XI. SPEECH ANALYSIS

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

A. RECOGNITION OF SILENCES

The class of speech sounds known as stops or plosives, e.g., /p/, /k/, /t/, /d/, /b/, and so on, are produced by a complete closure of the vocal tract followed by a sudden opening. During this movement the vocal cords may be vibrating. This vibration distinguishes voiced stops like /b/ and /d/ from voiceless ones like /p/ and /t/.

The acoustic correlate of this articulatory process is an absence of energy in a broad frequency band followed by a rapid onset of energy in all frequencies. For the voiceless stops, the absence of energy is more or less total; for the voiced stops, strong components are usually present in a region below 300 cps but above this region there is hardly any energy. This relative silence is thus one of the cues that permit us to identify stops.

The operation to be performed is that of determining when the input to a device falls below a certain low level and remains there for some minimum time. To recognize a critical level, a Schmitt amplitude discriminator is used after the speech signal, suitably filtered and amplified, has been rectified (full-wave).

In the first method tried, the output of the Schmitt circuit was given a certain amplitude which controlled the onset and timing of a highly unbalanced multivibrator. Impulses from this stage were fed to one side of a flip-flop. When the silence was ended, the flip-flop was returned to its first state and the multivibrator turned off. If a silence did not last long enough, the multivibrator was turned off before emitting its first impulse. In fact, at every half-pitch period, the signal could be expected to reach zero for a short time. It was, however, not possible to reset the multivibrator quickly enough, which led to cumulative timing errors. A new approach was, therefore, developed.

In the second method, the output of the Schmitt stage, a rectangular wave, is used to generate a negative-going triangle whose slope can be varied and which resets very quickly (with respect to audio frequencies) to its origin when the Schmitt tells it to. This triangle is coupled to a "simplified" Multiar (1). A pentode, normally heavily conducting, is cut off by the end of the triangle excursion. In the Multiar a transformer is used in a feedback path for rapid cutoff action, but this is unnecessary in this application, since the tube is cut on rapidly enough by the essentially vertical return of the triangle. It is this rapid return that marks the end of a silence. The location of the beginning of the silence does not have to be known with great accuracy for present purposes.

The circuit was tested with short sentences read by four English speakers (one

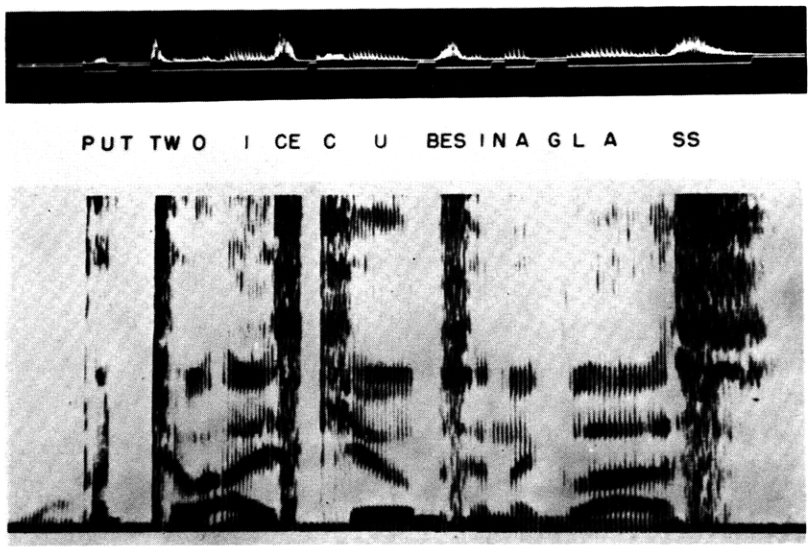

Fig. XI-1

Lower trace signals stops by "end of silence" mark. Note that /n/ is wrongly identified.

female and three male). Because of the relatively gentle slope of the filter available, the signal was highpassed at 1500 cps or 2000 cps to eliminate the voicing component. The critical timing was set at about 16 msec. This is the minimum time for which the amplitude must remain below the critical level before a silence can be noted. In other words, no silence shorter than the critical timing is registered. The hysteresis effects of the Schmitt circuit were arranged to be least for sine waves between 30 cps and 300 cps. A filmed record was made with a dual beam cathode-ray oscilloscope showing a rectified trace of the speech on one trace and the output signal of the "silence recognizer" on the other. See Fig. XI-1.

