SPEECH SYNTHESIS AS A RESEARCH TECHNIQUE

PIERRE DELATTRE FRANKLIN S. COOPER
ALVIN M. LIBERMAN LOUIS J. GERSTMAN
HASKINS LABORATORIES

Reprinted from
Proceedings of the VIIth International Congress of Linguists (1952)
LONDON 1956

1. Introduction

There is at present a renewed interest in the contributions to linguistic science which can come from acoustic studies. This is due in part to new instruments, developed during and after the way, which enable us to deal with the acoustic aspects of speech with a new facility — so much so that this new area has been termed "Acoustic Phonetics" by Martin Joos in his pioneering monograph.[1] The more popular label of "Visible Speech" has been applied by the Bell Telephone Laboratories to the best-known of these newer approaches.[2] And indeed the instrument which they developed does produce a "picture" from the speech wave. By studying this picture — this sound spectrogram — it is possible to see, more clearly than before, the acoustic patterns that correspond to the various linguistic units and to determine some of the relations between articulatory movements and sound.

However, we shall not deal further in the present paper with this essentially descriptive approach; rather we should like to discuss a different and newer research method which uses these spectrographic pictures as a basis for modifying or even synthesizing speech sounds, and thus enables an investigator to determine experimentally which aspects of the speech-wave are important to the listener; in short, to study speech by creating a variety of sounds and choosing those that are heard as speech.

2. Technique of Synthesis

The way in which these speech sounds are generated is just the reverse of the familiar decomposition of speech into its harmonic components. It is, if you will, a recomposition of those same harmonics to produce synthetic speech. Perhaps the easiest way to explain this procedure would be to recompose some typical syllables.

[At this point in the lecture, sound recordings and slides were presented to demonstrate how the syllables MI DO LA can be synthesized. See Appendix and diagrams in pocket.]

[1] M. Joos. Acoustic phonetics. *Language Monograph* No. 23, 1948 (Suppl. to *Language*, 24, No. 2).

[2] R. K. Potter, G. A. Kopp, and H. C. Green. *Visible Speech.* New York: D. Van Nostrand, 1947, 441 pp.

The sounds that were heard in the demonstration mentioned above were produced from spectrographic patterns by a machine constructed at the Haskins Laboratories in New York and called a pattern playback — it plays back the patterns of speech that are visible on spectrograms.

The principle of the playback is somewhat like that of a player piano. Spectrograms correspond to the perforated piano roll, and the individual sounds are pure tones rather than harmonic-rich notes from a piano. Briefly, the playback generates 50 harmonic tones 120 cycles apart, from 120 to 6,000 cps, in the form of beams of light modulated by a tone wheel. When a spectrogram, drawn with white paint, passes under the modulated light, each painted portion of the spectrogram reflects a portion of the light. This causes the corresponding harmonics to be heard after conversion of the light into sound by means of a phototube, amplifier, and loudspeaker. In the innermost circle of the tone wheel are four cycles of variation in film density, and from the center of the disk to the periphery, the number of cycles per revolution increases in steps of 4 cycles, reaching 200 cycles at the outermost, or 50th, of the concentric rings. The wheel rotates normally at 30 rps, and the light which passes through the wheel is therefore modulated at a fundamental of 120 cycles per second (corresponding to the innermost circle), and at each of the first 50 harmonics of that fundamental (corresponding to the 50 rings). It should be noted that the playback operates entirely on the basis of these 50 steady-state harmonics of the 120 cps fundamental, and does not resort to any generator of inharmonic sound — or noise — even for the production of highly fricative or plosive speech sounds; the noise-like and click-like sounds are produced, on the pattern playback, by brief tone-bursts scattered through restricted frequency-time ranges.

The very great advantage of the playback over other means of producing synthetic sounds is that it enables one to experiment with the dynamic aspect of speech — that is, with the rapid changes of formant frequencies in time — though it can also be used to deal with steady-state sounds. Thus, the acoustic counterparts of all linguistic units, from isolated phones to syllables and words, can be produced — and evaluated by ear — to determine the relation between acoustic stimulus and perception.

Our first experiments were to determine just how intelligibly the playback would talk. Spectrograms of sentences were played back, and the resulting sounds were presented to groups of college students. In spite of its monotone intonation, the playback speech was understood with rather few errors. Photographic spectrograms of human speech were used in one test; simplified, hand-painted copies of these spectrograms were used in the other. There was little difference in intelligibility.

[Some of these sentences, and the differences between playback speech from original spectrograms and speech from hand-painted simplifications, were demonstrated by recordings and slides.]

3. An Experiment With Vowel Transitions: The Voiced Stops

These experiments suggested a number of specific problems and a rather different approach to their solution. Instead of continuing with the simplification of the patterns found in spectrograms of connected speech, we have chosen to attempt the synthesis of individual speech sounds from the simplest possible patterns. Then, by varying systematically the components of these patterns and by presenting the resulting sound groups of listeners, we have attempted to determine which unitary configurations are essential constituents of speech patterns.

By proceeding in this way, we have been able to find a definite answer to the question of the importance of transitions in the perception of speech.[1] The fact that transitions exist in connected speech is immediately evident from spectrograms; that is, there are many instances in which the frequency positions of the vowel formants shift rapidly where vowels and consonants join. It seems clear that these transitions are the acoustic counterparts of the rapid articulatory shifts involved in passing from one speech sound to the next. The problem is one of interpretation: are these changes in the sound streams no more than the necessary transitions (as the name implies)

[1] A more detailed discussion will be found in: F. S. Cooper, P. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. Some Experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Amer.*, 1952, Vol. 24, pp. 597-606.

between steady-state sounds which serve to identify successive phonemes, or do they serve as important distinguishing characteristics of the phonemes? In other words, is the transition which one observes in a consonant-vowel combination merely a useless (or even confusing) residue, or may it be, in fact, a principal acoustic cue for the recognition of that combination?

Our exploratory work suggested the latter conclusion: we had, for example, found that the stop consonants in initial position were often characterized spectrographically by rapid transitions, and, conversely, that copying these transitions in the painted sentences seemed to convey the impression of a stop. However it seemed desirable to test the matter more systematically, and the stop consonants seemed to provide a useful starting point.

The design of the experiment is indicated in Fig. 1. Seven cardinal vowels (*i e ɛ a ɔ o u*) were used, each drawn with only two formants.[1] As can be seen in the figure, the extent of frequency shift (transition) of the second formant was varied in eleven steps (four transitions from a frequency above the steady-state position, one "straight", and six from below); the first formant had a constant "rising" transition in all cases. Part A of Fig. 1 shows these eleven transitions for one vowel, part B indicates the formant frequencies of the seven vowels, and part C shows one of the 77 "syllables" employed in the test.

The playback sounds derived from these patterns were assembled in random order and presented to groups of college undergraduates (not trained in phonetics) with instructions to identify the initial consonants as *b*, or *d*, or *g*. The listeners were in rather good agreement as to the particular transitions which corresponded to the individual stops, and it is therefore clear that transitions serve as cues for the stop consonants.

