XIV. SPEECH ANALYSIS

Prof. M. Halle
G. W. Hughes
J.-P. A. Radley

A. INVESTIGATION OF STOP CONSONANTS

The results of our first investigation of consonants, reported here last spring, were not altogether satisfactory. In particular, we felt that more accurate spectral data were needed.

1. Equipment

New equipment was designed to meet these major requirements: (a) It must permit the investigator to select a portion of any length at any place in the speech sample. (b) It must make it possible to obtain a spectrum (energy as a function of frequency) of the selected portion. (c) It must be possible to obtain the total energy in various wide frequency bands.

To satisfy the first requirement a gating circuit was built. Since our study samples are all on loops of tape we decided to trigger the gate by means of a pulse that is recorded on the loop at a place just preceding the sample. The gate is continuously variable in position and duration. Its maximum instability in position is less than 1 msec; in duration it is less than 0.1 msec. To obtain spectra, the gated samples are passed through a Hewlett Packard model 300-A wave analyzer whose frequency characteristics are adjusted to be flat for 170 cps and whose skirts have a slope of 64 db/decade. To measure the energy in wider frequency bands the samples are passed through a Spencer Kennedy model 302 electronic filter. By using the filter only as highpass or lowpass, an attenuation of 36 db/octave was obtained.

The output of either of these filtering arrangements is passed through a calibrated attenuator, amplified, and fed into a squaring circuit. The squarer was constructed in the Laboratory and is patterned after one designed by J. S. Rochefort (1). The circuit is accurate within 1 db over an input range of 40 db. The squared signal is then passed through an RC integrator whose output is kept constant for all measurements by adjusting the calibrated attenuator. The recorded data are the readings of this attenuator.

To minimize effects of tape noise and hum all measurements are taken with a fixed highpass filter inserted immediately after the gate circuit. Its critical frequency is 200 cps, below which there is an attenuation of approximately 24 db/octave.

2. Measurement Procedure

On tape loops the speech sample is available for measurement once every 3 sec. By comparing a full-wave-rectified display of the speech sample with a sonagram it is possible to locate the beginning of the stop "burst" with considerable accuracy. The

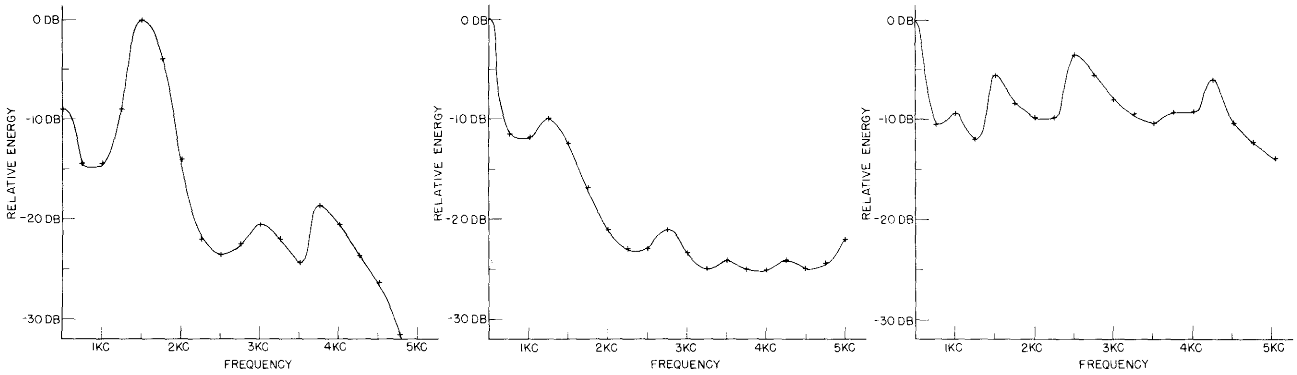

Fig. XIV-1
Energy spectra of voiceless stop consonants. Each of the stops was preceded by the vowel /a/.

length of the gate is then adjusted to 20 msec or to the beginning of the periodic oscillations that mark the beginning of the succeeding vowel, whichever is shorter. Spectra from 500 cps to 5000 cps and energy measurements in wider frequency bands are then obtained in the manner outlined above.

