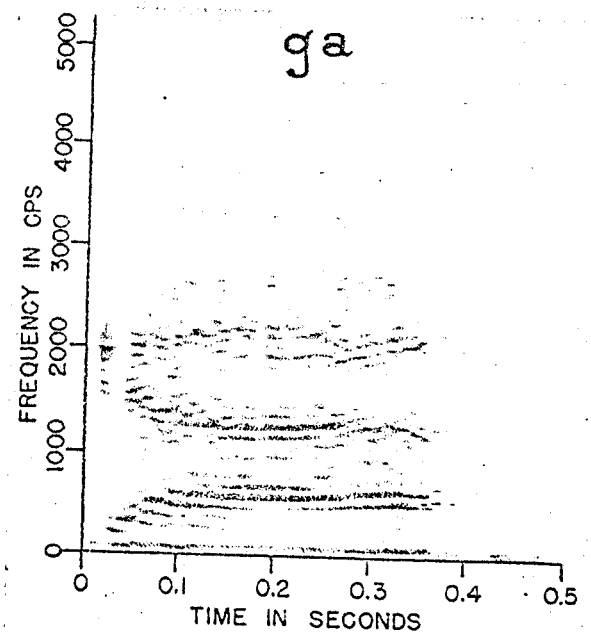

FIG. 7. A spectrogram of the spoken syllable *ga*. The listener's identification of the consonant is due to the burst of noise as well as to the transitions of the formants. (Reproduced by courtesy of *Psychological Monographs*.)

[1] Millard E. Ames, "A Wide-Range Deviable Oscillator," *Electronics*, 22, No. 5, 96–100 (May 1949).

[2] Gordon E. Peterson and Harold L. Barney, "Control Methods Used in the Study of the Vowel," *J. Acoustical Soc. Amer.*, 24, No. 2, 175–184 (March 1952).

important virtue of the hand-painted spectrograms is that their use permits the isolation of the effects of burst and transition simply by painting only one at a time.

Careful studies have been made of the effect of varying the frequency location of the burst, or the direction, extent, or duration of the various transitions. A discussion of these experiments is beyond the scope of this paper, but can be found in recent articles in the *Journal of the Acoustical Society of America*.[3,4]

The results of some of these experiments are given in Fig. 8, which shows the best transitions for b, d, g, with seven vowels. Note that the transitions are different for each consonant-vowel combination. It follows that if one were splicing a magnetic tape and were to cut off the *b* part of *bi* and try to transplant it onto an *a*, it would not fit, and we would not get the sound *ba*, but rather something quite unrecognizable. Similar difficulties would be encountered with most other consonant-vowel combinations: When one cuts the consonant away and adds it to another vowel, it is either unacceptable or it may even become another consonant. For instance, cutting the initial noise burst of *p* from *pi* and adding it to a vowel *a* could result in the sound *ka* rather than *pa*.[5,6]

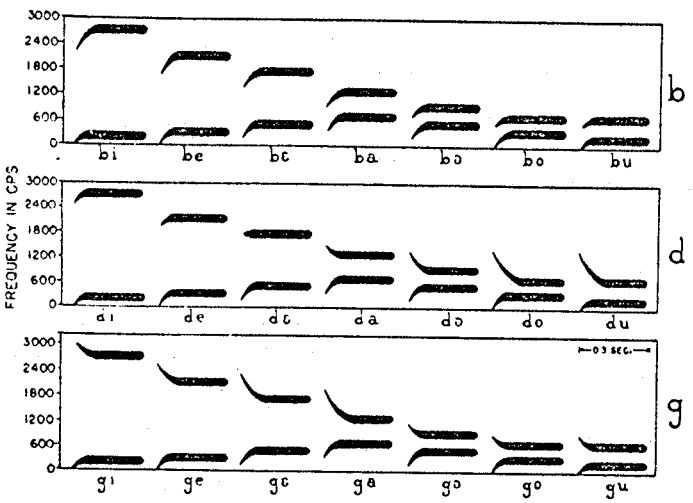

FIG. 8. Simplified spectrographic patterns for the consonants *b*, *d*, and *g* paired with each of several vowels. The transitions for a given consonant vary with the vowel which follows it. (Reproduced by courtesy of *Journal of the Acoustical Society of America*.)