The contours of Fig. 2 indicate which transition-vowel combinations were most clearly heard as each of the three consonants. It should be noted that the best transitions for a particular consonant (especially *d* and *g*) vary from vowel

[1] See: P. Delattre, A. M. Liberman, and F. S. Cooper. Voyelles synthétiques à deux formantes et voyelles cardinales. *Le Maître Phonétique*, 1951, No. 96, 30-36; P. Delattre, A. M. Liberman, F. S. Cooper, and L. J. Gerstman. An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, 1952, Vol. 8, pp. 195-210.

to vowel — although in a systematic manner — so that one cannot identify a particular transition with a particular stop in all cases.

4. Bursts of Sound as Cues for the Unvoiced Stops

Another cue which we have studied in a similar set of experiments is the burst of noise which regularly precedes a voiceless stop at the beginning of a syllable.[1] Spectrograms show that the energy of such a burst tends to be localized in frequency, but the spectrograms do not indicate an unambiguous correlation of this frequency region with the identity of the consonant. Our test syllables combined a brief burst of noise centered at one of 12 frequency positions (ranging from 360 to 4,320 cps) with one of the seven cardinal vowels mentioned above. The vowels were represented by two formants only and did not have transitions. The 84 syllables were presented to groups of students for identification of the consonants (as *p*, or *t*, or *k*) with the results shown in the dominance contours of Fig. 3.

Again, the cue given by a single variable — the frequency region of the burst of noise — provided distinctions on which nearly all the subjects were agreed. Also, for the bursts, as for the transitions, the identification of a particular consonant cue varied according to the vowel with which that cue was associated. For example, the burst at 1,440 cps was heard as *p* before *i*, as *k* before *a*, and as *p* again before *u*. (We have already seen that the identification of transitions depended on the vowel with which the transition occurred). In short, the perception of these stimuli, and perhaps also the perception of their spoken counterparts, requires the consonant-vowel combination as a minimal acoustic unit.

5. Rate of Transition and the Distinction Between Vowels and Consonants

In our discussion of transitions we have so far considered the effects on perception of variations in the direction and extent of the frequency shifts. Exploratory research with the

[1] A detailed account may be found in: A. M. Liberman, P. Delattre, and F. S. Cooper. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *Amer. J. Psychol.*, 1952, Vol. 65, pp. 497-516.

playback has indicated that perceived distinctions among speech sounds can also be produced by varying the rate of transition, i.e., the rate at which the frequency of the formant changes.

Three instances of this are shown on Fig. 4. In the first (top row), all three spectrographic patterns start at the formant frequencies of *u* and terminate at the formant frequencies of *a*. The first pattern shows a change from *u* to *a* with the rate of change kept relatively slow for both formants. The auditory result (from the playback) is a vocalic sequence with a very gradual change of colour from *u* to *a*, which can be transcribed approximately by *ua*. The second pattern shows an increase in the rate of change which is sufficient to transform the initial *u* into the semi-vowel *w*, so that the auditory impression of the whole pattern is now *wa* instead of *ua*. The third pattern shows a further increase in rate of change, to the point where the auditory effect of the initial part reaches the consonantal state of *b*, so that the total syllable is perceived as *ba*. Relative durations of the transitions which yield *ua*, *wa*, *ba*, are to each other as 6:2:1. (It should be noted that the transitions of *wa* and *ba* are drawn in curved shape. This was done in order to conform with spectrograms of human speech; if the transitions were drawn as straight lines, the duration of the changes would be somewhat shorter for similar — but less life-like — sounds).

In the second illustration of the effects of rate of change (middle row), the three patterns start at about the frequency positions of *i* and terminate at those of *æ*. The first pattern, when played back, is heard as *iæ*, the second as *jæ*, the third as *gæ*. The relative durations of the three changes are roughly the same as before, that is, 6:2:1. Again, the successive increases in rate of change cause the auditory impression of an initial vowel (*u*) to become that of a semi-vowel (*j*) and finally that of a true consonant (*g*).

In these two series of illustrations, a difference in rate of change of the frequencies of formants 1 and 2 was sufficient to produce perceived differences ranging from vowel to consonant: a relatively slow rate of change from one frequency to another will be heard as a vowel or diphthong, while a somewhat faster change over the same frequency range will produce a clear impression of a consonant.

The third illustration (bottom row of Fig. 4) of the effects of rate of change is suggested by the results of some preliminary research on *l* and *m*. Comparing the first and second patterns in the third row of Fig. 4, we see that both start at approximately the frequency positions of *æ* and terminate at those of *u*. They have the same first formants, but different second formants. The first pattern, with a gradual frequency change for the low formant, is heard as *æu*; the second pattern, with a sudden frequency change in the low formant, is heard as *æl*. The difference between a vowel, *u*, and a consonant, *l* is here due to a difference in rate of change of formant 1 only.

Now let us compare the second and third patterns (also in the bottom row of Fig. 4). They exhibit similar sudden changes in formant 1 — changes that apply to resonants in general — but they differ in the rate of change of formant 2: the second pattern, with a slow rate of change in formant 2, is heard as *æl*; the third pattern, with a more rapid rate of change of formant 2, is heard as *æm*. (It should be noted that if neither the quality of *l* nor that of *m* is very clear here, it is because the consonantal resonance of formant 1 (at the extreme right of each pattern) was set at an intermediate point between that of *l* and that of *m* — a resonance that can serve moderately well for both *l* and *m* but is not quite right for either. This intermediate resonance was chosen for the special purpose of demonstrating the effect of rate of change. With a slightly higher formant 1 frequency (for *l*), a better *l* would be heard; with a slightly lower formant *l* frequency (for *m*), a better *m* would be heard.)

[Sound recordings corresponding to the nine spectrographic patterns of Fig. 4 were played.]

It will be important to define precisely the articulatory movements which, in actual speech, produce the frequency shifts we have been discussing. Some of the relations between articulation and acoustic pattern are well-known,[1] and further possibly fruitful speculations might be attempted. For the present, however, it may be sufficient to note that in sound sequences, certain movements which change the volumes

[1] The present position, in regard to the relationships between articulation and sound spectrograms, is given by: M. Joos. *op. cit.*; H. K. Dunn. The calculations of vowel resonances, and an electrical vocal tract. *J. Acoust. Soc. Amer.*, 1950, 22, 740-753; P. Delattre. The physiological interpretation of sound spectrograms. *Publ. Modern Lang. Assoc.*, 1951, 66, 864-875; and P. Delattre. Un triangle acoustique des voyelles orales du français. *French Review*, May 1948, 21, 6.

and openings (hence, the resonant frequencies) of the mouth cavities (see Fig. 4) might correspond either to vowels or to consonants depending on whether they are made slowly or rapidly.

6. Some Speculations

We should like to turn now, to some speculations about the ways in which the acoustic patterns may be related to the linguistic structure of the language.[1] We have described briefly two experiments on the characterization of the voiced and voiceless stops by means of transitions and bursts of noise. These results (for the single vowel *a*), together with those of a related experiment on second-formant transitions as cues for the nasal resonants, can be displayed in a 3 by 3 array based on articulatory features (Fig. 5).

When the acoustic patterns are thus arranged, one finds that the acoustic data seem to fall naturally into place, with the distinctions among columns (place of articulation) being given by the transitions of the second formant, and the distinctions among rows (manner of articulation) by three markers, namely, rising transitions of the first formant for the voiced stops, bursts of noise for the unvoiced stops, and a steady resonant portion of low intensity for the nasal resonants.

[The playback sounds corresponding to the nine syllables shown in Fig. 5 were played, first row by row, and then column by column.]