The samples used in our measurements were syllables of Russian containing all of the voiceless stop consonants in all possible combinations before and after vowels. Each syllable was pronounced by two speakers, a woman and a man. There were 116 samples in all.

3. Discussion of Measurements

(a) Compactness (2)

Inspection of the spectra revealed that in the majority of cases /k/ was characterized by a single predominating maximum located below 3500 cps, while in /t/ or /p/ such maxima were usually absent (see Fig. XIV-1). The position of the maxima seemed to be relevant for distinguishing the so-called hard /k/ from the "soft" /k/: the former had their maxima below 1500 cps; the latter were all above 2000 cps, usually considerably higher. (This distinction corresponds roughly to the distinction between the k's in such English words as "coo" and "key.") The effects of the following vowel were clearly evident in the case of /ku/ and /ko/, where the maxima were below 1000 cps and had particularly high concentrations of energy, about twice that of /p/ and /t/. In all other "hard" /k/ the maxima are between 1000 cps and 1500 cps. Out of 84 /p/ and /t/ which were studied there was only a single case of a maximum in this region.

On the basis of these observations it was possible to devise rules for the mechanical identification of the spectra as being, or not being, /k/. These rules operate correctly in 82 out of 84 cases of /p/ or /t/, and in 27 out of 32 cases of /k/ (of which, however, 3 cases of failure seem to be the result of faulty recording or of faulty pronunciation on the part of the speaker). As the material increases it will be possible to simplify the rules, which are now somewhat inelegant.

(b) Gravity (2)

It is usually impossible to tell whether a certain spectrum is that of /p/ or of /t/ simply by examining it. We had hoped to distinguish between these two on the basis of spectral data alone. We noticed in some filtering experiments that /t/ could be transformed into /p/ only if frequencies above 1000 cps were effectively eliminated. An attenuation of 18 db/octave was insufficient but with an attenuation of 36 db/octave the effects were unmistakable. Accordingly, we decided to compare the energy below some critical frequency with that above some other critical frequency. We reasoned that best results would be obtained if the two critical frequencies were separated by at least an octave.

We investigated 84 samples in this manner and succeeded in establishing rather simple rules for distinguishing between /p/ and /t/. These rules make use of the fact that in /p/ the low frequencies predominate. The identification procedure is as follows. The gated sound is LP-filtered at 800 cps and HP-filtered at 1500 cps and 3000 cps. If the LP-filtered output exceeds that of the 3000-cps HP filter by 10 db or more, the sound is identified as /p/. If the LP-filtered output does not exceed the 3000-cps HP-filtered output by 10 db, the 1500-cps HP-filtered output is compared with the 3000-cps HP output. If the difference is more than 4 db the sound is identified as /p/. In this manner it was possible to obtain correct identification in 83 out of 84 cases. As in the case of the /k/ we feel that with an increase in the data it will be possible to simplify and generalize the rules.

M. Halle, G. W. Hughes

References

1. J. S. Rochefort, Design and construction of a germanium-diode square-law device, M. S. Thesis, Department of Electrical Engineering, M.I.T., 1951.

2. For an explanation of these terms see R. Jakobson, C. G. M. Fant, and M. Halle, Preliminaries to speech analysis, Technical Report No. 13, Acoustics Laboratory, M.I.T., May 1952.

B. SEGMENTATION

Central to linguistic analysis is the view that speech consists of discrete units, the phonemes. The purpose of the research reported here is to investigate possibilities of automatically segmenting the continuous acoustical wave into intervals containing one

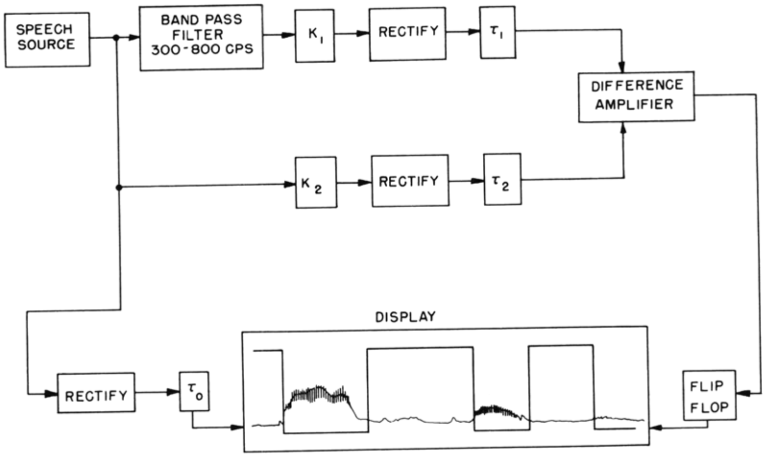

Fig. XIV-2
Block diagram of speech-segmenting circuit.

