1

The Magical Number Two and the Natural Categories of Speech and Music*

James E. Cutting

*Wesleyan University
and
Haskins Laboratories

Eleanor Rosch (1973) found that the Dani, a nonindustrial and nonliterate community in New Guinea, perceive certain colors and shapes in a manner functionally identical to American college sophomores. Her result is interesting because the Dani have no color terms other than those for light and dark and no terms for angular geometric figures. Her methodology is complex and not relevant to speech research; her discussion centers more on the general area of cognition than on perception, but her conclusion is central to my theme: there are salient stimuli in our environment which we perceive as prototypes of natural categories. In other words, our perceptual apparatus is geared to perceive certain stimuli better than others, and it warps a somewhat illfitting stimulus to be more like its natural prototype. Moreover, going somewhat beyond Rosch, there are distinct perceptual boundaries between these adjacent categories. The categories and boundaries are "natural" because they remain largely unmodified by learning or by environment. Rosch presented convincing evidence that natural categories exist in vision; here, I hope to demonstrate that they are prevalent in audition and are accompanied by equally "natural" boundaries. I will use findings of speech research to establish particular patterns of results indicative of "categorical" perception, and then search for them in music as well. Before presenting any data, however, I will discuss a theoretical framework in which to consider categories and boundaries.

THE MAGICAL NUMBER TWO IN SPEECH SOUNDS

For the last two decades a certain segment of the psychological community has been persecuted by an integer. The persistence with which this number plagues those of us interested in speech perception is far more than a random accident. There is, to quote a famous senator (and perhaps a more famous psychologist), a design behind it, some pattern governing its appearance. Either there really is something profound about this number or else we are all suffering from delusions of persecution. Our number, however, is not seven; it is two.

It is no mere trick that I choose to paraphrase the first paragraph of George Miller’s famous paper from *Psychological Review* (1956). Information processing has certainly burgeoned in the 20 years since his paper appeared, and talk of channel capacities and bits of information has since filled many books and articles. One may worry, then, that those of us interested in this smaller integer are somewhat misguided, if not stymied: perhaps each of us is only two sevenths of a proper psychologist, or perhaps our student subjects are only two-sevenths as bright as most. This is not the case (we hope). Whereas Miller is concerned with an upper limit of perceptual processing, we are interested in a lower limit. In addition, we are interested in the possible benefits derived from binary systems. In information-theory terms Miller is a three-bit researcher; we, on the other hand, are not even two-bit but rather one-bit researchers.

Psychologists and others have come very late to one-bit research, especially as it is relevant to language. Millennia before engineers and their computer science stepchildren thought in terms of binary electrical circuits, before physiologists discovered all-or-none neural firings, and before geneticists postulated dominant and recessive genes, Greek and Sanskrit grammarians were discovering the magical number two in distinctive features. These binary systems are fundamental to language: “the dichotomous scale is the pivotal principle of the linguistic structure” (Jakobson, Fant, & Halle, 1951, p. 9). Spoken language in particular is a house built on the number two (see Halle, 1957; Lane, 1966, 1967).

Consider some important binary oppositions in speech, using /ba/ as in bottle as a reference syllable. The slashes here act in a manner similar to the way a dollar sign denotes that numbers are American money: they indicate that the letters between them are spoken according to the International Phonetic Alphabet. It is reasonable that /ba/ should be considered a central utterance in a scheme of speech tokens. Unlike many speech sounds, the elements /b/ and /a/, and the syllable itself, are nearly universal to all languages of the world. A related syllable, /pa/ as in pod, is also nearly universal. Together, the two consonants /b/ and /p/ are a voiced–voiceless pair and differ only in the relative timing of the opening of the mouth and the initiation of pulsing in the larynx. For /ba/ the timing is nearly simultaneous in English, whereas for /pa/ there is a slight delay in the onset of voicing which is preceded by about a twentieth of a
second of whisper. This distinction is important because there is no speech sound, or phoneme, that is intermediate between /b/ and /p/.

