# Pitch Continuity and Speech Source Attribution

## C. J. Darwin and C. E. Bethell-Fox
### University of Sussex, Brighton, England

Two experiments showed that continuity of the pitch contour is an important perceptual indicator that a formant pattern comes from a single speech source. Subjects listened to repeatedly played formant patterns that changed smoothly between two vowels. When the pitch was a monotone, these patterns were heard as containing semivowels and liquid consonants; but when a discontinuous, steplike pitch contour was imposed on the patterns, they divided into two perceptually distinct speech sources, and the phonemic percept changed to a predominance of stop consonants. In the second experiment, this effect was shown to be due to changes in pitch itself, rather than to concomitant changes in amplitude or energy. Earlier suggestions that continuity of both formant structure and pitch contour are important for the perceptual coherence of speech were borne out.

When one listens to someone speaking against an irrelevant background of other voices, how does one know which sounds are part of the attended speech and which form the background? Differences in spatial location (Sayers & Cherry, 1957) and amplitude (Egan, Carterette, & Thwing, 1954) help the human listener to attend selectively to one voice among many, and a common pitch allows one to bundle together the individual formants of a single speaker (Broadbent & Ladefoged, 1957; Parsons, 1976). Although these dimensions of the speech wave are useful for sorting the elements of sound present at any one instant into potential *gestalten*, the listener must make additional assumptions to determine which bundles from successive moments of time belong to the same sound source. Speakers change their location, loudness, and voice pitch from moment to moment, so some rule must be used by a listener for these dimensions to be used successfully in natural situations.

This problem of the temporal coherence of auditory patterns is one which the gestalt psychologists saw as analogous to the visual problem of spatial grouping and subject to the same laws (Koffka, 1935, pp. 434–448). There has since been ample demonstration of the applicability of the gestalt laws of good continuity, and of proximity and equality, to the perception of tonal patterns and speech. A rapid sequence of alternating tones will divide into two perceptually distinct streams if successive tones differ sufficiently in frequency (Miller & Heise, 1950) or if one tone in a repeating pattern departs substantially in frequency from an established contour (Heise & Miller, 1951). A consequence of this segregation is that temporal order judgments between tones that are part of different streams are much more difficult to make than judgments between tones forming part of the same stream (Bregman & Campbell, 1971; Norman, 1967). A similar result has also been found using single formant patterns of variable resonant frequency with pitch held constant (Kinney, 1961) and for sequences of monotonic natural and synthetic vowels (Dorman, Cutting, & Raphael, 1975; Thomas, Hill, Carroll, & Garcia, 1970). A monotonic vowel sequence seems to be more likely to split into two streams if adjacent vowels differ markedly in their first, and to a lesser

extent subsequent, formant frequencies (Dorman et al, 1975). Cole and Scott (1973) describe a similar perceptual segregation with a consequent impairment of temporal order identification for repeatedly played sequences of consonant frication alternating with a steady-state vowel. If these streaming effects can occur with durations of elements considerably longer than phone durations in normal speech, why does speech not normally segregate? The answer is partly that the formant pattern generated by the continuously moving articulators generally changes smoothly, and smoothly changing sounds are less prone to streaming than those with abrupt changes (Bregman & Dannenbring, 1973; Cole & Scott, 1973; Dorman et al, 1975). Abrupt changes in the frequency of pure tones, or in the formants of speech when these form the elements of a repeating pattern, thus tend to favor the formation of perceptually independent streams apparently originating from two sound sources.

It is possible that other dimensions of the speech signal also contribute to its perceptual coherence. On the basis of an experiment that found that prosodic continuity strongly determined the number of intrusions from a spatially different sound source in a shadowing task, Darwin (1975) suggested that continuity in the pitch contour, along with formant continuity, contributes to the perceptual coherence of speech. In addition, Parsons (1976) has found it possible to separate two voices automatically by using the principle of pitch continuity to assign each bundle of formants to a particular speaker. Dorman, Raphael, Liberman, and Repp (Note 1) have shown that when adjacent speech syllables are spoken by different speakers, the phonemic percept can be different from that which occurs when they are both spoken by the same speaker. Our experiments used a different, but related, perceptual transformation to elucidate what it is that indicates to the listener that the speaker has changed.