As expected, the two criteria (level and time) used were necessary but not sufficient for the identification of stops. Different settings will mark silences before, or during, different classes of sounds. Almost all stops in the samples were marked by the end of a silence. Exceptions to this were a few very weak stops that did not reach above the critical amplitude (set just above the noise region) and the flapped /t/ as in "water," in which the silence is extremely short.

On the other hand, because of the very high cutoff frequency of the filter, sounds such as the nasals and the semi-vowels in which most of the energy is located in the low frequencies were judged to be silences. Better filtering might eliminate this.

J. -P. A. Radley

References

1. B. Chance et al., Waveforms, M.I.T. Radiation Laboratory Series, Vol. 19 (McGraw-Hill Book Company, Inc., New York, 1948).

B. THE NASAL CONSONANTS

The nasal consonants have in common with the vowels a spectrum that is dominated by a small number of resonance peaks, the so-called formants. In both the nasals and the vowels the lowest resonances (the first formant) contribute most of the energy. The vowels differ from the nasals in having higher first formants. Furthermore, if the spectrum contains formants above the first formant, these are much less attenuated in the vowels than in the nasals. The first formant of the nasals is generally below 300 cps. The lowest first formant in vowels is found in [i] and [u], where it is located at about 300 cps. The first formant in the other vowels is well above 300 cps. These facts suggested that a comparison of the energy below 300 cps with that above might yield information useful in distinguishing between vowels and nasals.

Our speech samples were isolated syllables consisting of a vowel preceded or followed by a nasal. These syllables had been recorded on tape by one female and two male speakers. Portions 50 msec long were gated out of the nasal and the vowel. The gated portion was passed through a Spencer Kennedy 302 variable electronic filter in which both sections were set either to 300 cps highpass or 300 cps lowpass. This gave us an attenuation of 36 db per octave. We found that a lesser attenuation (18 db per octave) yielded no results of interest. The output of the filter was passed through a variable precision attenuator, integrated and measured.

The following results were obtained: (The numbers in the table are the extreme values found when the lowpass output was subtracted from the highpass output.)

vowels (113 measured)	-8 db to +15 db
[i]	-8 db to +4.5 db
other vowels	-2 db to +15 db
[m] (37)	-15.5 db to -8 db
[n] (37)	-17.5 db to -3 db

If we omit the vowel [i] from consideration it is possible to separate the vowels and the nasals quite well by this criterion: the vowels give positive and moderately negative values, while the nasals give strongly negative values. The vowel [i] could be made to give more positive values if frequencies above 1000 cps were boosted. Such a weighting would affect the nasals but little, because they have very little energy in that region; however, it would greatly improve our chances of separating [i] from the nasals. In the next quarter we intend to investigate this further.

An attempt was also made to establish the difference between the two nasal consonants. For this purpose we measured spectra of these sounds (1). No consistent difference could be found between spectra of [n] and of [m]. See Fig. XI-2.