Another binary pair is /ba/ and /ma/, which differ in manner of production: /ma/ is nasalized, /ba/ is not, but otherwise they are identical speech sounds. When a child says "I have a cold id by doze" we can appreciate the effect of clogged nasal passages on the neutralization of this phonetic distinction. A third pair is /ba/ and /da/, which differ in place of articulation: /ba/ is labial, produced at the lips, and /da/ is alveolar in English, produced by placing the tongue on the alveolar ridge behind the teeth. Just as there is no speech sound between /ba/ and /pa/, there is none between /ba/ and /ma/ and none between /ba/ and /da/.

Until World War II these distinctions were based on little more than 3000 years of intuition about the nature of speech production. Psychologists, wary if not skeptical of intuition and typically more interested in perception than production, did not become interested in speech until the invention of the sound spectrograph. This device transforms sound into a permanent visual record of time, frequency, and intensity patterns. (See Potter, Kopp, & Green, 1947, for elegant and detailed examples of sound spectrograms.) Shortly after the invention of this auditory-to-visual transform came its inverse, a device known as the pattern playback, which transforms a visual display into sound. Through a period of interactive experimentation with these two devices many of the important acoustic cues were discovered which separate speech sounds from one another (see Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967, for an overview). Schematic spectrograms of the four syllables of particular interest here are shown in Fig. 1. Since the three pairs are logically orthogonal they are

![Schematic spectrograms of /ba/ (as in bottle) and three other syllables whose initial consonant differs from /b/ along one phonetic feature.](image-url)
displayed as if in three-dimensional space. These examples are exactly like those used for the pattern playback, and would be highly intelligible (if somewhat metallic and "unnatural" sounding) when played through that device.

Observe the acoustic differences between the syllable pairs. Although all pairs are very similar, /ba/ and /pa/, for example, differ in two ways. In /pa/ the first formant, or dark resonance band of lowest frequency, has been cut back from stimulus onset by about 60 msec. Also, the excitation pattern of the second, or higher, formant has changed. Instead of being excited by a periodic glottal source in the larynx, it is excited by aperiodic or noisilike turbulences in the mouth cavity. Natural speech tokens would typically have a third and other higher formants. Whereas the third formant carries some important phonetic information, the fourth and higher formants carry little or none; the first two carry the bulk of the linguistic information load and suffice for these syllables. The syllable pair /ba/-/ma/ differs mostly in the addition of steady-state nasal resonances to /ma/. They extend from just before to just after the release of constriction at the lips (which creates the formant transitions). For /ma/, however, the first formant transition is less prominent. The differences between /ba/ and /da/ are perhaps the smallest and conceptually easiest to visualize of the three pairs. In /ba/ the second formant glides upward in frequency at syllable onset, whereas in /da/ the second formant glides downward by about the same amount. It will be instructive to consider this pair in more detail.

Identifying

Humans have little success in producing speech sounds intermediate between /ba/ and /da/. Computer-driven speech synthesizers, on the other hand, can easily be programmed to produce these unlikely sounds. When a seven-item continuum of utterances is generated from /ba/ to /da/, the syllables array themselves as shown in the left panel of Fig. 2. When these seven syllables are randomly ordered and presented many times, and when listeners identify each as either /ba/ or /da/, we find our first empirical manifestation of the magical number two. Complementary identification functions show discrete perceptual categories as seen in the upper-left panel of Fig. 3. These are actual not idealized data. Notice that the first three stimuli in that array are almost always identified as /ba/, and that the last three items are almost always identified as /da/. (Stimulus 4 is perceived as /ba/ about half the time and /da/ the other half.) The stimulus differences appear to be perceived in a discrete rather than continuous manner.