The details of such an array will differ for the different vowels, but acoustic similarities among linguistically related sounds seem to be the rule.

The data on bursts and transitions tempt one to further speculation about the perceptual process and its possible dependence on a set of binary choices. That is, Fig. 3 shows that a burst of noise preceding a vowel was always heard as *t* when the center frequency of the burst was high, but otherwise as *p* or *k*:

$$\text{Bursts:} \begin{cases} \text{High } (+) = t \\ \text{Low } (-) = p \text{ or } k \end{cases}$$

Also, as may be seen in Fig. 2, minus transitions of the second formant were always heard as *b* (or as *p* in another series of

[1] See reference at p. 548, note [1], above.

tests with second formant transitions and " forced " choices among the voiceless stops), whereas plus transitions were heard as either *d*(*t*) or *g*(*k*), depending on the vowel. Thus, ignoring for the moment the distinction between voiced and voiceless stops which would, in any case, be given by other cues:

$$\text{Transitions:} \begin{cases} \text{Plus} = t \text{ or } k \\ \text{Minus} = p \end{cases}$$

This provides a possible basis for deciding among *p*, *t*, and *k* by only two binary choices when bursts and transitions are both available as cues:

$p = - -$ (low burst, minus transition)
$t = + +$ (high burst, plus transition)
$k = - +$ (low burst, plus transition)

Whether or not this means that satisfactory stop consonants can be synthesized simply on a " plus-minus " basis without attention to the exact placement of bursts, or the precise degree of transitions, must remain a rather speculative question until explicit tests of this point are made.[1]

We have discussed elsewhere the significance which we attach to the apparently simple fact that spectrograms of many familiar sounds yield distinctive " pictures."[2] This seems to imply a parallelism between auditory and visual perception of patterned information sufficient to justify one's thinking about, and experimenting with, auditory patterns (after conversion into spectrograms) as if they really were visual patterns, and hence subject to all of the laws of visual patterning. This is, in fact, the rationale of our research procedure. In working with speech sounds, however, we are dealing with a special category of auditory patterns which have been correlated in the listener's experience with the motor gestures used in speaking, gestures which are not always related in a simple one-to-one fashion to the acoustic patterns they produce.

[1] We should, perhaps, point out that the kind of binary scheme being considered here differs in several respects from the system put forward by Jakobson *et al.* R. Jakobson, C. G. M. Fant, and M. Halle. *Preliminaries to Speech Analysis.* Technical Report No. 13, Acoustics Laboratory, Massachusetts Institute of Technology, May, 1952.

[2] F. S. Cooper, A. M. Liberman, and J. M. Borst. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences*, 1951, 37, 318-325; and F. S. Cooper. Spectrum analysis. *J. Acoust. Soc. Amer.*, 1950, 22, 761-762.

Some of our data — in particular the results of the experiment on bursts as stimuli for *ptk* (Fig. 3) — suggest that the listener's perception is more closely related to the articulatory gestures than to the acoustic patterns.[1] This is to say, for example, that certain acoustic stimuli which differ but little, or not at all, may nevertheless be perceived quite differently if the patterns to which they belong happen to be produced by very different kinds of articulatory movements.

We should like to make the further general observation that we have come to doubt the wisdom of searching too eagerly for the acoustic invariants of speech, insofar as the term " invariants " may be taken to imply a one-to-one correspondence between successive acoustic features of the sound stream and the phonemes of the message.[2] For the perception of some speech sounds, as we have pointed out earlier in this paper, it seems that the acoustic unit (or pattern) must include at least a consonant-vowel syllable.

7. Synthesis of Words

In conclusion, we shall consider the synthesis of connected speech. It was natural that we should attempt, even at this early stage in our research, to synthesize sentences, not by copying spectrograms of human speech but rather by free-hand drawing, using principally the limited information that we had obtained about a few phonemes. Figure 6 shows the hand-drawn spectrogram of the sentence.

A Playback Can Talk Back. This spectrogram, which is one of many we have synthesized without having previously seen the corresponding spectrograms of human speech, produces highly intelligible speech when converted into sound by the pattern playback.

Little need be said about syllabic durations and loudnesses here; they are almost as unvarying as the pitch is monotone! Only the first syllable is markedly shorter than the others, and the third is less intense. As to the vowels, we may note that they use only two formants and that the frequencies of

[1] For further discussion on this point see: reference 5; and P. Delattre, F. S. Cooper and A. M. Liberman. Some suggestions for teaching methods arising from research on the acoustic analysis and synthesis of speech. Institute of Languages and Linguistics, Washington, D.C., Monograph Series Number 2, 1952, pp. 31-47.

[2] See reference 3 at p. 548, note [1], above.

those two formants benefit from extrapolations of the cardinal vowels (the only ones on which we have worked systematically) to the vowels of American English.

The most interesting features are those that furnish the cues to the perception of the consonants. These cues seem surprisingly simple, especially when we recall the apparent complexity of spectrograms of human speech:

p The burst of noise corresponding to the explosion of *p* has a low frequency when in sequence with a " clear *l* " (one whose second formant is relatively high).

l The length of the *l* in clusters such as *pl*, *kl*, etc., is less than half the length of an initial or intervocal *l*. The first formant transition between a resonant (*l*) and a vowel (*eI*) is so sharp that it can be omitted in the painting. A better example of this abrupt transition can be seen in the fourth syllable *kæn* between the first formants of *æ* and the resonant *n*.

leI The beginning of formant 2 in *eI* is bent toward formant 2 of *l*. This would apply to any vowel in sequence with an *l*: the formant 2 vowel transition is always " in continuity " with the formant 2 resonance of the *l*.

b Formant 1 of the vowels preceding and following *b* have minus transitions. (Minus means that the transition falls to, or rises from a lower frequency than that of the steady state of the vowel.) This minus transition applies to all voiced stops (*b d g*) and seems to make an important contribution to voicing. Formant 2 of the vowels preceding and following *b* also have minus transitions. This cue distinguishes labial stops (*p b m*) from dental or velar stops (*t d n, k g ŋ*). It should be noted that the minus transition of the vowel that precedes the intervocalic *b* is not indispensable — only that of the vowel which follows *b* is essential — but it helps in perceiving the word as an English word, since consonant anticipation is a characteristic feature of English.

æk Formant 1 of vowel *æ* in *bæk* ends in a very slight minus transition; this cue contributes to the voicelessness of the stops (*p t k*). However, the unvoicing effect is mainly produced, as in all the voiceless stops, by the burst of sound. Formant 2 of the vowel ends in a marked plus transition. This cue, with the vowel *æ*, distinguishes velar stops (*g k ŋ*) from dental or labial stops (*t d n, p b m*). The frequency of the main burst for *k* is located just above the frequency of formant 2 of the contiguous vowel, whatever the vowel may be. This cue aids in distinguishing from the other voiceless stops (*p, t*), but it probably contributes even more to the voicelessness of *k*.

kæ All that applies to *æk* on formant 1 transition, formant 2 transition and burst frequency, also applies to *kæ*, but in reversed sequence. The distance from the *k* burst to the vowel in *kæ* is sufficient to produce an effect of aspiration. In *æk*, the burst was placed close to the vowel in order to obtain the very implosive effect of the first element in a geminate; in absolute final, as in the last word of the sentence (*bæk*), the *k* burst is well separated from the preceding consonant.