From all this it appears that one cannot separate individual speech sounds and reunite them in a different order. This explains, for example, why a speech synthesizer depending on pre-recorded single sounds—to be combined à la typewriter—is likely to give unsatisfactory speech. Also, one would expect that a phoneme recognizer (speech typewriter) would work better if it utilized the shapes of transitions as well as spectral distributions of energy.

LOCUS

One of the rules which is most helpful in painting or synthesizing speech is the concept of the locus; this gives a simple way to account for the many different transitions shown in Fig. 8. Thus for a *d* preceding any vowel, we can see from Fig. 9 that the second formant transition appears to have started from a virtual point, or locus, at about 1800 cps; similarly, there is a locus for *b* and another for *g*.

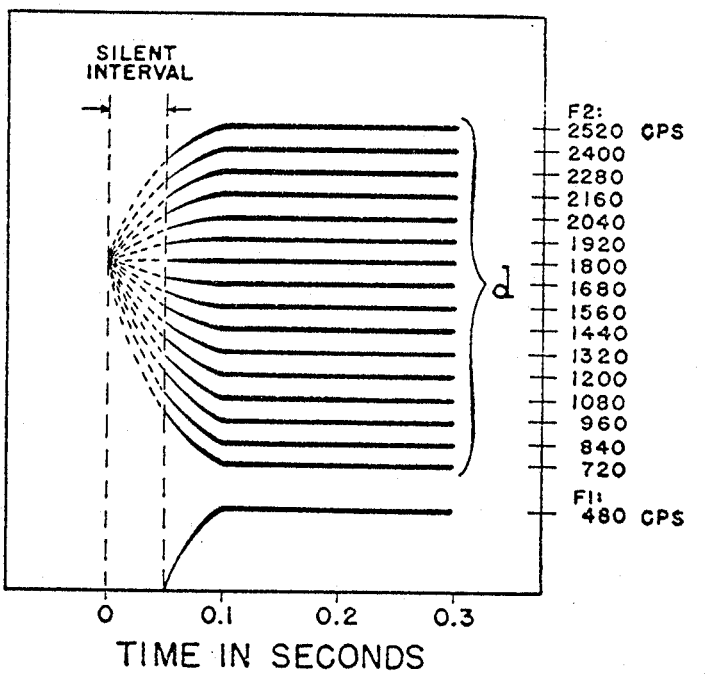

FIG. 9. Patterns which illustrate the locus principle. The second formant transitions of all vowels appear to originate in a virtual point (*i.e.* at one frequency for *d*, as is shown in the figure, at another for *b*, and at a third for *g*). (Reproduced by courtesy of the *Journal of the Acoustical Society of America*.)

[3] F. S. Cooper, P. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some Experiments on the Perception of Speech Sounds," *J. Acoust. Soc. Amer.*, 24, No. 6, 597-606 (November 1952).

[4] A. M. Liberman, P. Delattre, F. S. Cooper, and L. J. Gerstman, "The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants," *Psychological Monographs*, 68, No. 379, 1–13 (1954).

[5] A. M. Liberman, P. Delattre, and F. S. Cooper, "The Role of Selected Stimulus Variables in the Perception of the Unvoiced Stop Consonants," *Amer. J. Psychol.*, 65, 497–516 (October 1952).

[6] Carol D. Schatz, "The Role of Context in the Perception of Stops," *Language*, 30, 47–56 (1954).

Moreover, the locus for *d* is valid for other dental sounds, such as *t* and *n*, and is approximately correct for *s* and *z*. We can, indeed, build up a rather complete "periodic" table of speech sounds and their acoustic cues.[7] With this information, it is now possible to paint spectrograms by *rule*, that is, without reference to an actual spectrogram of the utterance. It is even possible to paint in dialects, as the recording of Alabama with a Southern or a French accent demonstrated.† The pattern playback can also produce non-speech sounds when one paints arbitrary figures instead of a spectrogram; likewise, by drawing short horizontal lines, comparable to slits in a piano-roll, one can make a kind of "music."

APPLICATION TO RECORDING AND TAPE EDITING

What are some of the practical uses of this knowledge? For one thing, it should help us to understand some of the odd things which happen in tape editing. Those who have had extensive experience report occasional changes in meaning when some portion of a word is removed or replaced. In practically all cases this might have been predicted. We have already seen that transitions cannot usually be transplanted from one vowel to another. For similar reasons, if a portion of a transition is cut off, the remainder may resemble the transition of another consonant, and it will sound like that consonant; or if a portion of a sibilant or fricative is cut off, one may also change the identity of the consonant. For instance, in one experiment, the word *sha* was recorded several times, and increasing portions of the initial noise patch were sacrificed by snipping away pieces of tape. When the progressively reduced versions were played, one noticed that the sound changed from *sha* to *tcha*, to *tia*, and to something between *ta* and *ka*.

