ACOUSTIC CUES IN SPEECH:

FIRST REPORT

Pierre DELATTRE

(This article, in its original French version, appeared in the journal Phonetica Vol. 2 (1958) pp. 108-118, 226-251.)

I. Introduction.

The last ten years will rank as important ones in the history of experimental phonetics. With entirely new research techniques at our disposal, especially those enabling us to manipulate artificial speech by electronic analysis and synthesis, it has been possible to single out the elements of speech. By varying the dimensions of these elements at will, we are in a position to judge the effects of such changes on perception, thus achieving dependable advances in the study of the physical nature of speech. Particularly important strides forward have been made in the study of those acoustic cues which underlie the sounds of speech and linguistic identification of them.

One the notion of formant was established, the greatest contribution of this decade unquestionably concerns the part played by formant transitions in the perception of consonants. (We shall see below how these transitions correlate with place rather than manner of articulation, i.e., they are used to distinguish among the various consonants of one class rather than among different consonant classes.) Neither the kymograph nor even the oscillograph had led us to suspect, with respect to stop consonants, for instance, anything more than an occlusion followed by a burst. Spectrographic evidence immediately brought out the fact that rapid frequency changes, the counterpart of articulatory movements, connect consonant with vowel centers. Our ability to state that these changes, which look rather transitory, do not play a 'side' part, but are at the very heart of consonant perception, is due to work with synthetic speech techniques. Over these last years more time has been devoted to studying the function of such transitions, to determining how to specify them, than to all other factors combined -- steady-state vowel formants, bursts, friction, affrication, stress, intonation, rhythm, etc.

Investigation of these cues is far from complete, but it is already at the point at which, by applying the acoustic rules that have come out either through definitive research, or exploratory studies, or extrapolation, it is possible to paint an 'artificial' spectrogram at the rate of a syllable a minute.

This spectrogram, moving before the electronic 'eyes' of a synthesizer -- somewhat like a player-piano roll -- produces quite intelligible speech.

Such an advance is obviously due to technical improvements originating during the war. They do not, however, represent a break with the past as we shall see by a brief survey of what came before.

Analysis of human speech waves, conducted electronically as by GEMELLI and PASTORI, or mechanically as by the abbé ROUSSELOT, or by simply auditory means as by Sir Richard PAGET, had enabled us to accumulate, well before the war, a stock of exact information about acoustic cues in speech. The many studies on the frequency and number of characteristic vowel notes were not far from yielding results. In our opinion, PAGET was the man heading in the best direction. He was the first to state that all vowels, not just front vowels, had at least two distinctive formants. In the area of consonants, PAGET had clearly perceived definite fricative and liquid frequencies, as much within the realm of periodic waves -- [l r ɳ] -- as the non-periodic -- [s ʃ]. He had put his finger on one of the most obscure cues for consonantal nasality (Human Speech, p. 95) -- a formant at about 250 cps which plays such an important role in today's synthesis in transforming a consonant of the type [b d g] into one of the type [m n ŋ] and contributes in a definite though lesser degree to the nasalization of vowels. He was even aware, so it seems, of the part played by transitions in consonant perception: "...inasmuch as [consonants] are produced by movements of the vocal organs (like the diphthongs) their resonances are characterized, not only by pitch, but also by their change and rate of change in pitch." (Human Speech, p. 124) And, as an example, he cites the [l] in [il], characterized by a falling transition of five half-tones, and the [l] in [ul] characterized by a rising transition of eleven half-tones, thus anticipating the recent 'locus' hypothesis which attributes the acoustic character these two transitions have in common (the thing that enables us to identify two such different transitions as a single phoneme) to the fact that both of them converge on one virtual frequency, an [l] -locus with transitory characteristics of the same sort: "...if the terminal [l] be sounded by itself, as a continuing sound, it becomes quite unrecognizable. The resonant change is the real characteristic, in spite of its great difference with different associated vowels." (Human Speech, p. 124) Finally, PAGET

had even foreseen the proprioceptive role of articulatory gestures in perception -- an hypothesis confirmed by the results of many synthetic experiments we shall mention in this report: "In this case [of l] the one constant characteristic is the movement made by the tongue....in recognizing speech sounds, the human ear is not listening to music, but to indications, due to resonance, of the position and gestures of the organs of articulation." (Human Speech, p. 125)

Later, it was Martin JOOS' merit to specify the possible importance of transitions: "Such identification of consonants by their effects upon contiguous resonants is apparently depended upon by listeners to a far greater extent than commonly supposed." (Acoustic Phonetics, p. 122) JOOS, however, deserves less credit for such hypotheses than PAGET, since for some years JOOS had had at his disposal that incomparable speech research instrument, the spectrograph, invented during World War II at the Bell Telephone Laboratories. On a spectrogram the formants stand out plainly as they move together and apart in the time dimension, reflecting in this way the continuous aspect of the various articulatory movements.

Terminology. Before going into the technical aspect of our subject, let us make explicit the terminology to be used here.

Frequency, a physical, acoustic concept, is objectively measurable in cps (cycles per second); it is subjectively perceived as pitch, a psychological concept. Intensity, a physical concept, is objectively measurable in db (decibels); it is subjectively perceived as loudness, a psychological concept. Duration, a physical concept, is measurable in msec. (milliseconds); it is subjectively perceived as length, a psychological concept.

On spectrograms of speech, formants are what we call certain areas of greater intensity. Formants average about 200 cps wide. On a three-dimensional spectrogram (frequency, duration, intensity) each formant is recognizable as a concentration of darkness moving up or down, in time, as its frequencies change. The concentration is solid for non-periodic sounds in general and for periodic sounds in wide-band display. It is not solid for periodic sounds in a narrow band display which reveals the individual harmonics.

Since this report is strictly about acoustic cues, we shall avoid, insofar as possible, the subject of correlations between acoustic and articulatory factors, which deserves a separate report. Let us simply mention that the resonance notes of the various phonating cavities do not correspond to the frequencies of the different formants in a direct and independent way. No single cavity is directly and independently responsible for the frequency of any one formant. Any formant depends, more or less, on the sum of the cavities, especially as the total (sum) cavity approaches the form of a tube of uniform cross-section. The 1st, 2nd, 3rd, etc., formants are the 1st, 2nd, 3rd, etc., modes of vibration of a tube closed at one end and open at the other; thus F1 (the first formant) frequency corresponds to the one-quarter wave length; F2 to the three-quarters wave length; F3 to the five-quarters wave length; etc. With the distance from the glottis to the lips at 17 to 18 cm. and the velocity of sound in air at 34,400 cm. per second, the frequencies of F1, F2, F3, etc., in a vowel articulated through a cavity of almost uniform cross-section, are, respectively, 500, 1500, 2500, etc., cps. (For F1, the calculation is: 34,400 divided by 17.2 divided by one-fourth; for F2: 34,400 divided by 17.2 divided by three-fourths; etc.) We rarely find formant frequencies in the proportion 1, 3, 5, etc.: They are usually higher or lower depending on whether contrictions in the phonating cavity correspond to nodes or loops in the respective modes of vibration. Nonetheless, we may say that the more the front and back cavities take definite shape (as, for instance, in [u]), the more acceptable a practical correlation is between F1 and F2 frequency changes and changes in front and back cavity volumes; at the same time, formants above F2 have less of an effect on perception, their intensities being too weak.

II. Present-day research techniques.

Spectrography. Spectrography gives us a visual image of speech broken down into its acoustic elements. This display is called a spectrogram. Using a spectrograph of the Kay Electric Company (Pine Brook, N. J.), it takes only a few minutes to make a spectrogram of 2.4 seconds of sound on a paper sensitive to an electrical spark. The displayed spectrum, much like a TV image, is made up of some 200 lines which appear darker as the intensity of a particular frequency rises above a particular level. The resulting image is three-

dimensional: in frequency (from bottom to top), in duration (left to right), and in intensity (degree of darkness). Before 1948 Bell Telephone's experimental spectrograms covered a frequency range of only 3500 cps (about the lower limit of [s] friction). Visible Speech, the well known book by POTTER, KOPP and GREEN of the Bell Telephone Laboratories, is illustrated mainly with such spectrograms. Kay spectrographs since 1948 cover linearly (not logarithmically) a frequency of 8000 cps -- 2000 cps per inch of height. The total picture of these 2.4 seconds is four inches high and twelve inches long.