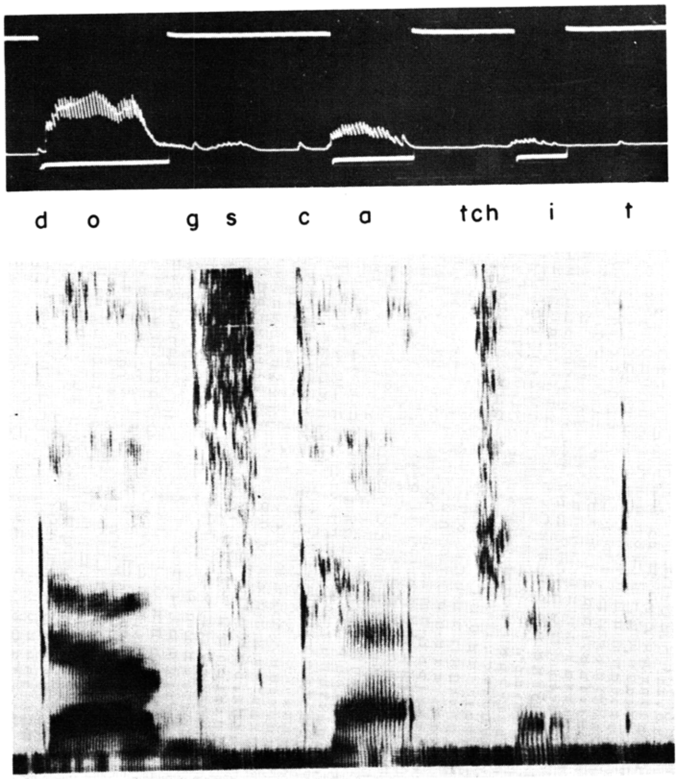

Fig. XIV-3

Above: A portion of film showing segmentation of vowels in the phrase "dogs catch it." Note that silence is interpreted as nonvowel in the binary result. Below: Sonagram of the same speech sample.

or more phonemes of certain types. The problem of segmentation is, of course, closely related to that of the identification of these phonemes, for a machine can only segment by recognizing that the characteristics of one phoneme have been replaced by those of another.

A circuit was devised to signal the transition from a vowel or sonorant (semivowel, liquid, and nasal) to a consonant proper or to a silence, and vice versa. The former group is characterized by having a larger fraction of the energy concentrated in a band from 300 cps to 800 cps. Figure XIV-2 illustrates the utilization of this property in the device to obtain a binary output. The bandpass filter was an SKL model 302. The circuit employs seven 6SN7 tubes and eight crystal rectifiers in two full-wave bridges. The gains, K_1 and K_2, and the amounts of smoothing, τ_1 and τ_2, of the rectified outputs, were adjusted to give good results on a few samples – tape recordings of a man and a woman speaking a series of sentences. The output of the segmenter and a rectified filtered trace of the original speech were photographed at the same time from the face of a dual beam scope (Fig. XIV-3).

The following observations were made: Sequences of several vowels and/or sonorants, and sequences of consonants proper and/or silence, are (of course) not segmented. Most trouble was given by /n/ and /m/: often a segmenting signal would appear in the middle of one of these, so that the phoneme was split in two. Sounds spoken too softly gave either no trigger or else had a rapidly fluctuating output. Difficulty was experienced with /z/.

A controlled investigation of the effects of varying K_1, K_2, τ_1, and τ_2 is planned. It is also proposed to use the same circuit to delimit fricatives, which have a concentration of their energy in the upper frequencies. The possibility of using the silence that precedes stop consonants as a segmentation signal is being studied.

J.-P. A. Radley