But although one cannot transplant a single speech unit from one place to another, it is often possible to do so with a syllable and thereby to change completely the meaning of a sentence. As one example, the word "reason" was changed to "treason" by rerecording "tree" and then using it to replace the first syllable of "reason." In this way, the sentence "He's got me up a tree, there must be some reason in this" was changed to " . . . there must be some treason in this." Correct spacing and a good splice are required. Fig. 10 shows spectrograms of "tree," "reason," and both the faked "treason" and the spoken word "treason." The splice is almost undetectable, even on the original spectrogram. The spectrograms of Fig. 10 contain portions of the sounds preceding and following the words. This was done in order to show any changes which might have occurred between syllables.

The realistic recording of the fricative sounds *s*, *sh*, *f*, and *th* has long posed special problems, for reasons which are more or less evident from the frequency and intensity characteristics of these sounds. In general, they consist of a patch of noise together with a transition to the following vowel. The noise portion of *s* is practically all above 3600 cps, while that of *sh* comes down lower but not below about 2000 cps. The noise segments of these sounds are rather long and intense, and they play a major part in their recognition; the transitions, although often quite prominent in the spectrogram, seem not to be very important for recognition. The *f* and *th* sounds have noise that is weak, at frequencies that are not clearly specified; the recognition of these sounds depends less than for *s*, *sh* on the noise portions and more on the transitions—especially in distinguishing the two from one another.

Unfortunately there is not much new to be said on the subject of how to record and transmit these sounds. To be sure, one must not introduce noise at the wrong frequency due to intermodulation distortion; also relative intensities are important, so one should not boost one frequency region too much. Finally, since some of the noise components are weak, one should maintain a good signal-to-noise ratio. All of this is, of course, well known.

OCTOPUS

Another speech synthesizer with which we have worked was designed on the basis of information gained in the course of our research. This device does not translate spectrograms directly into sound, but rather the operator sets the controls with the aid of spectrographic patterns. Octopus (a name chosen for reasons which may become evident from the description) consists of three formant generators of the kind shown in Fig. 6, connected in parallel. Each generator can be excited by a pulse train or by white noise, or by both simultaneously. The exciting signals, supplied by the circuit of Fig. 11, are amplitude-modulated before they are applied to the formant generators, thus setting the relative loudness of the formants. The buzz generator can also be varied in frequency by a control voltage.

Since all of the variables are voltage controlled, the instrument is potentially able to produce speech if one supplies all the required voltages at the proper time. One might obtain these control voltages from a painted tape, from a recording, or from a suitable analyzer which derives them directly from

---

[7] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic Loci and Transitional Cues for Consonants," *J. Acoustical Soc. Amer.*, 27, No. 4, 769-773 (July 1955).

† A demonstration was included in Mr. Borst's presentation at the Sound Creation session of Audio Engineering Society's Seventh Annual Convention.