However, one should not be overly impressed with the quantal nature of these complementary functions. Imagine an array of lines tilted at various angles like that shown at the right of Fig. 2. If we "read" these lines from left to right, Stimuli 1 through 3 might be considered "ascending" and Stimuli 5 to 7 "descending." Increments of physical difference between members of this visual array are exactly equal in angular degrees, just as increments in the /ba/-to-/da/
1. NATURAL CATEGORIES IN SPEECH AND MUSIC

FIG. 2 Schematic spectrograms of a /ba/-to-/da/ acoustic continuum, and a display of a companion array of slanted lines.

auditory series are equal in slope change of the second-formant transition. When the visual stimuli are mounted on cards and viewers asked to classify each as ascending or descending, we find nicely quantized identification functions shown in the upper-right panel of Fig. 3, with only Stimulus 4, the true horizontal, not a member of either category. Clearly the auditory and visual results are similar, and nothing would appear to be peculiar about speech.

As a further demonstration that identification functions should not be overemphasized, consider what happens when we ask the same listener-viewer to

FIG. 3 Identification functions for an array of speech syllables and an array of slanted lines (shown in Fig. 2) in two conditions: one assigning the items to two categories and the other assigning them to three categories.
classify the continua into three categories instead of two. The speech-syllable choices here are /ba/, “ambiguous” (not convincing as either stop consonant), and /da/; and the slanted-line categories are ascending, horizontal, and descending. Results are shown in the lower panels of Fig. 3. Both classes of stimuli yield similar identification patterns, with the third categories supplanting the old boundaries in the two-category tasks. From these results speech perception would appear to be no different from the perception of objects and events in other modalities. Moreover, the magical number two would seem irrelevant.

Two statements must be made before entertaining the notion that these conclusions are legitimate. First, the two stimulus series in question were judiciously selected. Few acoustic continua generated by a speech synthesizer appear to have phoneme boundaries with near-zero slope in the second-formant transition: /ba/-to-/da/ is closer to being an exception than the rule. Deviations from the peculiar regularity in this syllable array, such as that found in a /bl/-to-/di/ acoustic continuum (as in beam to deem), would be much more difficult to model in a visual continuum. Second, we should consider the nature of the middle categories in each set of responses. Intuitively they seem quite different. The middle visual category would appear psychologically more real than its neighbors. Indeed, the terms ascending and descending are derived with reference to horizontal. The middle speech syllable category, on the other hand, is a tenuous if not bogus domain. Certainly /ba/ and /da/ are not derived perceptually with reference to an ambiguous stimulus which is difficult if not impossible to pronounce. In short, we see horizontal lines every day; we do not “hear” ambiguous speech sounds. Just as with a Necker cube, the percept flips one way or the other: it is either /ba/ or /da/, and rarely anything else unless one asks the subjects to perform the unusual task of “ambiguating” the syllables as I have done.

**Discriminating**

Given these clues that a /ba/-to-/da/ acoustic continuum is perceived somehow in a unique and quantal manner, we should look to a second and more important manifestation of the magical number two—nonlinearities in discriminability.

If a listener-viewer is asked to compare two members of one of the arrays of stimuli used thus far, how accurate are her responses? For purposes of uniformity both arrays of stimuli are presented in a sequential discrimination task: the first stimulus is presented, followed by a silent or blank interval of one second, followed by the second stimulus (either identical to the first or two steps removed along the physical continuum). In this manner, along with item pairs which are identical, Stimuli 1 and 3 are compared; 2 and 4; 3 and 5; 4 and 6; and 5 and 7. Subjects are asked to report whether the two items are the same or different. Only the “different”-pair results are of interest here and are shown in Fig. 4; few errors occur on “same”-pair discriminations in this type of task.
FIG. 4 Two-step discrimination functions for speech syllables and slanted lines, superimposed on their respective identification functions.

Notice the sharp discrepancy between the two darker functions. The speech syllable data, shown in the top panel, demonstrate a sharp peak in discriminability at the Stimulus 3–Stimulus 5 comparison which rapidly tapers to lower-than-chance performance at either end of the continuum. The slanted line function, on the other hand, is at or near 100% performance throughout the stimulus range. (Smaller increments between slanted lines could easily decrease performance level, perhaps even create a peak near the horizontal, but significant troughs are unlikely to occur.)