In addition, the Kay spectrograph operates with two different filterings: 1) a wide-band filter (300 cps) which brings out the formants but conceals the individual harmonics that make them up; and 2) a narrow-band filter (45 cps) which brings out the individual harmonics but makes it harder to read the formants since the eye must now extrapolate to bring the separate harmonics together into solid formant bands. Several improvements have been added to these two sound displays. 3) At any point on the time axis, a profile or section of the amplitudes of all harmonics at that point can be made with either narrow or wide band filter. This cross-section (directly measurable in decibels) makes the intensities of the harmonics more explicit than does degree of darkness. 4) The total intensity can be displayed ('amplitude display') as a function of time by a moving line at the top of the spectrogram. This is also measurable in db from a zero line parallel to the time axis. 5) To get a better display of formant-frequency variations, it is possible to make spectrograms on a scale of 1200 cps to the inch (instead of 2000) which corresponds to the scale of the special spectrograph at Haskins Laboratories in New York. (This did not occur by chance: Haskins Laboratories furnished the specifications for this improvement to the Kay Company!) 6) To get a better display of fundamental frequency variations, it is possible to make spectrograms on a scale of 200 cps to the inch, a ten-fold amplification which causes the lowest 800 cps to occupy the entire vertical scale.

The spectrograph built by Haskins Laboratories for their own use is considerably more sensitive and flexible than the Kay machine. Spectrograms can be made on an infinite number of frequency scales and displayed linearly as well as

logarithmically (although this latter is less favorable to the eye). These spectrograms ordinarily have a frequency scale of 1200 cps per inch and a length of 79.2 inches for an eleven second duration. These are also the dimensions of the artificial spectrograms used in synthesis at Haskins Laboratories.

To compensate for the fact that in human speech there is an intensity drop of about nine db per octave frequency rise, spectrographs usually equalize upper intensities by that amount. In this way the higher formants are made quite as visible as the lower.

Analysis is the normal starting point in research. It enables us to make hypotheses to be confirmed in experiments with synthesis. Thus we may compare the spectrograms of two sounds the ear hears as different and see which formants have appeared, disappeared, changed in frequency, intensity, duration, shape, direction, etc. We may also attempt to see spectral differences resulting from separate articulatory changes. This method, however, is never entirely reliable. Just how far is it possible to modify the position of a single articulator while holding the others steady? How do you move the back of the tongue without changing the size of the pharynx? A cross-check using X-ray films is possible but not easily accessible: to take X-ray films of articulatory movements requires radiographic intensity of such magnitudes that a single subject may not be exposed for more than fifteen seconds a year without risk.

To check on what articulatory differences produce particular spectrographic differences an artificial vocal tract is needed. With this end in mind, first the Bell Telephone Laboratories (5) and then Massachusetts Institute of Technology (25) built electronic analogs of the vocal tract.

Analysis can thus only suggest what the speech cues may be. Synthesis must then provide verification. Example of erroneous conclusions based on analysis alone abound. Here are a few concerning the acoustic cues for nasality: Harvey FLETCHER attributed nasality of the addition of two formants, the lower at about 400-450 cps and the higher between 2169 and 3906 cps (Speech and Hearing, p. 63); Antti SOVIJÄRVI to three formants at about 2000, 2500, and 3000 cps (Die Vokale und Nasale der finnischen Sprache, Helsinki 1938, p. 161); Thomas

Using synthesis, it is possible to filter out the formants themselves, in spite of their changing frequencies.

Splicing. It is possible to cut up magnetic tapes and splice them back together either leaving out a middle segment or exchanging segments of several different specimens (or several versions of one type) among themselves. For instance, we can splice in an [m] where there was an [n], the friction of an [s] where there was an [f], etc., then test the results by ear. Since it is very difficult to make cuts at just the desired points, the results of splicings must always be checked by a spectrogram.

This method has its special advantages in testing how far the results of synthesis apply to natural speech.

White noise. Using noise in very wide or narrow frequency bands superimposed on speech during perception makes it possible to rate the comparative strength of acoustic features.

Synthesis. So far, Haskins Laboratories have built three speech synthesizers which we shall call SP, SV and SO. SP and SV were made to convert eleven second spectrograms (79.2 inches) into sound; the spectrograms may be 'natural' ones of the human voice as well as 'artificial' ones painted by hand. SO can speak only in isolated syllables. All three were made not to give the most realistic speech possible (phonographs and tape recorders take care of that), but to serve as good research instruments. The qualities demanded of them are flexibility and versatility: they should make it possible to separate out the many acoustic elements that make up speech and vary them in every dimension, all the while allowing us to hear the results of our manipulations. Descriptions of these research instruments are to be found in (4), (6), (18), (34). We shall limit ourselves to explaining how they are used in actual research.

Imagine a man's voice of middle register speaking in a monotone at 120 cps: all the partials of the voiced sounds over the spectrum are harmonics of the fundamental tone, grouped into formants selected by the buccal cavities. To imitate this masculine voice, SP possesses a gamut of 50 pure tones (sinusoidal waves) corresponding to the first 50 harmonics of a 120 cps fundamental (i.e., 120, 240, 360, .. , 6000 cps). These pure tones are activated by 50 light beams

with an individual width of a tenth of an inch. The 50 beams cover a width of five inches -- like a spectrogram -- and are arranged to correspond to the frequencies of the 50 harmonics of a natural spectrogram of the same size. As a spectrogram passes before the light beams at a given constant speed, all the pure tones corresponding to the formants present are activated. Each formant activates, on an average, three adjacent tones with the middle tone typically more intense than the two outer ones. Except at the lower end of the frequency range where the intervals are large, a formant of three adjacent tones played by itself sounds as a violent dissonance to the ear. However, two formants, that is, six tones in two groups of three adjacent tones, are heard as a perfectly good vowel -- no longer as a dissonance -- provided simply that the formant frequencies correspond to a vowel the listener is familiar with. To make an artificial spectrogram, a line is painted parallel to the time axis for each desired harmonic. The higher the frequency, the higher the line; the longer the duration, the longer the line (7.2 inches equals one second); the more intense the tone, the wider or brighter the line (up to one tenth of an inch per harmonic). In practice, a formant is painted as a solid line covering one channel (harmonic) completely and the two neighboring ones halfway. Any solid line produces a periodic sound. To simulate non-periodic sounds, random dotting is used. Friction noises so produced -- [s f] etc. -- are less natural than vowel sounds, though still satisfactory. Bursts -- [p t] etc. -- are painted as short lines, more or less vertical, about 600 cps wide, i.e., five neighboring channels, and come out rather satisfactorily. Variations in the frequency of the glottal tone cannot be imitated on SP with its fixed fundamental at 120 cps. It must be admitted, however, that if speech SP produces is intelligible without intonation, its worth is all the greater.

SV is a much improved instrument compared to SP, but in one sense less flexible though it plays the same spectrograms. 1) On SV, a single painted line automatically produces a complete formant (varying in intensity with the width of the line) -- and this so that the formants shall always be of a sort closer to the formants of natural speech, as much in wave damping as in phase relations of the various harmonics in the formant. (Phase plays no clearly distinctive role from the linguistic point of view though it seems to contribute to naturalness and thus to intelligibility.) Thus SV formants

constitute an improvement when it comes to naturalness. On the other hand, a certain degree of flexibility is lost since the individual harmonics are no longer under our control as on SP. 2) SV produces true noise. One solid painted line can be heard as a formant of periodic sound or as a formant of turbulent (non-periodic) sound. (On SP this effect of turbulence is achieved by random breaks in the periodic sounds.) The friction of fricatives, whispered vowels, [h], and aspiration are thus better produced on SV than on SP. 3) On SV the fundamental frequency can be varied. (Of course, when the fundamental rises in frequency, all the harmonics also rise, while the formant frequencies remain unaffected.)

SO is of a completely different type. No spectrogram is turned into sound; the control panel consists of switches enabling one to produce a syllable with intonation. SO was built to facilitate the study of transitions. The beginning and end (frequency and time) of three automatic formants (like SV but still more natural) are under the operator's control and they may be periodic or not. The rate of change of the transition curves (though not their shape), the intensity of each formant, as well as the duration of any sound segment are also under the operator's control. SO can produce at most eight segments of successive sounds. With all the switches set as desired, the entire syllable is sounded.