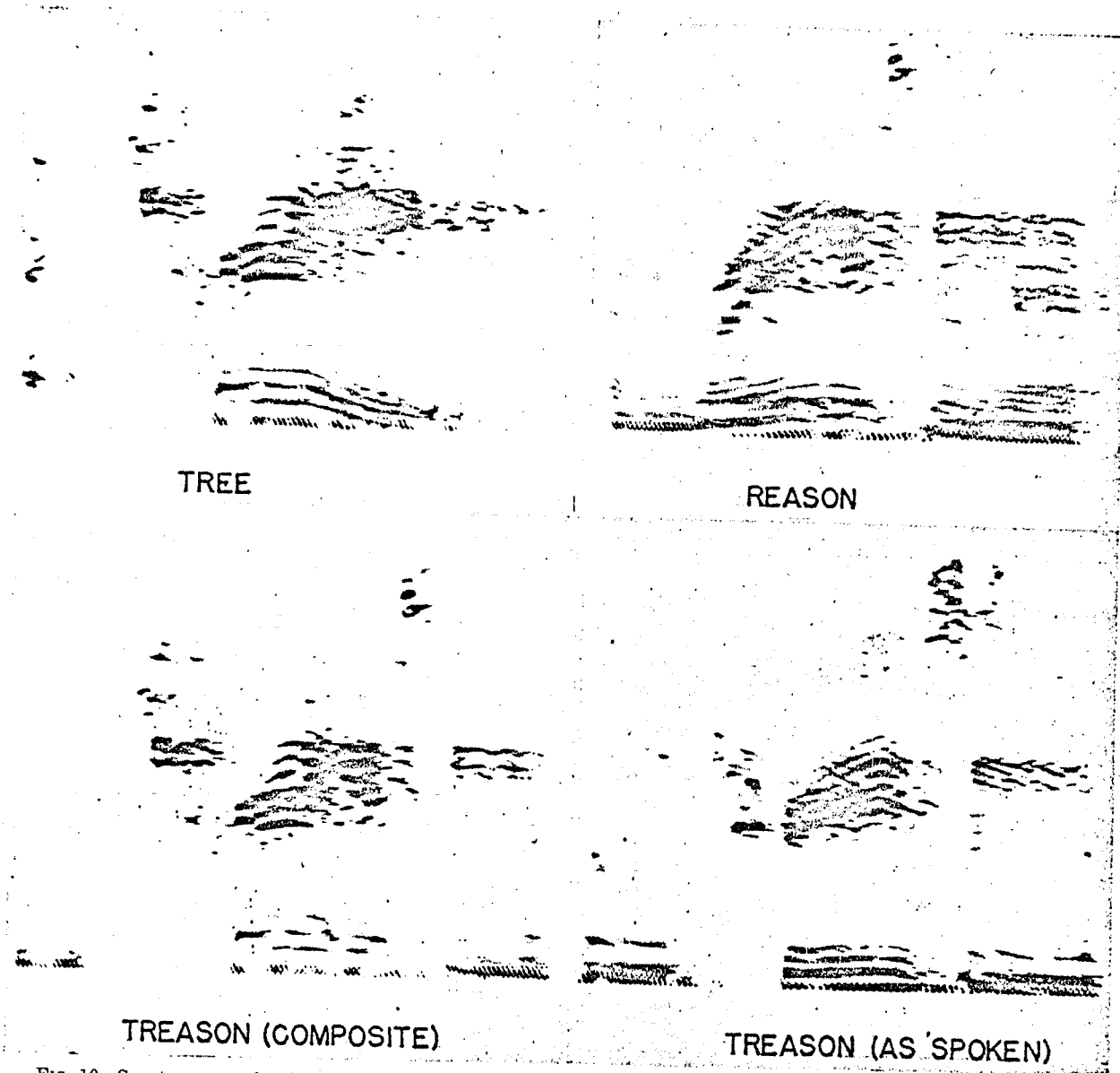

FIG. 10. Spectrograms showing how a composite "treason" was faked by tape editing. The sound was quite convincing and even the spectrogram does not clearly reveal the tampering.

speech somewhere far away. However, this particular instrument was built to save labor in painting the many involved patterns which are used in our experiments and which often vary in but one detail, and consequently it obtains its control voltages from a set of potentiometers.

The control voltages can be "programmed" or set up in advance for eight consecutive events, as shown in Fig. 12. A multiple-pole stepping switch then applies these preset voltages in the proper order. As an example, and taking but one variable—say the first formant frequency—there is a voltage divider for each event by which one can set the frequency of the formant generator. In addition, when the switch moves, the control voltage changes rapidly or slowly, depending on the time constant of an adjustable *RC* circuit. This permits one to make the transitions needed for many of the consonants. Also, successive events can have different durations, which are also set by potentiometers. The other variables can be controlled in like manner. Fig. 13 shows the control board with its many knobs. There is one knob for each variable, and they are all repeated eight times for the eight events.