Comparing these discrimination results with the two-category identification functions superimposed upon them, we see that for the speech items there is a correspondence between the crossover of the complementary identification curves and the peak in the discrimination function. Labelability changes inversely with discriminability: Items can only be perceived as distinct from one another when they have different names. This nonlinearity lies at the heart of the interest in the magical number two, it is called categorical perception, and it is in sharp contrast to the performance on the same task for the slanted line stimuli. Acoustic differences between speech Stimuli 1 and 3 or 5 and 7, for example, are typically inaccessible to the listener. However, the magical number
two is not as absolute as it seems here, with two discrete categories and a perceptually distinct boundary between them. It is necessary to consider certain systematic deviations from strict twoness.

2. PLUS OR MINUS TWO FUDGE FACTORS

Just as George Miller has his small margin for error, a fudge factor of plus or minus two, we speech researchers also have ours. It is considerably smaller; it is difficult to scale down to size in terms of numerical deviations from our magical integer; in fact, it may be difficult to determine whether it is actually positive or negative. What is clear is that it manifests itself, roughly, in two “sizes,” one trifling and the other not so trifling.

The trifling deviation from the magical number two, and the strict categorical perception it implies, can be seen using the same /ba/-to-/da/ continuum. When listeners are asked to attend to Stimuli 1 and 3, for example, both of which are identified as BAH nearly one hundred percent of the time, and they are asked to judge which of the two is more BAH-like—or in Rosch’s terms, which is the prototype—they will typically determine that Stimulus 1 fits the bill. They will do this, however, only after they have laughed at the experimenter for asking them to perform such a ridiculous chore, only after she has reassured them that there really is a difference between the items, and only after she has cajoled and exhorted them to do the best they can. Even after these machinations their performance is rarely close to being perfect. This special tutoring of within-phoneme-class acoustic differences does not appear to transfer to other stimuli, and often is not useful as pretraining for other tasks with the same stimuli. The fact that subjects can report differences between two different tokens of /ba/ is more a testament to what the human perceptual apparatus can do in an unusual situation rather than what it does do in a normal situation. In all fairness, however, it should be noted that this deviation from the magical number two is more trifling in size than in theory. Its discovery by Barclay (1972) was a blow to some stricter views of speech perception.

The not so trifling deviation from the magical number two requires another set of stimuli, and it is theoretically even more important than the first fudge factor. Shown at the top of Fig. 5 are the endpoints of an acoustic continuum of vowels from /i/ as in heat to /I/ as in hit. Between them, one can easily generate five intermediate stimuli, thus creating a seven-item array with equal increments of acoustic change between all members. Here, instead of changing slopes of transitions, the frequencies of entire resonances are changed, increasing in value from /i/ to /I/ for the first formant and correspondingly decreasing in value for the second and third formants. (The addition of the third formant here increases intelligibility, but the array would yield nearly identical results without it.) When these items are randomly ordered and presented many times to listeners,
FIG. 5 Schematic spectrograms of /i/ and /I/ (as in eat and it), endpoints of a seven-item acoustic array, and the discrimination functions for that array compared against the slanted lines and stop consonants.

Results show quantal identification functions similar to those shown at the top of Fig. 3 for consonants and for slanted lines. That is, Stimuli 1 through 3 are identified as /i/, Stimuli 5 through 7 are identified as /I/, and only Stimulus 4 is ambiguous between the two. Discrimination results, however, reveal a pattern unlike those for consonants or for slanted lines. They are shown in the lower panel of Fig. 5.

Notice that the vowel discriminations lie intermediate between the previously discussed consonant and slanted-line functions. There is a "peak" in the function at the Stimulus 3–Stimulus 5 comparison but the "troughs," or regions of poor discriminability, are not nearly as "deep" (close to zero percent performance) as those for stop consonants. The reason for this appears to be that within-category differences between two tokens of the same vowel remain in short-term memory long enough for accurate comparisons to be made. Performance on these comparisons, however, is still worse than for those made at the phoneme boundary. The reason for the trough–peak difference is related to the way in which the information is encoded. For speech stimuli the height of the peaks in any discrimination function can be taken as a measure of the strength of phonetically coded information, and the depth of the troughs in relation to those peaks can be interpreted as the relative strength of acoustically coded...
information. Only the phonetic code is relevant to short-term memory as it is usually defined; the acoustic code fades much more rapidly. Differences between the consonant and vowel functions at within-phoneme-category comparisons are one assessment of the magnitude of this not so trifling deviation from the number two. Some raw acoustic information about vowels is available for comparison purposes; practically none is available about the consonants.