Our own preliminary observations showed that the pitch contour of a set of repeatedly played, concatenated, steady-state vowels can influence the way in which these vowels are perceptually grouped. If, say, four different vowels, each lasting 120 msec, are played repeatedly, vowels 1 and 2 might be heard as one speaker, and vowels 3 and 4 as another. But if a pitch contour that, when played as pure tones, groups alternate notes together is imposed on these vowels, they regroup according to the pitch. Nooteboom, Brokx, and De Rooij (Note 2) have independently made a similar observation with alternating pitches imposed on a set of three vowels played repeatedly. They found that whether one or two voices were heard depended on the rate at which the vowels were played and on the pitch excursion in a way similar to that found with pure tones.

Our own observations led us to examine whether this grouping of formant patterns by their pitch contour would hold even if the formants changed smoothly from one vowel to the next. We thus asked whether pitch discontinuity would be able to override formant continuity in assigning sounds to speakers. The pilot experiment that we ran to test this used repeatedly played formant patterns that changed smoothly, via 60-msec formant transitions, between four different 60-msec vowel steady states. On these patterns were imposed two types of pitch, a monotone and a discontinuous, steplike pattern in which the pitch changed abruptly half way through each of the formant transitions linking adjacent vowels. Naive listeners confirmed our own impressions that the phonemic percept was very different for the different pitch contours. With a monotone, the predominant consonantal percepts were sonorants, as one would expect from a formant pattern moving from one steady-state value to another value via relatively long (120 msec) transitions. With the step-pitch contour, however, a large number of stop consonants were heard. Since in natural speech important cues for intervocalic stop consonants are a period of silence during closure and a rapid spectral change (of around 60 msec or less) before and after the closure, the perception of stops here indicated that the continuous formant pat-

terns were being perceptually divided into two distinct streams (with the point of division being the point of pitch change) and that each stream was presumed to be silent during the presence of the other. This account also corresponded with the subjective impression that there were now two speakers, one speaking at a pitch higher than the other's pitch.

## Experiment 1

If, as our preliminary observations indicated, discontinuities in the pitch contour can cause perceptual segregation of continuous formant patterns, then it should be possible to arrange formant patterns that will be heard as containing stop consonants with different places of articulation when the appropriate pitch contour is imposed on them, but that will not be heard as containing stops when the pitch contour is flat.

### Method

*Stimuli.* On each trial of this experiment subjects heard a repeating formant pattern that alternated between two steady-state values via linear formant transitions. The patterns used are shown in Figure 1. Each pattern had as one of its steady-state values formants appropriate to the vowel [æ] (726, 1,327, and 2,411 Hz) and this value, like the other steady-state value, lasted for 60 msec. For 60 msec on either side of this vowel, the formants changed linearly until they reached the values for the other steady-state vowel. The formant values for these other vowels were chosen so that the formant transitions 30 msec on either side of the [æ] vowel gave patterns appropriate to one of the three voiced stop consonants [b, d, g]. This could readily be achieved for [bæd] and [dæd] (patterns 1 and 2), but only an approximation to a [gæg] pattern was possible because a natural [gæg] starts and ends with second and third formants that almost converge. Continued linear extrapolation of these values would have resulted in the second and third formants crossing, a physically impossible event for the vocal tract. We therefore adopted a compromise pattern that had the second and third formants close together in the steady state. This compromise in fact led to a very weak /g/ percept which we attempted (unsuccessfully) to improve by lowering the frequency of the first formant to give a lower first-formant frequency onset.

Each of these three continuous formant patterns had imposed on it one of three pitch contours, which are also shown in Figure 1: (a) a monotone at 130 Hz, (b) a pattern which was low pitched (101 Hz)
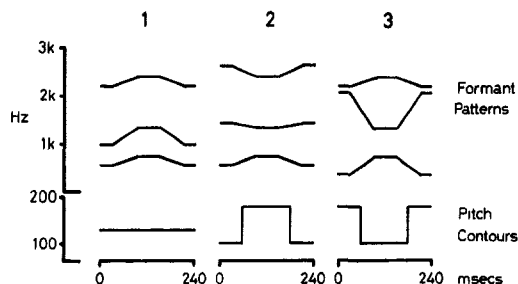


*Figure 1.* One cycle of each of the formant patterns and pitch contours used in Experiment 1. (The relative timing of pitch and formant movement is as indicated, but each formant pattern was used with each of the pitch contours.)

on the non-[æ] vowel (and during the 30 msec of transition on either side of it) and high pitched (178 Hz) on the [æ] vowel (and 30 msec on either side of it), and (c) a pattern that was like the second, but with the low pitch on the [æ] vowel and the high pitch on the other vowel.