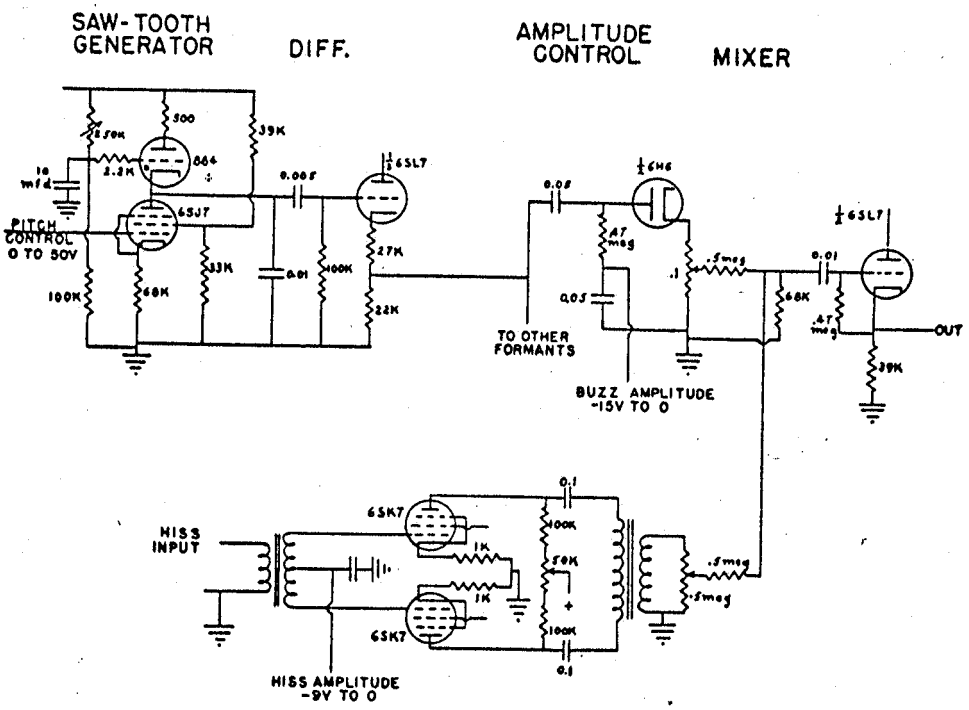

FIG. 11. Buzz generator and modulator which drive one formant generator of the speech synthesizer, Octopus.

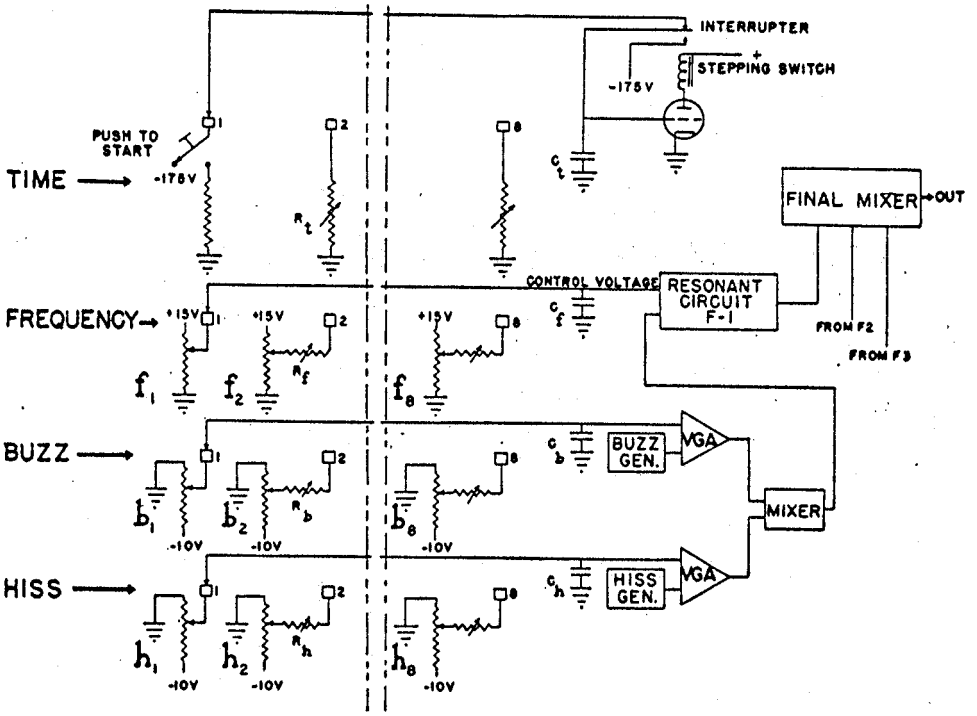

FIG. 12. Control voltages and stepping-switch connections of Octopus.

Some of the sounds, and the patterns corresponding to them, were demonstrated at the Speech Creation session by means of the following recorded commentary by André Malécot:

"We shall demonstrate some of the things that Octopus can be made to do by having it produce a pattern which will be changed one phone at a time in anagram fashion.

