F. SPEECH ANALYSIS[*]

Prof. M. Halle
G. W. Hughes
Carol D. Schatz
C. P. Smith

1. Analysis of Envelope Features

As stated in "Preliminaries to Speech Analysis" (M. I. T. Acoustics Laboratory Technical Report No. 13, May 1952), the phonemes of speech may be broken down into inherent distinctive features which are the ultimate discrete signals. One of these distinctive features which deals principally with the envelope of the sound has been termed interrupted vs continuant. Work is now in progress to determine a suitable measurement scheme which will identify this feature without fail. This feature appears as the principal distinction between the class of consonantal sounds called stops and those termed fricatives. Studies have been made of these sounds as they exist in English and Russian.

Listed below are the stops and fricatives of English being studied.

Unvoiced				Voiced			
Interrupted		Continuant		Interrupted		Continuant	
p	(pill)	f	(fill)	b	(bill)	v	(veal)
t	(till)	s	(sill)	d	(dill)	z	(zeal)
k	(kill)			g	(gall)		
t͡ʃ	(chill)	ʃ	(shell)	dʒ	(Jew)	ʒ	(leige)

A list of monosyllables containing one or more of these sounds is recorded onto continuous loops of tape to provide for repeated observation of the sound to be studied. A sharp trigger must then be recorded preceding the sound on the loop to synchronize the oscilloscope or any gating apparatus used.

Since the waveforms of the sounds of interest in this study are random and pulse-like in nature, the envelope of the waveform rather than its instantaneous value is considered. This presentation is accomplished as follows: (1) the sound is full-wave rectified; (2) the rectified output is smoothed in an RC network; (3) results are displayed on a long-persistence, slow-sweep-speed oscilloscope. A smoothing time constant of 40 msec has been the most satisfactory so far.

[*] The work reported in this section was also supported in part by the Carnegie Foundation and the Rockefeller Foundation.

Several criteria for definitive measurement of the continuant distinction vs the interrupted distinction have been investigated. This measurement must, of course, be related to the difference in onset times. Numbers calculated from data to be proportional to the rise time of the waveform itself, its envelope, or the integral of the waveform failed to distinguish adequately the borderline cases of [t͡ʃ] , and [s], or [z].

In all measurement criteria evaluated there was little difficulty in distinguishing the rapid onset feature of the voiceless consonants. Difficulties became apparent in the case of the voiced consonants.

The sounds on the right side of the table consist, in general, of the corresponding sound on the left side plus a large periodic or voicing component. Highpass filtering beginning at a frequency of 1600 cps has been found to be fairly successful in removing this voicing. However, further study is needed to separate more adequately the true stop or fricative component from the masking of the periodic component.

On the basis of still incomplete measurements the following procedure has been adopted:

1. The sound is passed through a highpass filter, rectified and smoothed, and then displayed on an oscilloscope.

2. The time at which the wave first reaches 20 percent of its ultimate peak value is noted. (a) If in a segment of 150 msec preceding this point there is a silence which in turn is preceded by sound, the sound in question is interrupted. (b) If there is no silence in this period, the sound in question is continuant. (c) If no sound precedes the silence, we proceed to part 3.

3. The time at which the wave shape reaches its maximum is identified. The maximum value in volts is recorded, together with voltages occurring at 15 msec intervals preceding this peak up to and including 105 msec before the peak. We then calculate N according to the following formula:

$$N = \left(\frac{V_1 + 2V_2 + 3V_3 + 4V_4 + 5V_5 + 6V_6 + 7V_7}{p} \right) 10$$

where p is the peak value of waveform in volts, and V_n is the value preceding peak by 15n msec. (a) If N is under 20, the sound in question is interrupted. (b) If N is above 65, the sound is continuant. (c) If N lies between 20 and 65, we proceed to part 4.

4. We measure the number of milliseconds the envelope of the waveform remains above 75 percent of the peak. A time constant of integration of 40 msec is used. (a) If this time is less than 40 msec, the sound is interrupted; if more, it is continuant.

By applying these criteria we have been able to distinguish correctly the consonants in about 300 syllables spoken by two males and one female. The number of failures is less than 5 percent of the total material.

G. W. Hughes, M. Halle

2. Frequency Analysis

In order to obtain detailed information about the stop consonants, a technique was evolved for measuring the spectral distribution of any small time-segment of the speech signal in terms of the average amplitude in the 32 filter bands of the speech analyzer (cf. Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., October 15, 1952). The special problems presented were those of selecting the segment of the syllable accurately and providing a means of measuring the average amplitude of the rectified speech wave in the segment.

The utterances that have been the subject of the study were recorded on loops of magnetic tape. To provide a synchronizing signal for the long-persistence oscillographic display, a tone burst was recorded onto the tape immediately preceding the beginning of the utterance. We constructed, using a high-speed relay switch, time delay and switching circuits which make it possible to select a small segment of the utterance for measurement. In order to measure the average amplitude of the rectified segment of the speech signal so selected, a balanced, low-drift RC integrator was built. This integrator contains polystyrene-dielectric capacitors to minimize leakage and soakage effects.

The segment of the speech signal for which the spectrum is to be determined is selected by observing the pattern of the rectified speech signal on the long-persistence cathode ray tube, and listening to the gated signal. The vowel and consonant portions of the signal are readily identified in terms of their characteristic patterns, even to the separation of the burst and the frictional modulation of the stop consonants. After adjusting the timing circuits in order to suppress all but that segment of the sound which is to be measured, each filter band is switched on in turn and a calibrated attenuator is used to adjust the level of the speech signal to a standard reference amplitude, as measured in the RC integrator. The attenuator readings, then, provide a measure of the relative distribution of amplitudes in the various filter bands.