æn Formant 2 of *æ* ends in a small plus transition. This cue, with *æ*, distinguishes dental stops (*n d t*) from velar and labial stops (*k g ŋ, p b m*).

n Formant 1 of *n* is lower than formant 1 of *l* (*pleI*). Formant 2 of *n* is " in discontinuity " with the formant 2 transition of the contiguous vowel. Both these cues distinguish nasal stops (*m n ŋ*) from oral stops (*p b, k g, t d*).

tɔk As was the case for the *k* sounds of *bæk* and *kæn*, the voiceless effect in *t* and *k* of *tɔk* are due principally to the bursts and in lesser measure to the reduced transitions of formant 1.

t The *t* burst is always high, but with the vowel *ɔ*, this burst does very little to distinguish *t* from the other voiceless stops (*k p*). More important, with *ɔ*, is the large plus transition at the beginning of formant 2. This cue serves, in fact, to distinguish the dental stops from the velar or labial stops with the vowel *ɔ*.

k Exactly the opposite occurs in the case of *k* with *ɔ*. Here, the small plus transition at the end of formant 2 — a minor cue with *ɔ* — does little to distinguish *k* from the other voiceless stops (*p t*); for *k* with *ɔ*, this distinction is due mostly to the frequency position of the burst. The frequency of the burst, as for *k* with other vowels, is centered just above formant 2. The space between the the vowel *ɔ* and the *k* burst is shorter than in an absolute final (e.g., the *k* of the last word, *bæk*), but longer than in a geminate implosive such as that of the third syllable, *bæk*. The *k* of *tɔk* is also implosive, but is not in a geminate.

bæk The last word *bæk* is similar to the third word except for the longer space between the vowel and the *k* burst. The increased space gives the effect of a " release."

It may be useful to return to the *k* of *bæk* to compare it with the *t* and *k* of *tɔk*. Two cues are used in each case, but with important differences in their relative contributions as distinguishing factors: For *t* with *ɔ*, the main cue is the transition; the burst is minor. For *k* with *ɔ*, the main cue is the burst position; the transition is minor. For *k* with *æ*, both the transition and the burst are good cues. This explains why the *k* in *kæ* is more distinct than the *t* in *tɔ* or the *k* in *kɔ*. To add distinctness to the *t* and *k* of *tɔk*, a transition in the third formant can be used.

[Recordings of this sentence, and of other words and sentences, were played].

The synthesis of connected speech, such as that we have just presented, was not a goal in itself; rather, its purpose was to test the experimental results obtained with the phonemes we have studied thus far, and, in general, to assess whatever progress we may have made toward an understanding of the acoustic bases for the perception of speech.

8. Appendix[1]

Synthesis of the syllables of MI DO LA

(Text of the London Demonstration)

The sounds will be displayed visually on a linear scale of harmonics [see any row Figs. 7, 8, 9], with the lowest frequencies at the bottom, and including the first 30 harmonics of a fundamental tone of 120 cps. Time is represented as moving from left to right. Intensity is shown by the width of each harmonic.

The first 30 harmonics to be used in our speech reconstruction will be played, as they are displayed visually in the three dimensions of frequency, time, and intensity [Row A].

Naturally, selected tones from the ones just heard can be combined to produce pleasant chords [B1].

But this type of chord does not occur in speech. Chords found in speech are mostly of a dissonant type [B2].

The three syllables MI DO LA, to be synthesized in small steps, will first be played fully reconstructed [B3].

The recomposition of these three syllables will start from the simplest elements in their vowels. (The two horizontal bars that are visible for *i*, *o*, *a*, correspond to formant 1 (the lower band) and formant 2 (the higher band) of these vowels).

Listen to three pure tones at the center frequencies of the second formant for *i*, *o*, *a* [C1].

Listen to three pure tones at the center frequencies of the first formants for *i*, *o*, *a* [C2].

Let us play together both notes of *i* [C3], both notes of *o* [C4], and both notes of *a* [C5].

The intelligibility is low, but the vowels can already be recognized.

Vowel formants for a male voice at 120 cps usually are composed of about three harmonics, and not just one as above. We shall add below and above the original notes the closest harmonics to each one. That will give us six notes per vowel, three for each formant.

Six notes that yield an *i* sound will be played successively [D1, D2]; then simultaneously [D3, D4]. (D4 should have nearly the same sound as D3. The difference is that with the D3 type of drawing the distribution of intensity among the harmonics can be drawn with more precision, while with the D4 type, the drawing can be done more rapidly.)

Six notes that yield an *o* sound will be played successively. [E1, E2]; then simultaneously [E3, E4].

Six notes that yield an *a* sound will be played successively [F1, F2]; then simultanously [F3, F4].

Compare the degrees of intelligibility for *i*, *o*, *a*, when composed of two notes [G1], and when composed of six notes for each vowel [G2].

Let us now reconstruct the syllable MI. Listen successively to the notes for a formant 1 and for a formant 2 that compose a nasal resonance such as that of a nasal vowel said with closed mouth [H1].

The same notes will be played simultaneously [H2].

The nasal resonance is not very intelligible by itself; but note that the corresponding resonance, produced in isolation by mouth, would not be very intelligible either.

Now, let us play in sequence, the nasal resonance and the vowel *i* [H3].

The result is intelligible, but poor, because the transitional part between the two sounds is not there. If we connect the nasal sound and the vowel *i* by a rapidly rising frequency for the second formant, the intelligibility is greatly improved [H4].

Some additional nasal resonance at a place where it is " in discontinuity " with the transition of the second formant will further increase the intelligibility [H5]. (This acoustic discontinuity seems to correspond in articulatory terms to the raising, or lowering, of the velum, which makes a sudden change in the acoustic pattern. " Continuity," on the other hand, would be required in passing from an oral consonant to an oral vowel, as in passing from *l* to *a* in the syllable LA [see K4].

Let us now reconstruct the syllable DO.

In order to perceive a *d* before *o*, we could imitate an explosion, composed of a brief burst of sound at a high frequency, including in the burst 5 or 6 contiguous harmonics. Listen to three such bursts in quick succession [J1].

Let us play one such burst before the vowel *o* [J2]. The

[1] A disk or tape recording of this, and the other sound demonstrations described in the paper, may be obtained, at cost, by writing to the Haskins Laboratories, 305 East 43rd Street, New York, U.S.A.

resulting consonant sounds voiceless. We shall add voicing in the form of a fundamental tone at 120 cps. Listen to such a fundamental tone [J3].

Let us play the burst with voicing before the vowel [J4].

The result is still quite poor. Let us try formant transitions. From *d* to *o*, the first formant has to rise [J5].

The second formant has to fall sharply in frequency [J6].

Let us play both transitions together [J7].

The results are superior to those obtained with a voiced burst of sound [J4], indicating that the perception of *d* before *o* depends perhaps more on the transitional effects than on the explosive effects.

We could add the burst to the transitions, but it will make little difference: compare DO with and without the burst [J8, J9].

Even the voicing seems redundant. Its perception may come largely from the rise of the first formant: compare DO with and without voicing [J9, J7].

Let us now reconstruct the syllable LA.

For the *l*, we have a resonant sound of the same type as for the *m* in MI. Only the mouth is more open, so the first formant will be higher for *l* than it was for *m*. Listen to the notes of *l* in succession [K1].

Listen to them played together [K2].