These sounds were produced by a serial formant synthesizer under control of a PDP-12 computer. The synthesizer control parameters were updated every millisecond and interpolated between adjacent millisecond values. The abrupt pitch changes were therefore made within 1 msec. The linear transitions of formant values in fact consisted of steps 5 msec in duration.

*Subjects and procedure.* Eight undergraduate students with no known hearing impairment listened individually over a loudspeaker to a tape consisting of 54 trials. Each trial was one of the nine possible stimulus patterns (3 Formant Patterns × 3 Pitch Contours) repeatedly played for 30 sec, 20 sec of which was a gradual increase and decrease in amplitude at the beginning and end, respectively. The synthesizer control program was modified to ensure that there were no discontinuities in the sound produced by the cyclical repetition of a single set of control parameters. The subjects were told to listen to the consonant(s) immediately preceding the /æ/ vowel and to indicate which they had heard by entering a confidence rating (from 1 to 5, indicating increasing degrees of confidence) in appropriate columns of their answer sheet. This sheet contained columns headed b, d, g, w, r, l, and y (these being the consonants heard most often in our own listening to the tape), with a final column left blank. They were free to enter other consonants into this column and also free to mark as many columns as they wanted on each trial. They were told that the consonant might change during the trial.

### Results

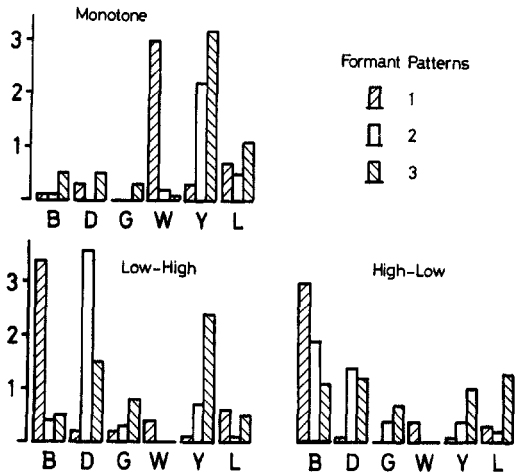Over 99% of the confidence ratings were made to the consonants b, d, g, w, l, and y,

Figure 2. Average confidence ratings for perceived consonants in Experiment 1. (The three pitch-contour descriptors—monotone, low–high, and high–low—refer to contours 1 through 3 of Figure 1, in that order.)

and so only their overall average confidence ratings are shown in the histograms of Figure 2. The main result that we anticipated was confirmed, since the ratio of stop consonant ratings to other consonant ratings is higher in the two conditions with a discontinuous pitch contour than in the condition with the monotone, although this effect was greater for formant patterns 1 and 2 than for 3. These effects were all shown by each of the eight subjects but also were assessed statistically by a four-way analysis of variance which had as dimensions perceived *manner of articulation* (stop vs. not-stop consonant), type of pitch contour (flat vs. step), formant pattern (pattern 1 vs. 2 vs. 3), and subjects. This analysis showed a significant interaction between perceived manner and pitch contour, $F(1, 7) = 56.9$, $p < .0001$, and a three-way interaction between these two dimensions and formant pattern, $F(2, 14) = 9.83$, $p < .005$.

There was also a large change in the place of articulation of the perceived consonant depending on the formant pattern used. This is hardly surprising for the predominant lateral and semivowel responses to the monotone patterns but is much more interesting for the stop-consonant responses

to the changing pitch patterns, since it indicates that the impression of a stop consonant is being produced by a segmentation of the formant pattern close to the point at which the pitch changes. All eight subjects had their largest b scores on formant pattern 1 and their largest d scores on pattern 2, although the g scores were divided between patterns 2 and 3. This latter effect appeared as an interaction between perceived place of articulation and formant pattern, $F(4, 28) = 29.5$, $p < .0001$, in an analysis of variance that had as dimensions perceived place of articulation (b vs. d vs. g), pitch contour (contour 2 vs. contour 3 of Figure 1), formant pattern (pattern 1 vs. 2 vs. 3 of Figure 1), and subjects. Although the two discontinuous pitch contours led to equivalent confidence ratings for stops as a whole, the contour which had the higher pitch on the [æ] vowel (pattern 2, low–high) gave a clearer effect of formant pattern on the perceived place of articulation, both for the (predominant) stops, $F(4, 28) = 6.03$, $p < .005$, in this analysis and for the other consonants reported, $F(6, 42) = 9.11$, $p < .001$, in an analogous analysis on the responses that were *not* stops (w, r, l, y).