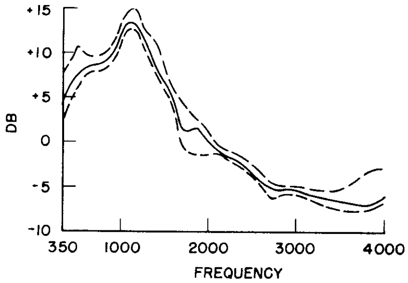

Fig. VIII-4

Average and peak deviations for the stop /k/ in 12 different utterances of the syllable /ku/.

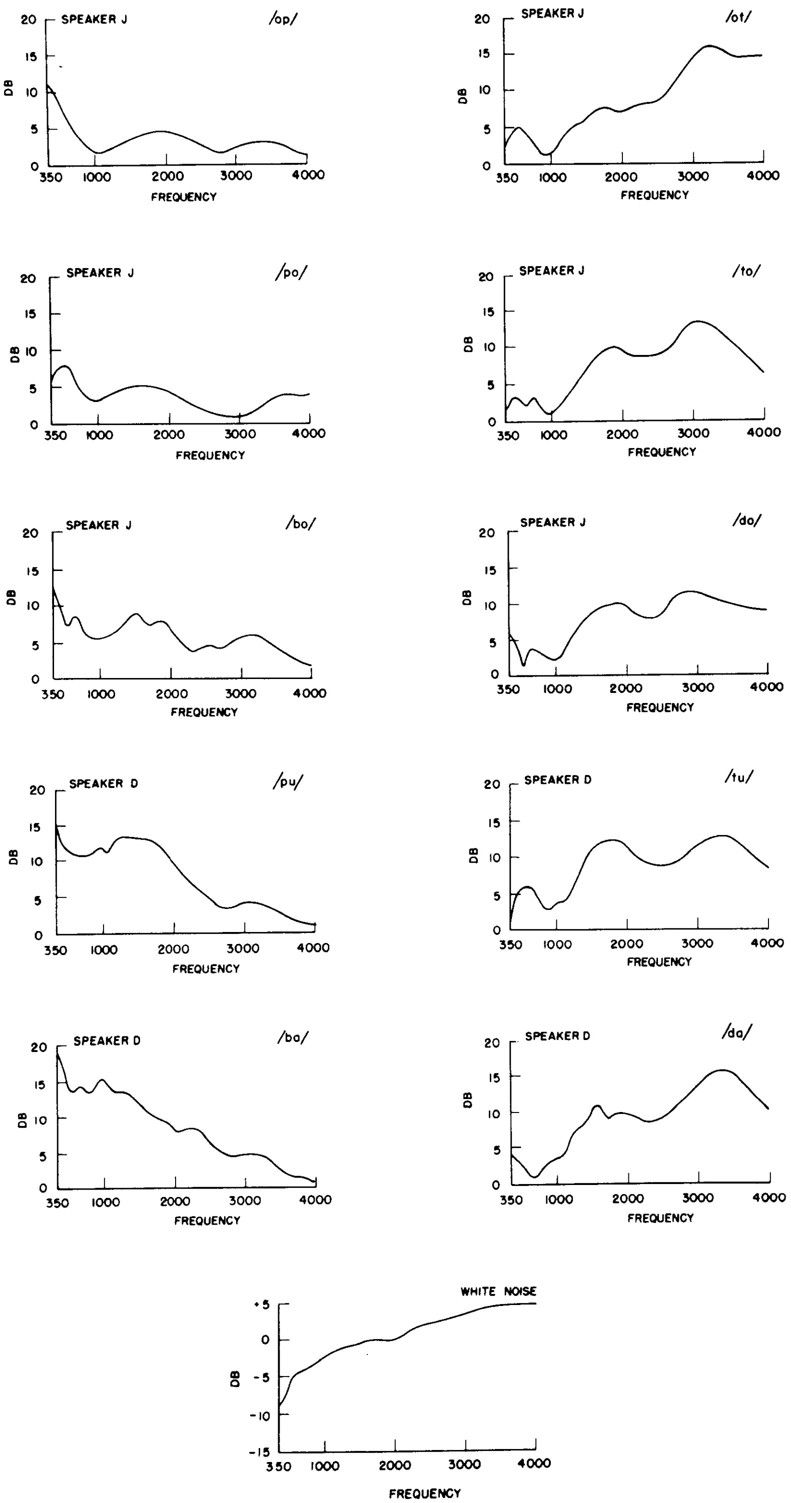

Fig. VIII-5

The spectra for /b/ and /p/ have a generally falling slope; the spectra for /d/ and /t/ have a rising slope.

In order to assess the variability of speech sounds measured in this manner, a comparison was made of 12 utterances of the syllable [ku] by a single speaker of American English. The burst and frictional modulation portions of the stop consonant [k] were measured for each of the 12 utterances, and the average distribution for the 12 utterances is plotted in Fig. VIII-4, as well as the peak deviations from this average.

The Russian hard /t/, /p/, /d/, and /b/ in all possible combinations with all the vowels were recorded on tape by two native male speakers. The syllables were then subjected to measurement according to the outlined procedure. Some of the spectra so obtained are given in Fig. VIII-5. It can be seen that /t/ and /d/ have more intensity in the upper frequencies, whereas /b/ and /p/ are stronger in the lower frequencies. Clearly marked maxima appear in some of the consonants; however, these maxima apparently do not correspond to the formants of the adjacent vowel.

C. P. Smith, M. Halle