Heard this way in isolation, it would be hard to recognize the resonance of *l*. But so would it if the sound were made in isolation by human voice.

Let us sound the *l* resonance and the *a* in succession [K3].

The result is already satisfactory. But it can be improved if some transition is added between the steady state of *l* and the steady state of the vowel. This transition, in contrast with that of *m*, must be " in continuity " with the resonance of the second formant of the *l* [K4].

Let us sound the three syllables in succession: MI DO LA [L1].

From the point of view of linguistic perception, we cannot improve very much on this. We could however add a third formant to DO and LA, in order to equalize the " voice " of the three vowels (*i* having more high notes than *o* or *a* in its two essential formants) [L2]. (Proper transitions in the third formant will often improve the intelligibility of the

consonants and, indeed, the third-formant transitions appear in some cases to make truly important contributions to the perception of these sounds).

We could further improve the realism, the naturalness, of the three syllables by adding voice formants in the high-frequency region where, it seems, one aspect of voice differentiation is located [L3].

Note, however, that the use of only two formants gives a satisfactory degree of intelligibility. Compare the three syllables, played in reverse order [L3, L2, L1].

The final visual patterns of MI DO LA that you have seen and heard [L1], are hand-painted schematic spectrograms that show only the principal acoustic cues necessary for the perception of these syllables. Spectrograms made from human pronunciation of MI DO LA would, of course, be considerably more complex, and the important acoustic cues would in most cases be more difficult to see and specify.

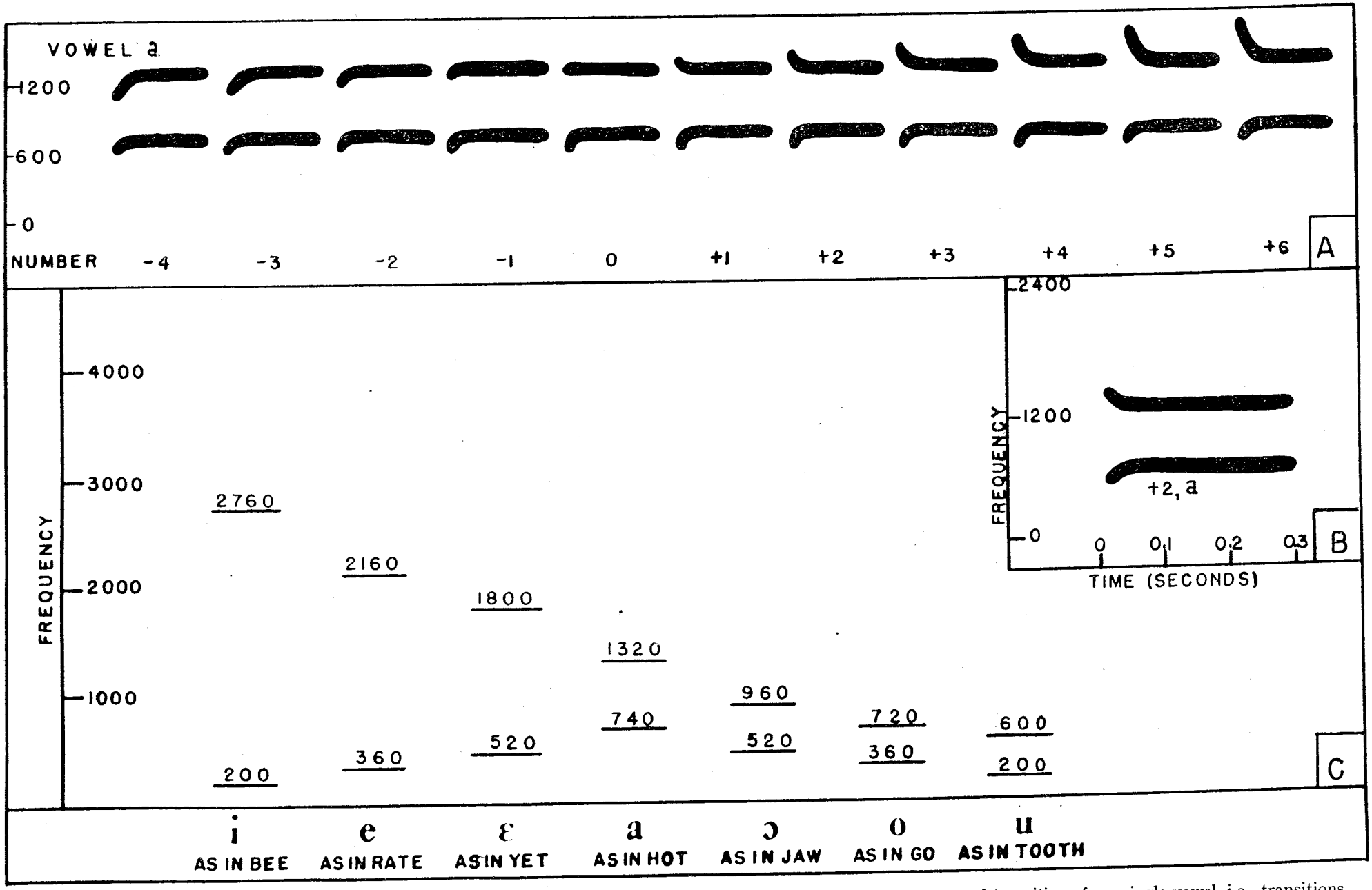

FIG. 1.—Spectrographic patterns of the sounds used in a test of second-formant transitions. A: The complete range of transitions for a single vowel, i.e., transitions from —4 to +6 for *a*. B: One of the 77 test stimuli. Listeners were asked to identify the initial consonant (due to the transition) as one of the voiced stops, *b*, *d*, or *g*. C: Formant frequencies of the seven two-formant vowels.