*Discussion*

When a continuous formant pattern is heard repeatedly, the consonants perceived in that pattern change profoundly depending on the pitch contour imposed on the pattern. The effects that we have observed can be explained by assuming that the two different components of the step-pitch contour form two perceptually distinct streams, so that the formant pattern is interpreted as being the alternated speech of two differently pitched speakers, rather than that of a single speaker whose voice has the required pitch discontinuities. This segregation has the effect of producing illusory silences in each stream during the portions of the signal attributed to the other stream, and these silences are in turn interpreted, along with the rising first formant in our sounds, as indicating the presence of a stop consonant. Further-

more, the place of articulation of this perceived stop consonant depends on the pattern of formant movement at the time that the pitch changes.

Although the effect is striking in the stimuli we have used here, it is not clear from this experiment whether the effect is in fact due to changes in pitch or rather to some associated variable such as the energy or loudness of the resulting sounds. Here, no attempt was made to control for the increased energy of higher pitched sounds (although different vowels were equated at the lower pitch for equal subjective loudness). The step in energy caused by a sudden change in pitch is considerable, since the pulse generator which provided periodic excitation to the formant filters had a constant pulse width (and consequently a variable duty cycle). Doubling the pitch would thus double the time-averaged energy. The next experiment examines the effects of energy discontinuities and of pitch discontinuities when an attempt has been made to keep energy and peak amplitude constant.

## Experiment 2

### Method

*Stimuli.* Two of the formant patterns (patterns 1 and 2) from the previous experiment were used, with various amplitude and pitch manipulations as follows:

1. Equal peak – flat pitch: With the pitch kept at 130 Hz for the entire stimulus, each 5-msec-duration line of the synthesizer control parameter table had its amplitude individually adjusted so that the resulting sounds all had the same peak amplitude displacement when measured on an oscilloscope screen.

2. Equal peak – step pitch: With the pitch changing from 101 to 178 Hz half way through the transitions into the [æ] vowel and back again half way through the transitions out of it (pitch contour 2), the same equalizing procedure was used on each 5-msec line of control parameters.

3. Equal energy – flat pitch: Similar to number 1, with each segment adjusted to give the same VU reading after bandpassing between 45 and 3,500 Hz.

4. Equal energy – step pitch: The same pitch contour as number 2, with the same energy equalizing procedure as number 3.

5. Half peak – flat pitch: The same as number 1, except that the peak waveform amplitude was halved from half way through the transition out of the [æ]

until half way through the transition into the [æ] (in other words, the [æ] was louder than the other vowel).

6. −6 dB – flat pitch: The same as number 3, except that the VU reading was reduced by 6 dB in a manner similar to number 5.

These six different manipulations applied to each of the two formant patterns were played six times each to eight new subjects under conditions similar to those in Experiment 1. The only substantive differences were that the trial duration had to 5-sec fade in, 10-sec steady, and 5-sec fade out, to make the experiment less boring, and all stimuli were bandpassed between 45 and 3,500 Hz.

### Results

Since the scoring method used in Experiment 1 does not clearly indicate whether the response that appears with the highest average confidence is also actually given as the most confident one on any particular trial, the data in this experiment were scored both by the same method as Experiment 1 and also by assigning on each trial a score of 1 only to the response with the highest confidence rating.

The main result of Experiment 1 was confirmed, since regardless of whether the sounds had been equalized by peak or by energy, both scoring methods showed that each of the eight subjects had a higher score for stops than for other consonants with the step-pitch contour and a higher score for other consonants than for stops with a monotone. On the other hand, when amplitude discontinuities were introduced, there was no significant change in the relative distribution of stops compared to other consonants (by Wilcoxon tests).