We pursued this question further and attempted to synthesize some vowels and the

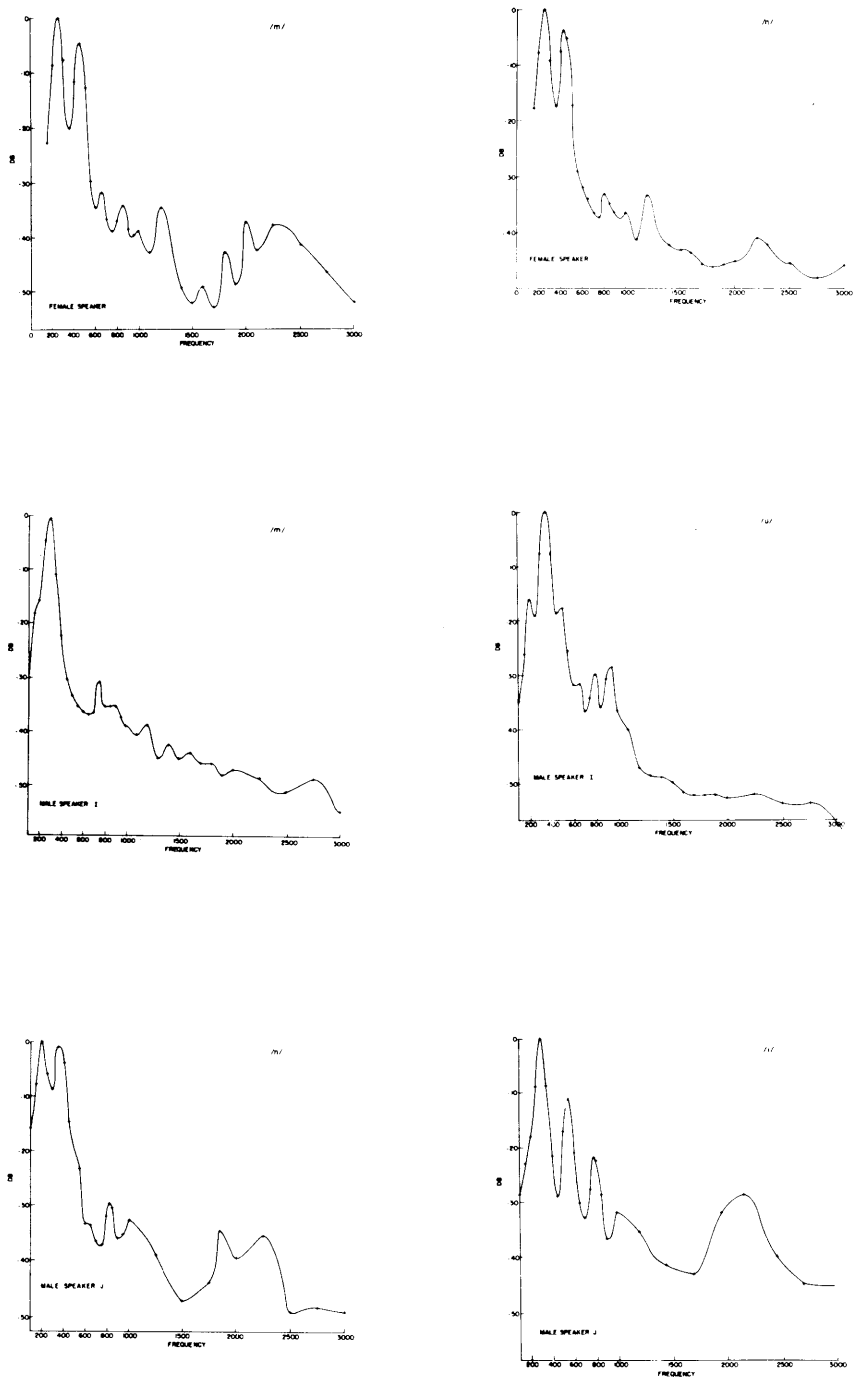

Fig. XI-2

Spectra in the first row were obtained from the syllables /mi/ and /ni/. Spectra in the second and third rows are of the nasal and the vowel in the syllables /mu/ and /ni/, respectively. Note the predominance of the low frequencies in all spectra.

nasals by means of oscillators. A good [m] could be synthesized by a single oscillator tuned at 150 cps. When a second oscillator tuned to 300 cps was added, no appreciable difference was noted until the output of the second oscillator was substantially equal to that of the first. Under these conditions [u] was perceived. In order to obtain [n] it was necessary to introduce a very weak component (-40 db below the fundamental) at about 2300 cps. When this component was increased in intensity to -17 db a sound much like the German [ü] was heard. The best [n] was synthesized by introducing two fairly strong components (-20 db) at 600 cps and 750 cps in addition to the above-mentioned weak component at 2300 cps. It was, however, observed that the 2300 cps component was much more essential in the perception of [n] than the 600 and 750 cps components.

M. Halle, G. W. Hughes

References

1. Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1954; July 15, 1954.