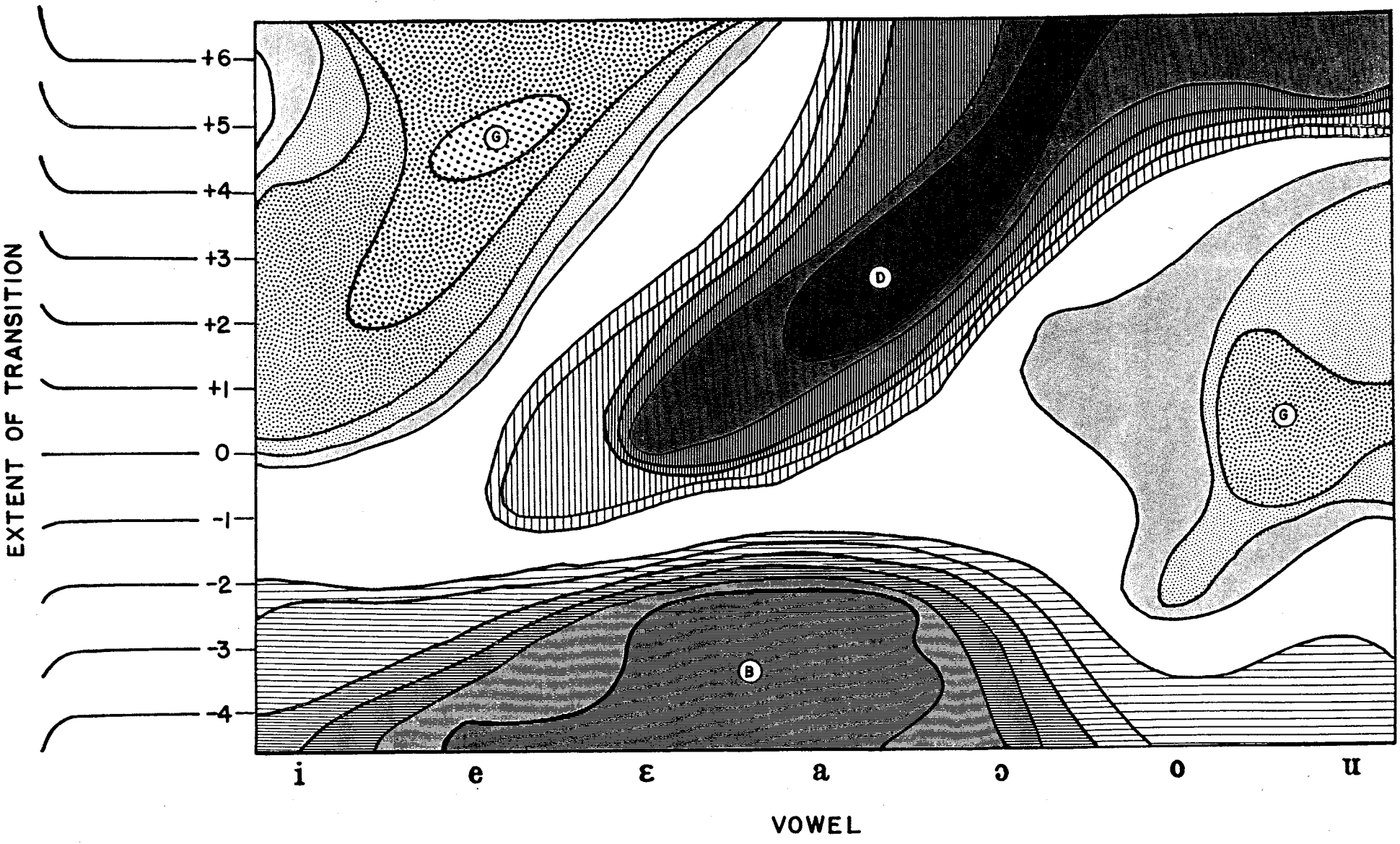

FIG. 2.—Results of the test on second-formant transitions. The contours indicate the extent to which judgments of one of the three voiced stops exceeded the sum of judgments of the other two stops for the various combinations of vowels (x-axis) and transitions (y-axis). For example, a +5 transition with *e* was heard as *ge* by a large majority of the listeners; also +3 with *u* was heard as *gu*, but not by quite so large a majority.

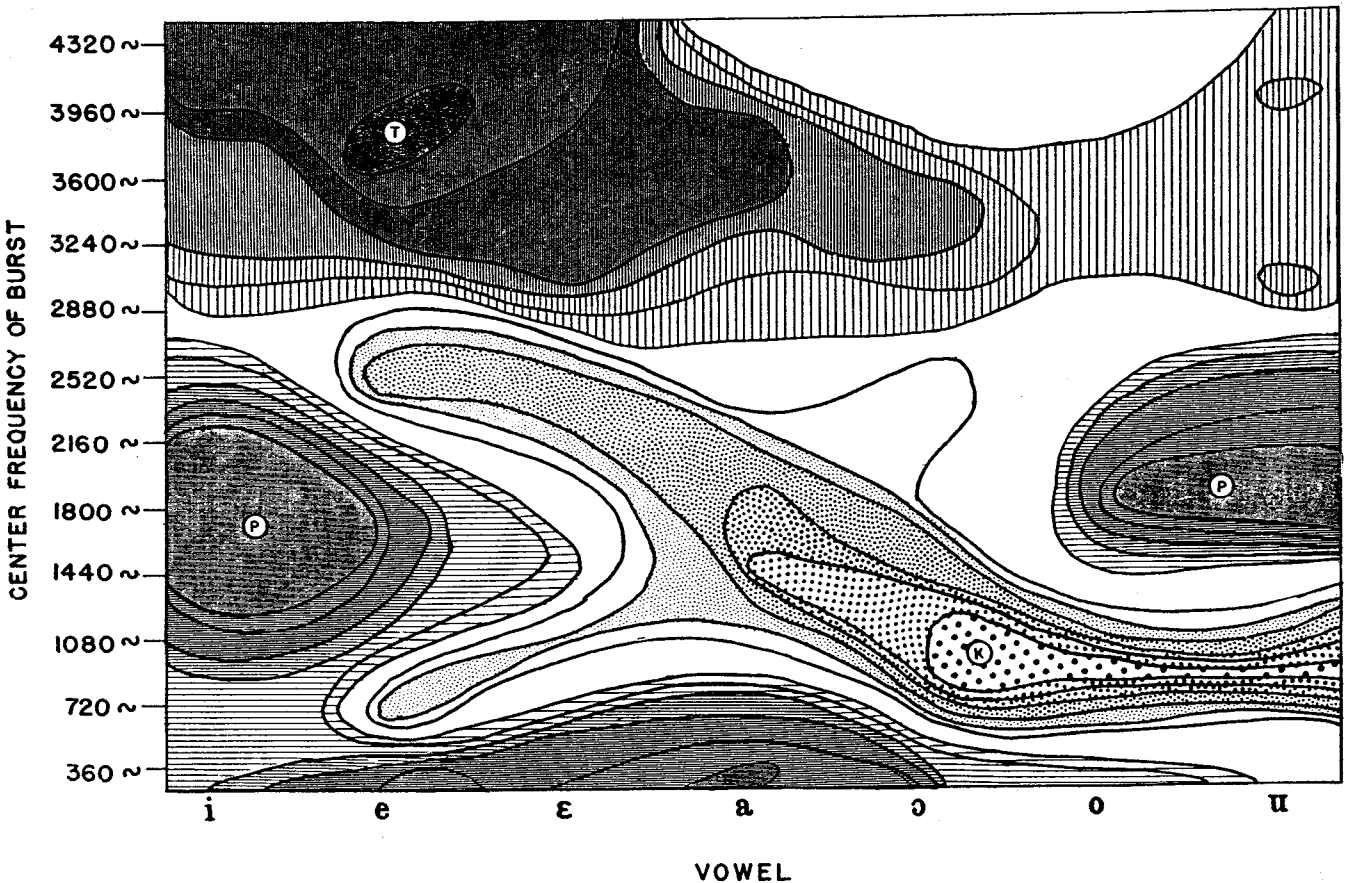

FIG. 3.—Results of a test on the unvoiced stop consonants in which bursts of noise in various frequency regions were paired with two-formant vowels. The contours indicate the extent to which judgments of one of the three unvoiced stops exceeded the sum of judgments of the other two stops for various combinations of vowels (x-axis) and bursts of sound (y-axis).