A more detailed analysis was carried out using the same scoring method as in Experiment 1 and pooling of the data similar to those used in the first experiment. This analysis showed that the perceived manner of articulation (stop vs. other) changed significantly with pitch contour (flat vs. stop), $F(1, 7) = 101.0$, $p < .0001$, with no interaction of this effect with method of amplitude equalization ($F < 1$) or with formant pattern ($F < 1$). With the step-pitch contour, there was a significant change in the place of articulation of a perceived stop with change in formant pattern,

Table 1

*Confidence Ratings for Perceived Consonants in Experiment 2*[a]

| Pitch contour | Perceived consonant groups | | | |
|---|---|---|---|---|
| | b, d, g | w, r, l, y | m, n | Others |
| Flat | .8 | 4.1 | .1 | .3 |
| Step | 4.0 | 1.3 | 0 | .2 |

*Note.* Ratings have been summed within each group of consonants. Values are from 0 to 5, indicating increasing degrees of confidence.
[a] Averaged over formant pattern for sounds with either equal energy or equal peak amplitude.

$F(2, 14) = 14.5$, $p < .001$, although this interacted with the method of amplitude equalization, $F(2, 14) = 6.7$, $p < .01$, so that the change in place of articulation with changing formant pattern was more marked when the peak amplitudes were equalized (Condition 2) than when the energy was equalized (Condition 4). These effects are displayed in Tables 1 and 2. When only the amplitude or the energy of the sounds was changed, and the pitch kept flat, however, there was no significant overall increase in the report of stop consonants, although there was a weak interaction of perceived consonant manner with formant pattern, $F(1, 7) = 9.6$, $p < .025$, which is illustrated in Table 3. The interesting aspect of this amplitude-step data is the high ratings given to nasal consonants. The most confidently heard consonant in

Table 2

*Confidence Ratings for Perceived Stop Consonants in the Pitch-step Conditions of Experiment 2*

| Equal peak/energy | Consonant | | |
|---|---|---|---|
| | b | d | g |
| Pattern 1 | | | |
| Energy | 3.3 | .1 | 1.0 |
| Peak | 3.7 | 0 | 0 |
| Pattern 2 | | | |
| Energy | 1.7 | 2.3 | .3 |
| Peak | 1.2 | 3.3 | .1 |

*Note.* Values are from 0 to 5, indicating increasing degrees of confidence.

Table 3

*Confidence Ratings for Perceived Consonants in Experiment 2 for the Amplitude-step and Energy-step Conditions According to Formant Pattern*

| Formant pattern | Perceived consonant groups | | | |
|---|---|---|---|---|
| | b, d, g | w, r, l, y | m, n | Others |
| 1 | 1.4 | 2.7 | 1.7 | .2 |
| 2 | .9 | 1.5 | 3.3 | .1 |

*Note.* Values are from 0 to 5, indicating increasing degrees of confidence.

the two amplitude-step conditions of pattern 1 was /n/ and together with /m/ was frequently heard in the two amplitude-step conditions of pattern 2. This contrasts with all the other conditions of this and the earlier experiment, in which responses were very largely confined to the stops, liquids, and semivowels.

General Discussion

The impression that a stop consonant is present in a continuous formant pattern when abrupt changes in the pitch contour are introduced does not appear if abrupt amplitude changes are made instead. This indicates that concomitant amplitude changes were not responsible for the pitch effect that we found in Experiment 1. Moreover, a sharp reduction in amplitude is reported as a nasal consonant, an observation no doubt related to the lower amplitude of nasal murmur than of oral continuants in real speech. It seems unlikely, then, that amplitude discontinuities on their own can produce perceptual segregation, at least to the extent that it occurs for the pitch changes that we have used. This is reassuring, given the abrupt amplitude changes that are common in speech. It might be argued that an abrupt change in pitch can cue a stop consonant directly rather than, as we are arguing here, via auditory streaming. This seems unlikely for three reasons. First, our own impression while listening to the sounds (the subjects were not specifically asked) was that stops are heard as a consequence of the occurrence

of streaming, rather than independent of it. Second, while it is well established that a brief period of silence is a powerful cue to an intervocalic stop, even in the absence of any formant transitions (Bastian, Eimas, & Liberman, 1961; Lisker, 1957), changes in pitch have never been shown to have any effect on whether sounds are perceived as stops, although they have been shown to have a weak effect on voicing (Fujimura, 1961; Haggard, Ambler, & Callow, 1971). Third, the first few repetitions of the cyclical pattern are not generally (to our experienced ears) heard as stops. Like auditory streaming, the effect needs some repetitions before it appears, at least with the simple two-pitch patterns used here. Pitch continuity, therefore, emerges as a powerful determiner of the perceptual coherence of the speech of a single speaker.