"The first pattern will produce the word *box* and each successive set of adjustments will result in a slightly different word. The accompanying chart (Fig. 14) shows the entire demonstration graphically.

"For the first word, which is *box*, Octopus has been set up to play the pattern as it is shown at the top of the chart.

"To change *box* to *socks*, we change the transitions of F1, F2 and F3 at the beginning of event 4, eliminate the voice bar of event 3 and substitute hiss at 3600 cps.

"To change *socks* to *sacks*, we change the steady state frequencies of event 4.

"To change *sacks* to *sack*, we need only eliminate the hiss of event 8.

"To change *sack* to *sash*, we eliminate the burst of event 7, change the transitions of event 5, and add hiss at 2400 cps in event 6.

"To change *sash* to *gash*, we eliminate the friction of event 3 and change the transitions at the beginning of event 4 in formants 1, 2, and 3.

"To change *gash* to *gas*, we eliminate the terminal transitions, those of event 5, and move the hiss from 2400 cps up to 3600 cps.

"To change *gas* to *guess*, we change the steady-state frequencies of event 4.

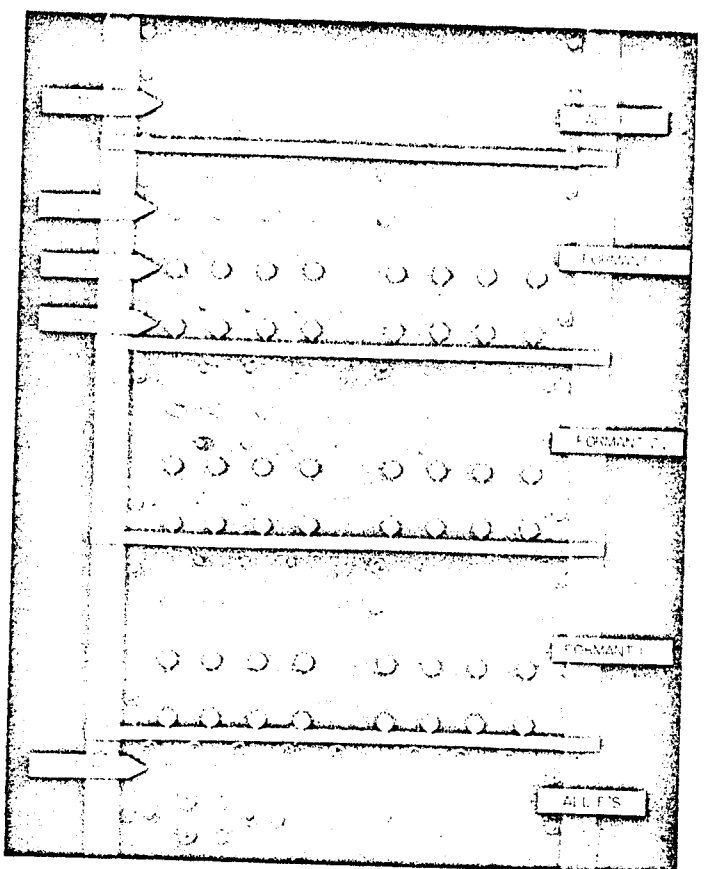

FIG. 13. Control board of Octopus, with functions indicated.

	WORD	TRANS.	SPECTRAL DISPLAY
1	box	baks	
2	socks	saks	
3	sacks	sæks	
4	sack	sæk	
5	sash	sæʃ	
6	gash	gæʃ	
7	gas	gæs	
8	guess	gɛs	
9	yes	jɛs	
10	"	"	SAME

FIG. 14. Spectrographic patterns and the corresponding words, which were spoken by Octopus in a recorded demonstration.

"To change *guess* to *yes* we change the rates of transition at the beginning of event 4.

"The following sounds, based on the same pattern, illustrate intonation variations of the word *yes*. The final yes is played with hiss substituted for buzz, as it would be when the word is whispered."

Finally, it was shown that the synthesizer could produce a singing voice. In fact, this is somewhat easier than producing the speaking voice, since the pitch variations required in singing are given by the musical score, whereas the intonation patterns used in speaking constitute a topic on which much additional research is needed.

This work was supported in part by the Carnegie Corporation of New York and in part by contract work for the U.S. Department of Defense.