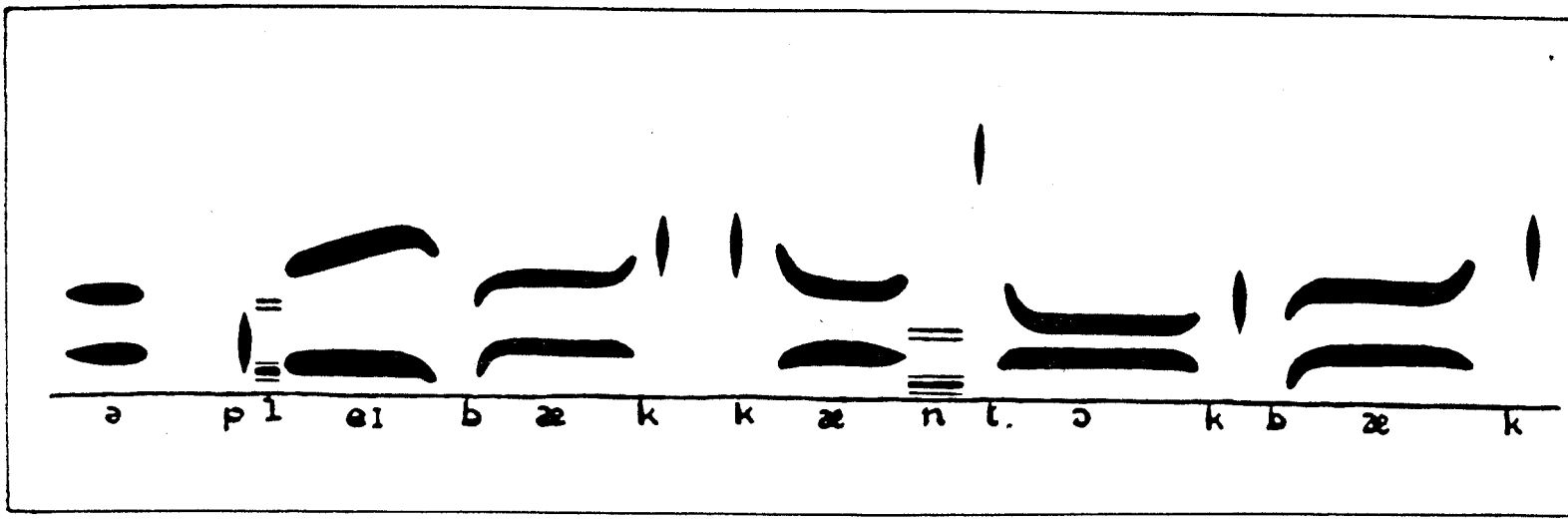

FIG. 6.—Synthesis by rule. An example of a sentence painted directly from the typewritten text without reference to spectrograms of th spoken sounds. Consonants and vowels are drawn in accordance with the results of the systematic tests described in the text.

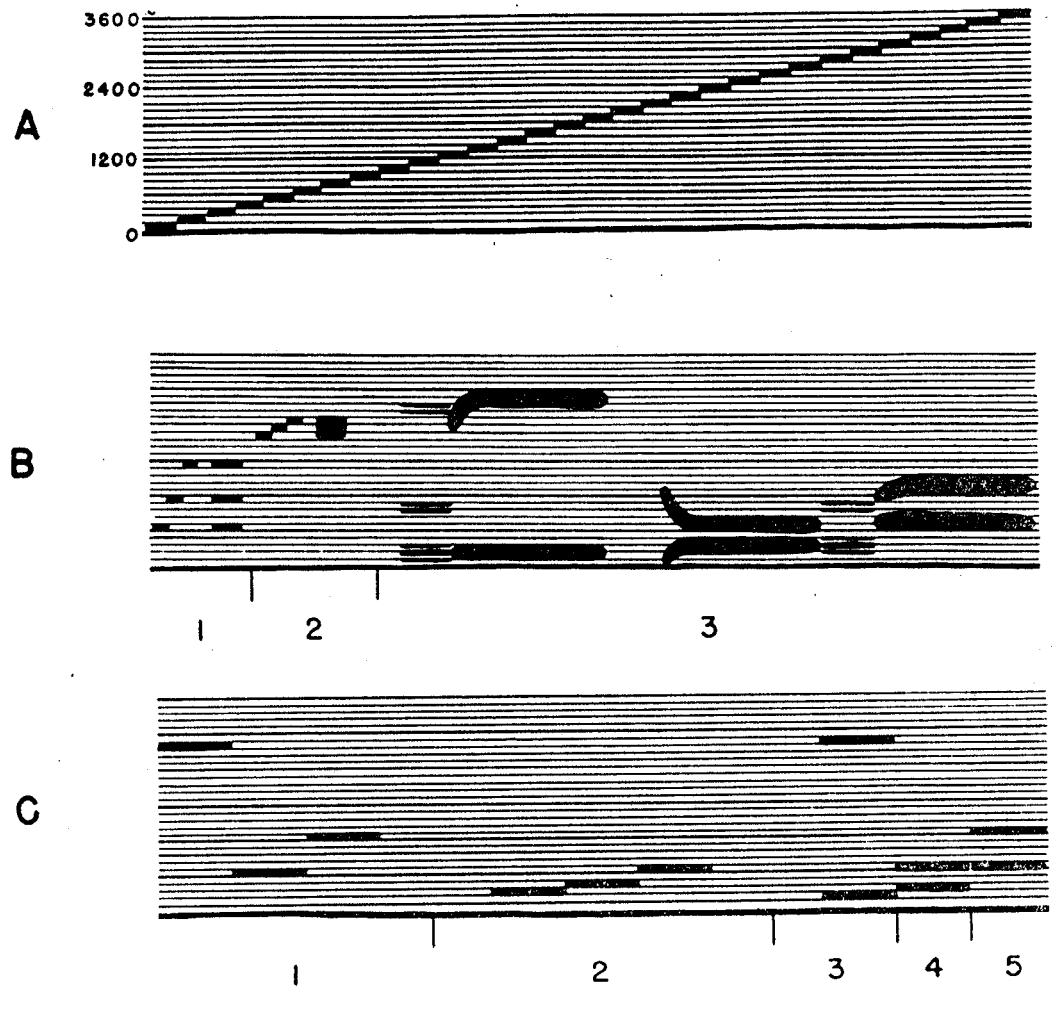

FIG. 7.—Spectrographic patterns corresponding to the sound demonstration described in the Appendix: Parts A-C.