Another synthesizer, by the name of PAT, built in England, has just been put into experimental use (47). It produces four formants (automatic ones as on SV and SO), periodic and turbulent sounds, and changes of fundamental, hence harmonic frequency to give intonation. As a research tool, PAT is of the same kind as SP, SV and SO (though less flexible) in the sense that the acoustic variables of speech are manipulated separately, either by control switches or by profiles on a projection plate, and not of the analog type on which the 'articulatory' variables of a vocal tract are manipulated to determine what changes are wrought in the spectrum. (In analogs, generally speaking, for one articulatory modification, changes in all the formants are observable.) Moreover, PAT resembles SP and SV in that it follows formant changes in time and thus produces phrases; analogs have so far produced only sustained, isolated sounds.

III. The results.

Bibliography: In section XVI we have listed the works which, unless we err or omit, have contributed to our knowledge of acoustic cues in speech over the past ten years -- roughly 1947 to 1957 -- even if only by their erroneous conclusions. The items are numbered chronologically.

The first question to be asked when faced with a spectrogram is: "Over these 8000 cps of formants, where in such a wealth of acoustic detail are the distinctive features from a linguistic standpoint?" As soon as the first Haskins Laboratories synthesizer was ready and tests on artificial speech coming from 'natural' spectrograms were satisfactory, exploratory work set out to answer this question. The many formants were successively suppressed (one at a time, then in groups) and the results of such omissions were submitted for identification by ear (11), (12), (13), (18). This soon made it apparent that aside from a few turbulent sounds -- especially dental and alveolar friction and bursts -- the lowest three formants, often even the lowest two formants, included all the principal distinctive features. With this much to go on, an attempt was made to see how far hand-painted spectrograms could simplify the still rather complex look of the two or three lowest formants of a natural spectrogram without losing intelligibility when reconverted back into sound. Simplification was also carried to a point where there was a partial loss of intelligibility. Thus, when the curves from the three distinctive formants were successively replaced by straight lines, the simulated effect for the ear was of tightening the jaw (F1) or the tongue (F2). By slowing the machine down, or speeding it up (which does not change the fundamental) the manner of articulation was partly changed: a voiceless consonant became voiced, a stop turned into a fricative or a semivowel, a liquid or semivowel changed into a vowel or diphthong. Varying frequencies for turbulent waves and giving other directions to F2 and F3 transitions mainly changed the place of articulation. Broadly speaking, by manipulating every imaginable aspect of a spectrogram, it was possible to single out the acoustic cues, determine to what extent their dimensions could be changed, and so specify their individual roles in perception.

IV. Explosives, or oral stops.

As a consonant class, the oral stops are chiefly distinguished by the amount of interruption in the vocal sound, the shortness of the intense turbulence or burst that follows, and by the rapid transitions leading into a following vowel or from a preceding one. This has been the class most studied, probably because it has appeared to be the most challenging. (The fricatives seemed so simple to synthesize that little attention was originally paid them.)

Bursts. The first systematic experiment with SP was designed to study the effects of initial voiceless stop bursts (16). The syllables to be identified were made up of a synthetic burst followed by a synthetic vowel with two formants of three harmonics each. The bursts looked like vertical ovals 600 cps wide and 15 msec. long. They were assigned twelve different frequencies from 360 to 4320 cps and combined one at a time with each of seven cardinal vowels [i e ɛ a ɔ o u] giving a total of 84 synthetic syllables. These syllables were then played in random order from a tape recording to 30 phonetically untrained subjects with a request to identify them as one of [p t k]. The results were clear. High bursts, from 3000 cps up, were identified as [t], the rest as [k] or [p] depending on whether they were located just above the beginning of F2 -- [k] -- or elsewhere -- [p]. Examination of the results further showed that the effect of burst frequency was not independent of the vowel. In one particularly striking case, a burst located at 1440 cps was heard as [p] before the vowel [i] and as [k] before the vowel [a]. On the other hand, bursts of extremely different frequencies were heard as the same consonant. Thus, on the one hand, a single sound can be identified in two different ways; on the other, two very different sounds can be identified as the same. Two hypotheses were already postulated which would be borne out by later experiments: 1) The smallest acoustic unit in speech is the syllable. 2) If there is an invariant enabling us to tell one place of articulation from another, it lies in the articulatory gesture rather than the acoustic feature. We may suppose the acoustic form of speech to be perceived not directly, but indirectly by reference to an articulatory gesture which is the same for several different acoustic values.

Bursts were studied synthetically in two other works (51), (52). The bursts were combined not with straight formants (steady-state vowels) as in the preceding experiment, but with consonant-vowel transition curves of the kind seen on spectrograms.

In (52) we have a comprehensive study of all appropriate combinations of three variables: F2 transitions, F3 transitions, and bursts. Only the American vowel [æ] was combined with these three variables. Twenty-six subjects were asked to identify as one of [b d g] 294 synthetic syllable patterns. Seven burst frequencies were combined with seven F2 transition curves to form 49 syllables. These 49 syllables were also synthesized with each of five different F3 transitions, making 245 further patterns. The F1 transition used was of a shape appropriate for voiced stops. These 294 syllable patterns with burst could then be compared with similar patterns without burst. The results were completely in agreement with the experiment described above (16), but go even further. High frequency bursts favor [d]-judgments, low frequency bursts (except the very lowest of the seven) favor [g]-judgments, first (going down the frequency scale) at the expense of non-burst [d], then at the expense of non-burst [b]. The best [g] have a burst just above the F2 transition (cf. (10)). Finally, the very lowest burst, which favors neither [d] nor [g], only slightly favors [b]. It must not be concluded from this that perception of labial place of articulation depends only on transitions -- we know that in the absence of transitions certain burst frequencies cause a quite clear labial place of articulation to be heard (16). We should simply conclude that in the perception of the labial place of articulation the part played by bursts is no doubt a less important one than that played by transitions.

In this over-all picture of the three consonants, burst effects are weak as compared to transition effects, and this in spite of the fact that the bursts in these experiments were probably more concentrated in frequency than they are in natural speech. We must not forget, however, that only the vowel [æ] was used. With a rounded vowel, the part played by bursts in perceiving place of articulation would no doubt be much greater.

Some information about bursts, from the viewpoint of manner of articulation, is to be found in a study of affricates (51). Since one of the distinctions between the class of affricates and the class of stops lies in the duration of the turbulent sound, it appears that a consonant is identified as a stop and no longer an affricate when the duration is less than 30 msec. The study was done with splicing as well as synthesis.

The first important splicing experiment (19) set about to verify the results of (16) in natural speech. Verification was positive. Tapes of the syllables [ki ka ku] were cut right after the burst, then the cut portions were spliced to the vowels [i a u] without transitions. Among other results, we find that the burst of [ka] combined with [i] was identified as [pi] by 93% of the subjects, with [u] as [pu] by 99% of the subjects. So just as in the synthetic experiment (16), we have the same burst heard as [k] or [p] depending on whether it is combined with [a] or [i].

We find more splicing with stops (and other consonants) in a study which literally put the commutation principle to the test (36), and the results are of the same sort as in (16). Each time the vowel following a constant consonantal element is changed, perception of the consonant changes too. Moreover, the results of any splicing commutation are predicable from what is now known about transitions.

A detailed analysis of the Danish stops [p t k b d g] before all Danish vowels (23), shows us the whole complicated picture an analysis typically furnishes with respect to intensities, durations and other energy concentrations over the frequency scale. How many of these features are distinctive? The test of synthetic reconversion into sound alone could tell us. (Thus the fact that a [p]-burst is not concentrated in frequency, but extends over almost the whole frequency range of the spectrum, does not necessarily mean that it plays no part in distinguishing place of articulation. Synthesis might show that certain portions -- varying with the vowels -- of this widespread noise could play such a part.) Generally speaking, the hypotheses presented at the conclusion of the study, on the role of bursts and transtions in the perception of stops, are not in agreement with the results later obtained

by synthesis (21), (26), (52). We should note, however, that the 'speculations' of (11) and (13) are not in agreement with the later results of (21), (26), (52) either. At that time neither the principle of the F2 locus nor the role of F3 transitions were yet known.

Two studies comparing final stops with and without release arrived at comparable results. In one (37), the informant recording said the stops without releasing them; in the other (45), releases were cut out after the recording was made. (The reader will realize that by omitting its release, a final consonant is deprived of burst as well as any embryonic transitions that may follow, so all that is left to tell the place of articulation are the implosive transitions that precede it.) The most interesting results are those that show, in both studies, that the consonants suffering most from lack of burst are [k g] with [u] (and with back rounded vowels in general). This result indicates that perception of place of articulation for [k g] with back rounded vowels depends a great deal on the burst and very little on the transitions. Moreover the fact was to be foreseen: the F2 transition of [k g] with [o u] moves not toward the velar locus but toward the labial locus; once [uk] is deprived of its burst or release, we should hear [up] or simply [u] -- and this is just what happened in the perceptual tests of (37), (45). In (37), [k] without burst or release is poorly perceived with a dark [l] or [r] as well; now, the first and second formants of dark [l] and [r] are very close to those of [o].