Has the effect that we have demonstrated here under very artificial conditions any significance in situations more representative of real speech communication? Our sounds have all been multiple repetitions of a short formant pattern and a short pitch pattern, so that subjects can predict extremely easily what is coming next in the pattern. Our own observations indicate that, like the streaming found with pure tones, a number, albeit small, of repetitions are required before distinct perceptual segregation occurs. If the effect is relevant to natural situations, then we should be able to demonstrate that segregation can occur without the repetition of a fixed pattern of either formants or pitch movement. When random tokens of the three different vowels used in the first experiment were alternated with [æ] and suitable formant interpolations were carried out, then the impression of stop consonants being present when a repeating step-pitch contour was imposed was, to our ears, just as strong as when the formant pattern was entirely predictable. How the effect is influenced by differences in the pitch contour is currently being investigated; but our expectation is that rather brief exposure to a pitch movement should be enough to cause a pitch discontinuity to be interpreted as a different sound source. We are presently using pitch contours similar to those used by Heise and Miller (1951) to investigate this, and initial results indicate that a repeating-pitch pattern is not essential for a stop consonant to be perceived on a syllable whose pitch deviates from an established contour, although repetition or even previous exposure to the repeating pattern appears to augment the effect.

As we remarked earlier, abrupt pitch changes do not occur often in natural speech, and an intriguing possibility is that the factors that determine auditory streaming with both speech and pure tones are also to be found in production limitations. Perhaps we hear sequences of tones split into different streams because the human vocal apparatus is unable to produce such rapid alternations in pitch.

## Reference Notes

1. Dorman, M. F., Raphael, L. J., Liberman, A. M., & Repp, B. *Some masking-like phenomena in speech perception.* Haskins Laboratories Status Report on Speech Research, SR 42/43, 1975, 265–276.
2. Nooteboom, S. G., Brokx, J. P. L., & Rooij, J. J. de. *Contributions of prosody to speech perception* (IPO Annual Progress Report No. 11, 34–54). Eindhoven, Holland: Institute for Perception Research, 1976.

## References

Bastian, J., Eimas, P. D., & Liberman, A. M. Identification and discrimination of a phonemic contrast induced by silent interval. *Journal of the Acoustical Society of America,* 1961, *33,* 842.

Bregman, A. S., & Campbell, J. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology,* 1971, *89,* 244–249.

Bregman, A. S., & Dannenbring, G. L. The effect of continuity on auditory stream segregation. *Perception & Psychophysics,* 1973, *13,* 308–312.

Broadbent, D. E., & Ladefoged, P. On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America,* 1957, *29,* 708–710.

Cole, R. A., & Scott, B. Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology,* 1973, *27,* 441–449.

Darwin, C. J. On the dynamic use of prosody in speech perception. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception.* Berlin: Springer-Verlag, 1975.

Dorman, M. F., Cutting, J. E., & Raphael, L. Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *1*, 121–129.

Egan, J. P., Carterette, E. C., & Thwing, E. J. Some factors affecting multi-channel listening. *Journal of the Acoustical Society of America*, 1954, *26*, 774–782.

Fujimura, O. Some synthesis experiments on stop consonants in the initial position. *Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology*, 1961, *61*, 153–162.

Haggard, M. P., Ambler, S., & Callow, M. Pitch as a voicing cue. *Journal of the Acoustical Society of America*, 1970, 47, 613–617.

Heise, G. A., & Miller, G. A. An experimental study of auditory patterns. *American Journal of Psychology*, 1951, *64*, 68–77.

Kinney, J. A. S. Discrimination in auditory and visual patterns. *American Journal of Psychology*, 1961, *74*, 529–541.

Koffka, K. *Principles of gestalt psychology*. London: Kegan Paul, Trench, & Trubner, 1935.

Lisker, L. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 1957, *33*, 42–49.

Miller, G. A., & Heise, G. A. The trill threshold. *Journal of the Acoustical Society of America*, 1950, *22*, 637–638.

Norman, D. A. Temporal confusions and limited capacity processors. *Acta Psychologica*, 1967, *27*, 293–297.

Parsons, T. W. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 1976, *60*, 911–918.

Sayers, B. McA., & Cherry, E. C. Mechanisms of binaural fusion in the hearing of speech. *Journal of the Acoustical Society of America*, 1957, *29*, 973–987.

Thomas, I. B., Hill, P. B., Carroll, F. S., & Garcia, B. Temporal order in the perception of vowels. *Journal of the Acoustical Society of America*, 1970, *48*, 1010–1013.