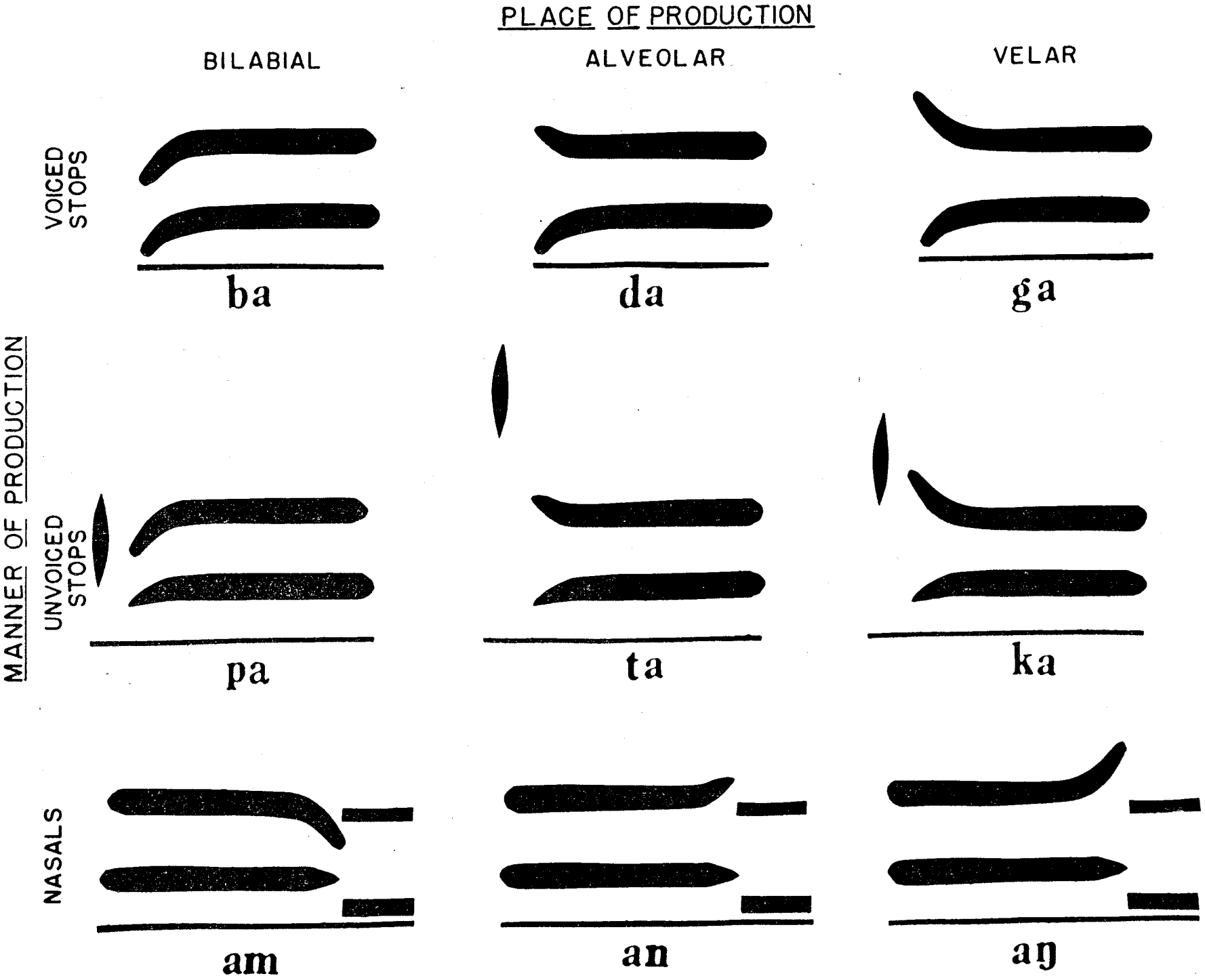

FIG. 5.—Relationships between acoustic and articulatory features of the stops and nasal resonants (with the vowel *a*). The articulatory array is conventional; the spectrographic patterns entered on it are those which were preferred (for the nine syllables) by the subjects in a

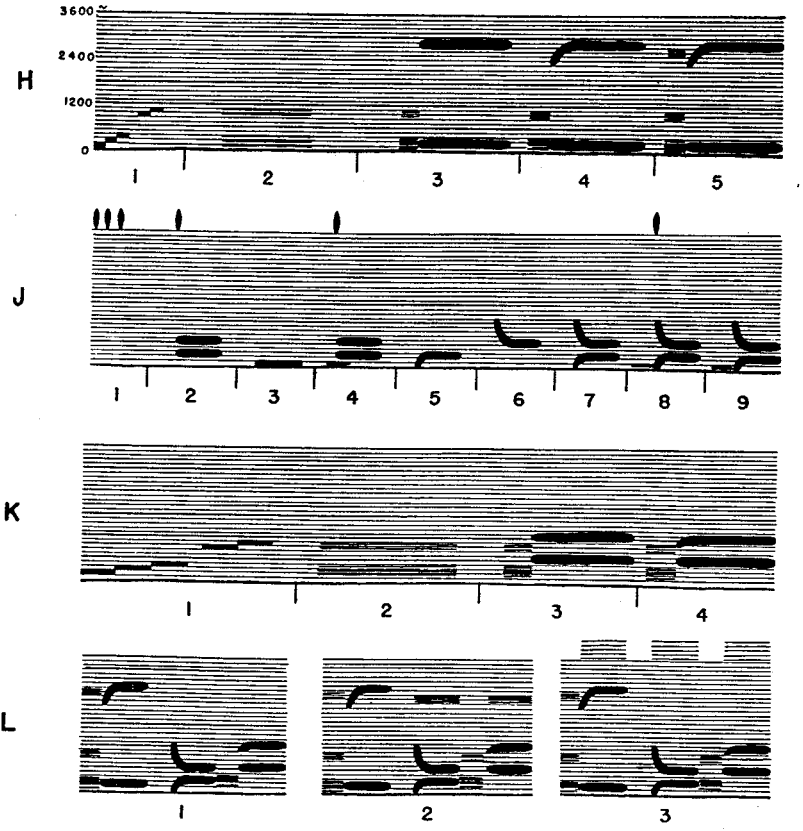

FIG. 9.—Spectrographic patterns corresponding to the sound demonstration described in the Appendix: Parts H-L.

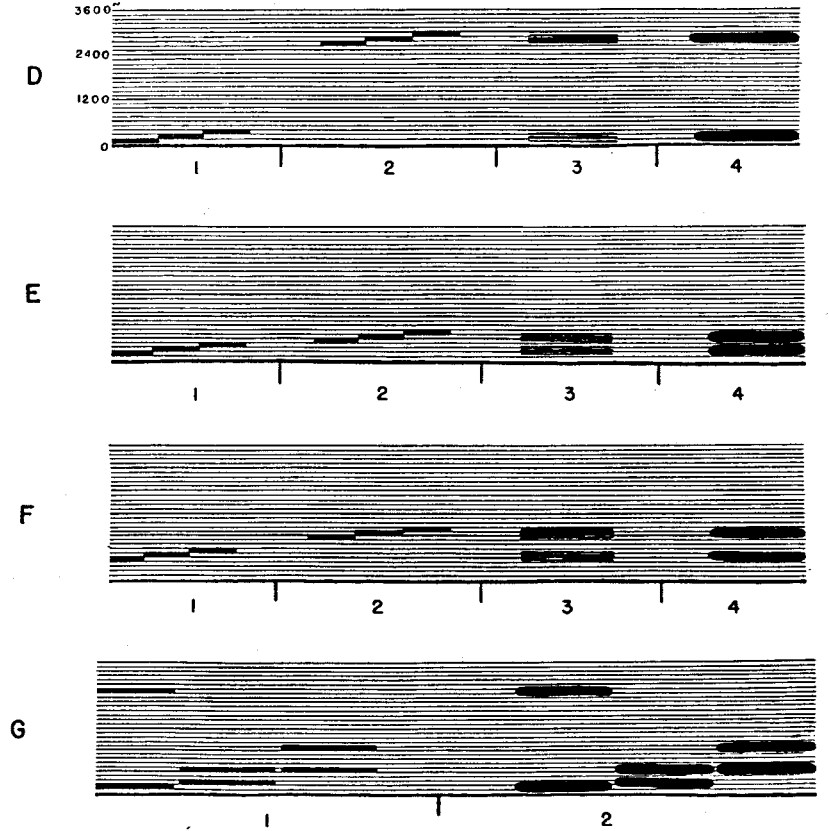

FIG. 8.—Spectrographic patterns corresponding to the sound demonstration described in the Appendix: Parts D-G.