Study (45) goes on to investigate bursts by analysis and filtering. It is established (but rather unsurely) that separated from contexts are identifiable. Then, on the basis of the intensities and frequencies of the bursts of [p t k b d g] before and after six vowels representing the various articulatory positions, an attempt is made to discover by filtering, two pairs of binary features which would allow them to be identified in isolation. Would these two pair of distinctive features pass the test of synthesis? Whatever the answer may be, the article contains valuable information about the spectral properities of bursts and, roughly speaking, this information is in agreement with the results obtained by synthesis (16), (52). For [t d] the frequencies are high; for [p b] they are low; and for [k g] they lie in between, spread over a wide range of frequencies as they follow the F2 transition which varies from 3000 to 600 cps.

Stop transitions. So far no acoustic cues have been found outside of the first three formants. Let us call the transitions of these formants T1, T2, T3. Cues found for T2 and T3 correlate almost entirely with place of articulation (as does the burst frequency). The cues found for T1, on the contrary, correlate with manner of articulation. The former is thus a distinction among consonant classes; the latter a distinction between voiced and voiceless consonants.

T1. Through analysis of spectrograms it was noticed very early that a higher F1 (in frequency) corresponded to a more open vocal tract (2), (8). Applied to consonants, this should mean that the more open a voiced consonant is, the higher its T1 should begin. But no systematic investigation of this correlation has been made for consonants, and we shall have to bring together ideas scattered through several studies.

In studies on T2 (21) and (52), and on T3 (52), it was necessary to have a rather quick T1, beginning as low as possible, in order to obtain voiceless stops. (We do not know whether this actually corresponds to zero or 120 cps, the fundamental on SP.) To achieve nasal stops in (21), we apparently had to start T1 at the frequency of FN1 (the lowest formant of the nasal occlusion -- around 250 cps) and connect it vertically with the beginning of the adjoining vowel, which means that to the eye T1 looks straight and starts on a level with the adjoining vowel. In study (26) on the F1 locus for stops, straight F1 frequency variations combined with a curved F2 indicate that the lowest starting point for T1 is the best one for stops, and that as this point rises in frequency we move toward perception of the more open consonant classes. Examination of spectrograms of fricatives indicates, in general, a higher T1 starting point than is the case for stops. With initial liquids and semivowels, we find in (49) that T1 should begin rather high -- near 400 cps on the average -- to avoid any perception of a stop.

It can be seen that much remains to be done in specifying the part T1 plays in distinguishing among classes of consonants.

Transition speed and T1 duration also contribute to class distinctions. These two factors, varied together for T1 and T2 (33), enabled us to distinguish among the following three classes: vowels, semivowels, voiced stops. By changing the

duration and speed of T1 and T2, [u] went to [w], then to [b]; [i] went to [j] then to [g]; and if we had had French listeners we would no doubt have found that varying the same factors would have sent [y] to [ɥ] to [d]. The change from semivowel to consonant stands out better then the change from semivowel to vowel. The change from [b] to [w] occurs at a transition lasting some 40 msec.; from [g] to [j] at 50 to 60 msec.

The implosive and explosive forms of T1 and T2 are shown in (35) as contributing to the perception of syllable break (respectively, after and before the consonant). The investigation was carried out by synthesis.

Finally, we shall see below that certain dimensions of T1 appear to contribute to the distinction between [p t k] and [b d g] commonly called 'voiced-voiceless' (32).

T2. With a rather short duration (or quick speed), T2 are probably the strongest cues for distinguishing among places of articulation. Except for [k] followed by a rounded vowel, T2 works better than bursts. This is understandable in view of the fact that like vowel formants, they are much louder than the voiceless burst sounds. Stop T2 have an average duration of 50 msec.; this tends to be shorter in labials and longer in dentals followed by back vowels.

T2 dimensions contributing to place of articulation identification are: 1) its direction, said to be positive if T2 goes above the adjacent vowel F2, and negative if it goes below; and 2) the frequency difference between its beginning and the point at which it joins F2 of the vowel. (This dimension is usually given as a multiple of 120 cps in works from Haskins. Thus a -3 transition reaches a frequency 360 cps below the corresponding vowel formant.)

A comprehensive T2 study by synthesis (21) followed shortly upon study (16) on bursts. It included eleven T2 variations each combined with seven cardinal vowels [i e ɛ a ɔ o u], and this repeated for voiced, voiceless and nasal occlusives for a total of 231 artificial spectrogram patterns, identified by 33 subjects when converted into sound. No bursts were used in these patterns. Voice was achieved by having T1 start at zero (or 120) cps; voicelessness by cutting back on the beginning of T1; and nasality

by having T1 start on a level with F1 and by adding three nasal formants in the occlusion.

The extremely complex results indicate a different F2 not only for each place of articulation, but also for each vowel at each place of articulation. Furthermore, the nasal stop results were quite similar to those for voiced and voiceless stops.

Locus. Looking for a place of articulation invariant, we noticed that all T2 perceived as labial converged virtually on a low frequency (no matter what the vowel of the syllable was), that all T2 perceived as dental (or alveolar) on an intermediate frequency, and that all T2 perceived as velar (or palato-velar) on a high frequency. (This left a small area of ambiguity: before the back rounded vowels [ɔ o u], no T2 was clearly perceived as velar-- a problem since solved.) The name 'locus' was given to this virtual point of convergence of the transitions associated with one perceived place of articulation.

Specifying loci frequencies has been the object of extended synthetic research. A locus correlating with each place of articulation was determined, not by extrapolation from curved T2, but by varying straight formants, thus avoiding errors that curves could have introduced. By varying a straight F2 (zero T2) from the top to the bottom of the frequency scale and combined with a fixed T1 of a shape to produce a voiced stop, we obtained a [g] when the straight F2 was at 3000 cps; then lowering the frequency of the straight F2, [g] was lost and a [d] began to be heard which reached maximum perceptibility at 1800 cps; as the frequency of the straight F2 continued to drop, [d] faded out and a [b] began to be heard which reached its maximum at about 700 cps. Next, it was necessary to determine the duration T2 ends from their respective loci. This was done by repeated cuts on transitions starting from the same locus. We arrived at an average duration of 50 msec. So specified, the locus furnished a practical invariant for place of articulation. It has made it possible to define a stop transition without reference to any particular vowel combined with it. A stop T2 can be described as having a duration of about 50 msec. and extending towards a locus of place of articulation, perceptible by a virtual line that would reach this locus in

another 50 msec. The frequency differences among the real transition ends perceived as one place of articulation are obviously due to articulatory anticipation of the adjacent vowel. (The articulatory correlation of the three loci, as well as the non-applicability of the velar locus to rounded vowels, has been clearly established on X-ray moving pictures, but we do not have space here to deal with the physiological correlates of acoustic cues.)

Research in progress indicates that velar stops before rounded vowels in natural speech have burst frequency as their main acoustic cue for place of articulation. If, in natural speech, the T2 in a syllable such as [go] does not move toward a high velar locus, the reason is that rounding keeps the frequency low at the beginning of the transition. In artificial speech, however, a [go] can be obtained without burst simply by painting a positive T2 directed towards a velar locus at 3000 cps and long enough to extend beyond the dental locus at 1800 cps.

In (22) there will be found a presentation of the locus concept as understood personally by a visitor to Haskins Laboratories. We must state here, however, that the hypothesis attributing to the locus the frequency of the vocal resonator before the release of the consonant is no longer accepted.

An electronic analog of the vocal tract attempted to verify the locus concept (42). This analog simulated three articulatory variables with which it synthetically produced sustained sounds (of a vowel-like nature): point of tongue constriction, degree of tract constriction, and degree and extension of labial constriction. Thus it does not produce consonants, but may nonetheless be used to study consonants by noting the successive effects of each adjustment on spectrograms. The results so obtained by [b d g] transitions are approximately in agreement with the respective loci provided it is understood that throughout the study the term 'locus' has been confused with 'transition beginning'. There is nothing unusual in the fact that transition beginnings (and not loci) vary in anticipation of the vowel — an anticipation that was recognized, but assumed (in (42)) to be much sharper than it actually is, especially for [b], according to X-ray films of articulated [b d g].

Before leaving the subject, we shall note that the notion of locus need not be restricted to stop transitions only, but may, perhaps, be extended to all consonant transitions. It would seem that [f] has the same locus as [p], [s] as [t], etc., or at the very least, something very close to this.

In (52), already mentioned in connection with bursts, T2 variations were very carefully examined in combination with T3 variations, or burst, or both. Before the vowel [æ] and with a constant T1 of a shape to give voiced stops, seven variations in T2 were studied: -6, -4, -2, 0, +2, +4, +6. These are the same dimensions as in (21) with the odd transitions left out for simplicity's sake. The results entirely confirm those of (21). By themselves, the transitions of the first two formants (without burst or T3) are enough to distinguish among [b d g]. [b] depends most on T2, [d] least -- [d] depends on T3 much more than the other two do. [b]-judgments are about 100% at -6, -4 and -2, then fall sharply. At zero [d] reaches almost 90% and at +2 almost 100%, then [d] falls suddenly and gives way to [g] which reaches 95% at +4 and 100% at +6.

The results of T2 studies by analysis agree perfectly with those by synthesis, but are naturally less precise. It is just because spectrograms are so difficult to read, especially with respect to transitions, that synthesis can be of such help.

Four T2 studies by analysis are worth noting.

There are many remarks about T2 in (1) involving the notion of 'hub' which may be taken as a forerunner of locus even though the two notions actually differ considerably.

Analyses of T2 in (3), previously mentioned in the introduction, shrewdly foresaw the role that synthesis was to confirm and make specific.

Good T2 analyses are to be found for Danish stops in (23) indicating that the locus of Danish labials is not so low as for Anglo-American and Romance labials.

Finally, T2 analyses in (23) entirely confirm the results of (21), (26), (52) obtained through synthesis.

T3. Nothing has yet been published bearing especially on T3, but we may say that the results of a detailed study in press are in agreement with those of (52) which we shall summarize below, while pointing out that they apply only to the vowel [æ].

The T3 question is quite a bit simpler than that of T2 since F3 lies at approximately the same frequency for all vowels (it is slightly higher for a good cardinal [i]). In the main, we may say that T3 is positive for dentals, and negative for labials and velars. Among the negative T3, all contribute toward labial more than toward velar perception, and the higher among them contribute more toward velars than the lower. Perception of dental place of articulation owes a great deal to T3 (with certain vowels, perhaps more than to T2); perception of labial place of articulation less; perception of velar place of articulation still less. (For labials, T2 dominate, for velars either T2 or a burst.)

In (52) in which, on the one hand, five variations of T3 (-4, -2, 0, +2, +4), and, on the other, seven burst frequencies are combined with each of seven T2 variations, we have an opportunity to compare the effects of T3 with those of bursts. Generally speaking, for [d] and [b] the contribution of T3 is distinctly greater than that of bursts; for [g] the reverse is true: the contribution of bursts is greatest.

According to analytic comments in (23), T3 for Danish stops differ slightly from the above. The analyses of (45) are approximately in agreement with the above. (However, the difficulty in making out T3 curves on spectrograms is well known.)

V. Fricatives.

It has been established in (51) that fricatives, as a consonant class, are distinguished in part from affricates and stops by the duration of the noise (turbulent sound) as well as by the rate at which the initial intensity of the noise rises. Noise duration is relatively greater and rate of intensity rise relatively slower for fricatives (see below for affricates). The part played by transition speed as a class distinction has not been systematically studied. It is

certain, for instance, that between the rapid transitions of a [b] and the slow transitions of a [w], there exist intermediate transitions corresponding to [v], and it will be necessary to determine the respective roles played by T1, T2, and T3. In this same distinction of fricatives as a class, the part played by the frequency of the beginning of T1 is also worth studying.

Since nothing has appeared concerning the cues that distinguish among the various fricatives, we shall provide only general hypothetical indications, based in part on a communication abstracted in JAS 26:952. According to synthesis, these cues are found in the upper transitions (T2, T3) and in the friction noises.

The parts played by T2 and T3 in place of articulation perception are no doubt comparable to those for stops, but a systematic study has not been undertaken. T2 and T3 should thus be describable in terms of loci correlative with place of articulation; however, it is to be anticipated that the part they play will be less important than in stops inasmuch as friction noises are louder than bursts.

We should perhaps divide the fricatives into three subclasses distinguished among themselves by friction intensity, spread of friction frequency, and transitions: [sʃ] probably have great intensity and intermediate spread; [θf] weak intensity and great spread (almost the whole spectral frequency on Kay spectrograms); [çx] medium intensity and a narrow spread. The part played by transitions should not be negligible, since in synthesis the same friction (ambiguous towards 3500 cps) may be heard as [s] or [ç] depending on whether it is linked to the vowel by T3 (positive -- dental) or by T2 (positive -- palato-velar).

Within these three classes, distinctions are simple: [s] is distinghished from [ʃ] mainly by friction frequency ([s] has a lower limit of about 3500 cps, [ʃ] of about 2000 cps); [θ] is distinguished from [f] mainly by transitions (which are more or less on the dental locus for [θ] and the labial locus for [f]); [ç] is distinguished from [x] by both friction frequency and transitions.

The fricative [h] is probably characterized by a short turbulent sound at the frequency of F2 (and perhaps F3) of the contiguous vowel -- hence by the absence of transitions and of F1. Since the glottis is wide open, the friction distinctive of [h] is, no doubt, that resonating in the cavity above the point of vowel constriction. This is in contradistinction to whispered vowels which may be supposed to resonate throughout all the cavities above the glottis, the point of constriction producing the turbulence being the vocal cords themselves. This is why though voiceless, whispered vowels still show as F1.

One work by analysis and filtering should be cited, (31). Here are studied spectra of isolated [f s ʃ v z ʒ] friction in all positions and as spoken by different subjects. The results confirm what was said above concerning the distinction between [s] and [ʃ] as being a difference in friction frequency. A new factor is introduced: [f] often, but not always, displays a very high energy concentration at about 8000 cps. This should be verified by the first synthesizer to reach that frequency.

VI. Affricates.

In study (51), already cited, affricates were examined from the standpoint of manner of articulation. The thing that distinguishes them from fricatives and stops lies in the turbulent sound, besides which there is the fact that affricates, like stops, have an interruption (occlusion) which fricatives do not.

Two acoustic cues were found: noise duration and rate of increase in noise intensity (measured by the duration over which the intensity rises from the beginning of the noise -- let us call it 'rise-time').

Roughly speaking, affricate noise (after the occlusion) as compared with fricative noise is shorter in both total duration and rise-time. Compared with stops, affricate noise is longer is duration. Using average rise-time values, a voiceless fricative is perceived when the total noise lasts at least 110 msec., a voiceless affricate when the total noise lasts less then 50 msec., and a voiceless stop when the total noise lasts <u>at the very most</u> 30 msec.

Affricate place of articulation cues have not yet received systematic study, but it is clear that cues similar to those for fricatives and stops will be found: first in the transitions (loci); then in the frequencies of the friction noise. It is likely that noise intensity and frequency spread are also tied in with the question.

VII. Nasal stops.

We are classing the nasal consonants [m n ŋ] among the stops because they share with oral stops both T2-T3 form (speed) and direction. In addition, closure of the nasal passages at the nostrils does not prevent production of [m n ŋ]; an opening to the outside at the nostrils -- small at best -- is thus non-distinctive. What is distinctive is the occlusion in the vocal tract and the connection of nasal with buccal cavities over the lowered velum.

The acoustic cues for manner as well as place of articulation can be seen rather clearly in (21), a study carried out by synthesis.

Manner. Nasal stops are distinguished from oral stops: 1) by the shape of T1 which seems to begin on a level with FN1 (around 250 cps) and rises vertically to the level of the contiguous vowel instead of starting from zero or 120 cps, as it does in the voiced oral stops; and 2) by the nasal formants during the occlusion -- which replace the complete silence of voiceless oral stops, or the very low tone of voiced oral stops (this corresponds to the fundamental and occasionally somewhat to the second harmonic heard through the mouth and pharynx walls). In the experiments of (21), the nasal formants (during closure) were the same for all three consonants [m n ŋ] since exploratory work had indicated that they played only a weak part in place of articulation distinctions. They were located at 240, 1020 and 2460 cps. The higher two were extremely weak in intensity (about 15 db less than for a normal vowel at the same frequency) and contributed little to the nasality of the consonant. The first nasal formant at 240 cps was only slightly weaker than that of a normal vowel (some 6 db less) and had a strong nasal effect on perception. Hence it appears that nasal manner in consonants depends on the shape of T1, on a steady-state formant at about 250 cps, and on T2 and T3 shapes similar to

those of oral stops at the same place of articulation. (The importance of the nasal formant at about 250 cps was mentioned for the first time in (20).)

Place of articulation. 1) The part played by transitions, defined with respect to loci correlative with place of articulation, is the same as for oral stops. It is a strong factor but not the only one. 2) The frequencies of the nasal formants (during closure) above the one at 250 cps all play something of a part in place of articulation perception -- weak, but real. Research coming after (21) has indicated that labial place of articulation perception is favored by the presence of a weak F2 (during closure) located between 1000 and 1500 cps, and by the absence or weakness of F3; perception of dental and velar place of articulation, by the presence of an F3 at about 2300 cps (in addition to the F2). No clear cue distinguishing dentals from velars has been found for nasal formants.

The role of nasal formants (during closure) has been studied a good deal by splicing. In (39) nasal occlusions of [m n ŋ] were interchanged before their various transitions as well as spliced in before the transitions following the bursts of [b d g]. The results confirm the fact that place of articulation cues lie almost entirely in the transitions; the nasal formants play an almost negligible role in this respect initially and a slightly more apparent though still very weak one finally.

The electronic analog of the vocal tract has also been used to produce [m n ŋ] synthetically (50). Place of articulation distinctions were perceived (81%, 61%, 62% respectively for [m n ŋ]) -- better perceived than segments of human nasal formants in (39): 96%, 36%, 12%. It should be said that in (50) judgments were made by nine trained subjects, in (39) by 50 phonetically untrained subjects. Spectral analysis of the nasal sounds produced by the analog confirms the importance of F2 in distinguishing [m] from the other two, and indicates the possibility of distinguishing [n] from [ŋ] by a formant above 3000 cps.

VIII. Liquids and semivowels.

Inasmuch as American [w j r l] have certain spectral qualities in common reflecting their articulatory degree of opening (which averages greater than for stops, affricates and fricatives), they were studied, as a group, in initial position (49). The [r] in question here is a retroflex apical and palatal continuant without trill.

Manner. Distinguishing them acoustically from the other consonants, these four appear to have the following in common: 1) A steady-state F1 of relatively high frequency (near 400 cps on the average); this distinguishes them from the nasals in which the lowest formant never exceeds 250 cps. 2) Steady-state formants above F1 of greater intensity than those in nasals, yet weaker than in vowels. 3) Transition continuous with the steady-state formants (nasal transitions may be discontinuous with the nasal formants). 4) Relatively slow transitions (on the average, about 100 msec., whereas stop transitions average about 50 msec.).

Place. These four consonants are distinguished from each other by the frequencies of the steady-state formants above F1, and by the transitions leading from the steady-state portion to the adjacent vowel. These transitions may be described in terms of a locus which would be their virtual point of convergence relative to the perception of a particular place of articulation. 1) [w] is distinguished from [r l] and [j] by the F2 loci: low for [w] (about 700 cps); intermediate for [r l] (about 1100 cps for a palatal [r], 1300 cps for an alveolar [l]; high for [j] (around 2700 cps). 2) [r] and [l] are distinguished from each other by T3 loci which are relatively lower for [r] (around 1500 cps) and higher for [l] (around 2500 cps). T3 has no special effect on [w] or [j]. To these can be added three minor differences which no doubt also contribute in distinguishing the four consonants: 3) A transition duration of 100 msec. is acceptable for all four but a slightly shorter duration favors [l] over [r] and a slightly longer duration [r] over [l]. 4) Steady-state formants are less indispensable for [j w] than for [l r]; and the average steady-state duration of [j w] (30 msec.) is shorter than for [l r] (60 msec.). 5) [j] is improved by friction at an appropriate frequency, which is not the case for the other three. This tends to bring [j] in line with the

fricatives, with which it indeed belongs if it is simply the voiced counterpart of [ʂ].

IX. Syllabic consonants.

No systematic study has yet appeared. Work in progress by synthesis is investigating the effects of reduction in F2 intensity and of shape of the preceding transitions as manner of articulation cues that would distinguish the consonants [l r m n ŋ] as syllabics from vowels with formants at almost the same frequencies. For instance, American [mɛdl] "medal" is distinguished from [mɛdo] "meadow" by these two cues.

The same investigation is studying formant frequencies as cues distinguishing the various syllabic consonants among themselves.

X. 'Voiced-voiceless'.

As we continue to discover new acoustic factors for what is called 'voicing' or 'voicelessness' in consonants, the presence of a fundamental (first harmonic in the spectrum of vocal cord vibrations), which was the basis of these terms, assumes a smaller place and we prudently come to replace the term 'voice-voiceless distinction' by '[p-b] type distinction'! According to (23), the [p b] distinction in Danish cannot depend on the presence or absence of vocal cord vibrations since the cords vibrate for neither. Here, however, we shall continue to use the terminology 'voiced-voiceless' -- arbitrarily, or with a perceptual meaning.

1) The sign of voicing usually present on spectrograms during the occlusion of voiced oral stops is called a 'voice bar' in (1). With wide-band filtering at 300 cps, it is the displayed fundamental to which a more or less intense second harmonic has been added, depending on the subject. The contribution of this voice bar to the perception of voicing has been amply confirmed by synthesis. Thus in (52) to make the syllables [ba da ga] perceived as 'voiced', they were preceded by a line painted in the first harmonic channel which caused the fundamental to be heard for 60 msec. In synthesis it is possible, however, to make voicing heard by several other

means if the fundamental is absent. This can also be done with splicing. In (48), when the occlusion of [b] in "ruby" (containing a fundamental) was replaced with a piece of unused magnetic tape of the same duration, "ruby" rather than "rupee" was still heard (by both Anglo-american and Romance ears). To make perceived voicing disappear, it is not enough to leave out the fundamental, further changes are necessary too -- for instance, lengthening the occlusive silence (see below under 7)).

2) With fricatives -- which are articulated with something of a buccal opening -- voicing may include, in addition to the voice bar, a low intensity, neutral vowel paralling the friction. The effective contribution of this factor has been confirmed by synthesis.

3) The simple presence of T1 seems to contribute very strongly to the perception of voicing and its absence to voicelessness. As early as (21), voiceless stops were produced synthetically with a very reduced T1. Work in progress is studying partial suppression of T1 as a factor for voicelessness in intial, final, and both implosive and explosive intervocalic position.

4) Aspiration, that is, acoustically, omission of T1 and presence of turbulent (non-harmonic) instead of periodic (harmonic) sound during the first 50 or 60 msec. of T2 and T3, contributes distinctly to the perception of voicelessness. It is to be noted that if T1 also has a turbulent sound, there is no longer a voiceless effect, no doubt because the result is then not a period of aspiration, but rather of whispered vowel. Aspiration is thus apparently similar to the consonant [h], and like [h] is probably articulated with an open glottis so that only cavity resonances above the vowel constriction have an effect on perception; on the contrary, since a whispered vowel has all the formants of a non-whispered vowel, its constriction should be at the glottis and all cavities above the glottis should have an effect on perception.

5) T1 transition speed was studied in (32). A T1 duration of 20 msec. or slightly less caused stops to be heard as voiceless, and a duration of 50 msec. or slightly more made them heard as voiced.

6) Presence of a T3 contributes slightly to voicelessness. Since the artificial spectrographic patterns of [b d g] used to study T2 variations in (21) had no T3 it was not necessary to have a fundamental; conversely, because the patterns used in (52) did have a T3, a fundamental had to be added.

7) The relative duration of intervocalic consonants (or final consonants with vocalic release) is a very strong factor. In (48), splicing on natural speech varied the duration of stop occlusions: "ruby" went to "rupee" between 60 and 100 msec. without a fundamental and between 80 and 120 msec. with a fundamental. In (27) it is possible to see among other things the effect of friction duration. With a constant vowel duration in a syllable of the type VC, judgments went from 100% [juz] to 70% [jus] when the duration of friction was increased to 50 to 250 msec.

8) Relative intensity of noise is also a factor in voicelessness. In (16), syllables synthesized by combining burst and vowel without transition were heard as voiceless. In (52), the presence of bursts in syllables that were meant to be voiced to American ears required a fundamental followed by a relatively slow T1. In (27), variation in friction noise intensity had some influence on the perception of voicing, though it was very slight.

9) The influence of vowel duration relative to duration of following consonantal friction was the main subject of (27). The technique of synthesis made it possible to combine several vowel durations with several friction durations while the transition factors were held constant. The results were clear (though they apply only to fricatives): the longer the duration of the vowel, the greater the chances the consonant will be perceived as voiced. Thus for a certain constant friction duration, judgments went from 100% for [jus] to 65% for [juz] as the vowel went from 50 to 200 msec.

10) We may mention, finally, a study in which consonants were identified after distortion by filtering and the addition of noise (24). The distinctions 'voiced-voiceless' and 'nasal-oral' stood up much better than place of articulation distinctions.

XI. Oral vowels.

A systematic study of oral vowels by synthesis was the subject of two articles: (7), (15). In (7) are to be found the formant frequencies necessary for synthesizing the sixteen main cardinal vowels with just two formants. These sixteen vowels were chosen by ear from 235 appropriate combinations of F2 variations with a constant F1 and of F1 variations with a constant F2. On the vowel diagram formed by placing F1 and F2 frequency values on the ordinate and abscissa, it is rather curious to see that [æ] is aligned with [i e ɛ], while [a] is aligned with [yøœ].

In (15) the following basic facts were established:

1) In synthesis, two formants suffice to clearly characterize vowel color, even nasal vowel color.

2) Human vowels, however, are often identified by three formants. Put another way, in human speech F3 does play a part in identifying certain vowels, in fact all vowels with a rather high F2, that is, rather close F2 and F3. These are usually front vowels.

3) In perception, there is a relative equivalence between two formants located close together and a single formant at the average of their frequencies. Thus back vowels can be identified by a single formant at a frequency intermediate between F1 and F2 (F3 is very weak for back vowels and contributes little more than 'naturalness', since it is very far from F2 in frequency). Similarly, when F2 and F3 are close together, as in front vowels, perception of the two together is almost equivalent to perception of a single formant at a frequency between them.

4) For synthetic 'front' vowels with two formants, F2 frequency is between the F2 and F3 frequencies of natural vowels with the same color. Thus, natural [i] color with formants at 250, 2500 and 3000 cps can be synthesized with two formants at about 250 and 2750 cps. (Of course it is synthesized even better with three formants at the same frequencies as the natural vowel!)

5) Formants above 3000 cps for [i] and above 2500 for other vowels play practically no part in the linguistic characterization of vowels. Their main contribution is in characterizing the color of individual voices.

6) Individual variations in formant intensities have two different effects depending on whether the formants are close together or far apart in frequency. If the two formants whose relative intensity is being varied are far apart from each other, the vowel color becomes weaker as the intensity difference increases until at last all linguistic identity fades away and is replaced by a musical identity (generally a dissonances of contiguous sounds); if the two formants are close by each other in frequency, the vowel changes color becoming more and more in perceptual effect like a vowel with only the one formant that was kept at its original intensity.

7) When F1 alone is reduced in intensity, the color change is perceived as tending to the nasal. (The first indication that the acoustic cue for vowel nasality lay in the weaker intensity of F1 is to be found in (15).)

Study (38) of American vowels by filtering out frequencies above 670 cps (in an attempt to omit formants above F1) concluded that the distinctive acoustic cues for these vowels are F1, F2 and duration (in two steps). Duration does seem indispensable for distinguishing two vowels like [e] and [ɪ] which have almost the same formant frequencies. (We must note, however, that the role of F3 was not included in this study.)

Vowel formant frequencies have been studied analytically for several languages. For instance, American vowels in (1), (3), (14) and especially (10); French vowels in (2) and (8); Danish vowels in (23); Swedish vowels in (35); Polish vowels in (43); Japanese vowels in (44).

The electronic analogs described in (5), (25), (29) have produced good synthetic vowels and so contributed not only to specifying articulatory and acoustic correlations but also to verifying the linguistic value of the results obtained by synthesis.

Finally, a theory proposed in (3) has just been confirmed by synthesis in (47). Linguistic identification of vowels probably does not depend entirely on the absolute frequencies of the formants, but on the frequencies relative to a speaker's total formant structure and this may vary slightly from one person to another, as was already indicated by divergences among men, women and children (increasingly higher frequencies in that order) established in (10).

XII. Nasal vowels.

The cues for vowel nasality were discovered through synthetic techniques, (15) and especially (20), and later confirmed by the electronic analog of the vocal tract (40) and by an analysis of nasal Japanese vowels (41).

The first cue, the only one that can actually change an oral into a nasal vowel independently of other cues, is a reduction in F1 intensity. To get synthetic French nasal vowels, a reduction of 12 to 15 db relative to the normal F1 intensity of the oral vowel is necessary.

The second cue (second in importance) is a formant at about 250 cps, which we shall call FN1 (first nasal formant). This is in all likelihood the formant which occupies the first place in the occlusion of nasal consonants. We know that it makes a considerable contribution to vocalic nasality because when FN1 is present, only a small decrease in F1 intensity is required to cause the vowel to be identified as nasal. By itself, however, FN1 nasalizes a vowel only slightly.

The remaining cues, not always visible on spectrograms, are quite weak and have almost negligible perceptual effects. They are, chiefly, a formant at around 1000 cps and another around 2000 cps.

Hypothetically, the reduction in F1 intensity may be attributed either to the great damping action of the fibrous cavities of the nose which apparently act only on low frequency waves on a level with F1 (40); or to anti-resonances which would suppress some of the F1 tones (41). Since the nasal cavities have a rather constant volume, it would be necessary for the buccal cavities, especially the pharyngeal cavity,

to accommodate their volumes to make the F1 frequencies accord with the anti-resonances.

The damping hypothesis is supported by the fact that efforts on the part of the analog to produce nasal vowels originally failed. Simply adding a third cavity only produced an additional formant at about 1000 cps, and the vowels were not perceptibly nasalized by it. In order to produce a vowel that could be heard as nasal and would display an F1 of very low intensity, the nasal cavity had to be considerably damped.

On the other hand, the anti-resonance hypothesis, which assumes accommodation of the cavities, is supported by the fact that in all nasal vowels the weak F1 tends towards the same frequency of about 500 cps (41). This would explain the evolution of all nasal vowels toward the same degree of aperture (i.e., half-open): Over the course of the history of French, [in yn un] became approximately [ɛ̃ œ̃ ɔ̃] and [ã] became a vowel closer to [ɔ̃] than to [ã].

XIII. Prosodic features.

A beginning has been made on studying the prosodic elements of speech such as stress, rhythm, and intonation, by synthesis. Important results are to be expected since the objective factors of duration, intensity and frequency can be varied independently and the results of separate and combined variations judged subjectively by the ear.

A pioneer study has attempted to compare the effects of duration and intensity variations (without including frequency as yet) in perception of the location of English stress. Words like "object" were used, which are taken as substantives when stress lies on the first syllable and as verbs when it lies on the second. When the subjects identified the word as a substantive it could be allowed they had preceived stress on the first syllable, when as a verb, on the second. The results: both factors contributed to the perception of stress location, but duration more so than intensity (28).

Studies taking the frequency factor into account are under way.

XIV. Theoretical background.

In conclusion we shall cite the more general studies between 1947 and 1957 that have summarized and ventured theoretical statements concerning acoustic cues in speech. In chronological order (1) comes first with its wealth of fine spectrograms (though they usually represent slow speech). Next we have (3) with its insights programming the direction of present research. Then there is (9) proposing distinctive features, defined as binary oppositions, from which the acoustic cues follow. Since this 'preliminary' work relied on too hasty an analysis of spectra, it will have to be revised when acoustic and physiological research have determined what the cues actually are. In (11), (12), (13) it is possible to follow the rapid development of synthetic research; however, the hypotheses, based on partial results, are themselves partly faulty since they precede the discovery of the F2 locus concept and of F3 transitions. Finally, in (46), theoretical construction proceeds to a more advanced level. Here we find views based on ten years of investigation into the respective roles played by acoustic and articulatory phenomena in speech perception. In particular, the suggestion is made that acoustic waves are preceived not directly, but rather indirectly through reference to the articulatory gestures.

XV. Conclusion.

Though the progress made over the past decade is impressive, we are still far from possessing a complete and reliable picture of the acoustic cues in speech. Still remaining to be done: 1) Thorough studies on factors as yet barely touched upon; and 2) research completing the work already done on the better known factors, examining them now in a broader range of contexts. Investigation on various cues will be pursued simultaneously in each of the different laboratories, but progress will nonetheless be slow. In the complete investigation of a single cue, several years usually elapse between the time the cue is first isolated and the time when the definitive tests have been analysed. We shall submit another report as soon as the next significant advance has been achieved.

XVI. Bibliography.

1. POTTER, R. K.; KOPP, G. A. and GREEN, H. C.: Visible Speech (Van Nostrand, New York 1947).
2. DELATTRE, P.: Un triangle acoustique des voyelles orales du français. French Rev. 21: 477-484 (1948).
3. JOOS, M.: Acoustic Phonetics (Waverly Press, Baltimore 1948).
4. COOPER, F.: Spectrum Analysis. J. Acoust. Soc. Amer. 22: 761-762 (1950).
5. DUNN, H.: Calculation of Vowel Resonances, and an Electrical Vocal Tract. JAS. 22: 740-753 (1950).
6. COOPER, F. S.; LIBERMAN, A. M. and BORST, J. M.: The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech. Proc. nat. Acad. Sci., Wash. 37: 318-325 (1951).
7. DELATTRE, P.; LIBERMAN, A. M. and COOPER, F. S.: Voyelles synthétiques à deux formants et voyelles cardinales. Maître Phonet. 96: 30-37 (1951).
8. DELATTRE, P.: The Physiological Interpretation of Sound Spectrograms. Publ. Mod. Lang. Assoc. Amer. 66: 864-876 (1951).
9. JACOBSON, R.; FANT, C. and HALLE, M.: Preliminaries to Speech Analysis, the Distinctive Features and their Correlates (Acoustics Laboratories of MIT, Cambridge, Mass. 1952).
10. PETERSON, G. and BARNEY, H.: Control Methods Used in a Study of the Vowels. JAS. 24: 175-185 (1952).
11. DELATTRE, P.; LIBERMAN, A. M.; COOPER, F. S. and GERSTMAN, L.: Speech Synthesis as a Research Technique. Proc. 7th Int. Congr. Ling. London 1952; pp. 555-561 (1952).
12. DELATTRE, P.; COOPER, F. S. and LIBERMAN, A. M.: Some Suggestions for Language Teaching Methods Arising from Research on the Acoustic Analysis and Synthesis of Speech. Rep. 3rd. ann. Round Table Meet. Linguist. Lang. Teach. 2: 31-47 (1952).
13. COOPER, F. S.; DELATTRE, P.; LIBERMAN, A. M.; BORST, J. M. and GERSTMAN, L.: Some Experiments on the Perception of Synthetic Speech Sounds, JAS. 24: 597-606 (1952).
14. PETERSON, G.: Information-Bearing Elements of Speech. JAS. 24: 629-636 (1952).

15. DELATTRE, P.; LIBERMAN, A. M.; COOPER, F. S. and GERSTMAN, L.: An Experimental Study of the Acoustic Determinants of Vowel Color; Observations on One- and Two-Formant Vowels Synthesized from Spectrographic Patterns. Word 8: 195-211 (1952).
16. LIBERMAN, A. M.; DELATTRE, P. and COOPER, F. S.: The Role of Selected Stimulus-Variables in the Perception of the Unvoiced Stop Consonants. Amer. J. Psychol. 65: 497-517 (1952).
17. DURAND, M.: De la formation des voyelles nasales. Studia Linguist. 7: 33-53 (1953).
18. COOPER, F. S.: Some Instrumental Aids to Research on Speech. Report of the Fourth Annual Round Table MLLT. 3: 46-54 (1953).
19. SCHATZ, C.: The Role of Context in the Perception of Stops. Language 30: 47-57 (1954).
20. DELATTRE, P.: Les attributs acoustiques de la nasalité vocalique et consonantique. Studia Linguist. 8: 103-109 (1954).
21. LIBERMAN, A. M.; DELATTRE, P.; COOPER, F. S. and GERSTMAN, L.: The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants. Psychol. Monogr. 379: 1-14 (1954).
22. DURAND, M.: La perception des consonnes occlusives: problèmes de palatalisation et de changements consonantiques. Studia Linguist. 8: 110-123 (1954).
23. FISCHER-JØRGENSEN, E.: Acoustic Analysis of Stop Consonants. Misc. Phonet. 2: 42-59 (1954).
24. MILLER, G. and NICELY, P.: Analysis of Perceptual Confusions Among some English Consonants. JAS. 27: 338-353 (1955).
25. STEVENS, K. and HOUSE, A.: Development of a Quantitative Description of Vowel Articulation. JAS. 27: 484-494 (1955).
26. DELATTRE, P.; LIBERMAN, A. M. and COOPER, F. S.: Acoustic Loci and Transitional Cues for Consonants. JAS. 27: 769-774 (1955).
27. DENES, P.: Effect of Duration on the Perception of Voicing. JAS. 27: 761-764 (1955).
28. FRY, D.: Duration and Intensity as Physical Correlates of Linguistic Stress. JAS. 27: 765-768 (1955).
29. HOUSE, A. and STEVENS, K.: Auditory Testing of a Simplified Description of Vowel Articulation. JAS. 27: 882-887 (1955).

30. MALMBERG, B.: The Phonetic Basis for Syllable Division. Studia Linguist. 9: 80-87 (1955).
31. HUGHES, G. and HALLE, M.: Spectral Properties of Fricative Consonants. JAS. 28: 303-310 (1956).
32. DURAND, M.: De la perception des consonnes occlusives, questions de sonorité. Word 12: 15-34 (1956).
33. LIBERMAN, A. M.; DELATTRE, P., GERSTMAN, L. and COOPER, F. S.: Tempo of Frequency Change as a Cue for Distinguishing Classes of Speech Sounds. J. exp. Psychol. 52: 127-138 (1956).
34. BORST, J.: The Use of Spectrograms for Speech Analysis and Synthesis. J. Audio Engng. Soc. 4: 14-23 (1956).
35. MALMBERG, B.: Distinctive Features of Swedish Vowels; Some Instrumental and Structural Data. For R. Jacobson, pp. 316-321 (1956).
36. FISCHER-JØRGENSEN, E.: The Commutation Test and its Application to Phonemic Analysis. For R. Jacobson, pp. 140-151 (1956).
37. HOUSEHOLDER, F.: Unreleased ptk in American English. For. J. Jacobson, pp. 235-244 (1956).
38. MILLER, G.: The Perception of Speech. For R. Jacobson, pp. 353-360 (1956).
39. MALECOT, A.: Acoustic Cues for Nasal Consonants; an Experimental Study Involving a Tape-Splicing Technique. Language 32: 274-284 (1956).
40. HOUSE, A. and STEVENS, K.: Analog Studies of the Nasalization of Vowels. J. Speech Dis. 21: 218-232 (1956).
41. HATTORI, S.; YAMAMOTO, K. and FUJIMURA, O.: Nasalization of Vowels and Nasals. Rep. Kobayashi sci. Inst. 6: 226-235 (1956).
42. STEVENS, K. and HOUSE. A.: Studies of Formant Transitions Using a Vocal-Tract Analog. JAS. 28: 578-585 (1956).
43. JASSEM, W.: The Formants of Sustained Polish Vowels; A Preliminary Study. Study of Sounds; pp. 335-349 (Chiyoda, Tokio 1957).
44. HATTORI, S.; YAMAMOTO, K., KOHASI, Y and FUJIMURA, O.: Vowels of Japanese. Rep. Kobayashi sci. Inst. 7: 69-79 (1957).
45. HALLE, M.; HUGHES, G. and RADLEY, J. P.: Acoustic Properties of Stop Consonants. JAS. 29: 107-116 (1957).
46. LIBERMAN, A. M.: Some Results of Research on Speech Perception. JAS. 29: 117-123 (1957).
47. LADEFOGED, P. and BROADBENT, D.: Information Conveyed by Vowels. JAS. 29: 98-104 (1957).

48. LISKER, L.: Closure Duration and the Intervocalic Voiced-Voiceless Distinction in English. Languages 33: 42-49 (1957).
49. O'CONNOR, J. D.; GERSTMAN, L.; LIBERMAN, A. M.; DELATTRE, P. and COOPER, F. S.: Acoustic Cues for the Perception of Initial /wjrl/ in English. Word 13: 24-44 (1957).
50. HOUSE, H.: Analog Studies of Nasal Consonants. J. Speech Dis. 22: 190-204 (1957).
51. GERSTMAN, L.: Cues for Distinguishing among Fricatives, Affricate, and Stop Consonants. Diss. (New York University 1957) (Research done at Haskins Laboratories, New York).
52. HOFFMANN, H.: A Study of Some Cues in the Perception of the Voiced Stop Consonants. Diss. (University of Connecticut 1957) (Research done at Haskins Laboratories, New York).