One feature of these discrimination results which I have not discussed so far is the effect of the duration of the silent interval between the two stimuli in the sequential discrimination task. For the vowel stimuli, this interval is vital for determining the depths of the troughs in the functions. If the interval is shortened from one second to a quarter of a second, listener performance on within-category comparisons of Stimuli 1 and 3 or of Stimuli 5 and 7 will increase to as much as 85 or 90%. If, on the other hand, this interval is lengthened to as much as two or three seconds, listener performance on these same comparisons will decrease to as low as 40–50%. There is no such effect of silent or blank interval on consonants, and probably none for slanted lines (although I have not done the experiment). For consonants in particular the within-category acoustic information appears to be lost prior to the onset of the second stimulus in a pair to be discriminated, regardless of how short that interval may be. Perception of consonants, then, is almost instantaneously phonetic. Practically no raw acoustic husk remains in memory. In Rosch’s terms, our perceptual apparatus warps all stop consonant stimuli which fall within a phoneme category, changing them into a perceptual prototype. Once that prototype is internally registered the nonprototypic vagaries of the stimulus are largely inaccessible to consciousness except through experimental exhortations like those discussed earlier. A more careful account of these phenomena and the extent to which they occur has been developed by David Pisoni (1973, 1975; Pisoni & Lazarus, 1974; Pisoni & Tash, 1974), and the interested reader should refer to those articles.

There are many psychological differences between stop consonants and vowels. I feel that it is incorrect, however, to say one class of phonemes is more “speech-like” than the other. Nevertheless, since the perception of stop consonants adheres more to the magical number two, it is those stimuli that we will consider in depth in this chapter. Before continuing with the stops, however, we must consider the possibility of categories and boundaries elsewhere in audition.

3 THE MAGICAL NUMBER TWO IN NONSPEECH SOUNDS

For years, those of us interested in speech and the number two (implied by the phenomenon of categorical perception) have called those auditory events which are not speech “nonspeech.” Underlying the use of this handy and linguocentrically biased term is the belief that speech is somehow different from all other
auditory events, just as it is different from the perception of slanted lines. Most of us, I think, still believe this to some degree. I do. But, whereas we used to be armed with an arsenal of empirical data supporting the uniqueness of speech processing, lately we have begun to strip ourselves of these findings. Central to our arsenal and seemingly most invulnerable to attack was categorical perception.

Shortly after the initial formulation of categorical perception—that it involves a discontinuity in discrimination functions for stimuli equally spaced along a physical continuum—arose the issue of "acquired distinctiveness." Expressed as a question in its simplest form: are the discrimination peaks learned for stop consonant stimuli? Do children, for example, acquire the distinct speech categories, or are they innate? A dozen years after the question first arose, it was answered conclusively and that answer is discussed in Section 4. At the time, however, it was not possible to test young infants, so the question was asked in another form: can nonspeech discrimination peaks be acquired through training? Perhaps the process of acquiring categories and boundaries follows a developmental trend: initially, all stimulus pairs may be equally discriminable to the untrained listener and her discrimination function would be "flat" and moderately above chance level throughout the range of the continuum; only later, with training, would a peak appear in this function, and perhaps the troughs would correspondingly drop within each category.

Harlan Lane, an early proponent of this view, found that certain subjects listening to certain complex nonspeech sounds (spectrographically inverted speech patterns) acquired discrimination peaks through a simple training procedure (Lane, 1965). Pisoni (1971), in a careful replication with similar stimuli, found this to be true for a few selected subjects, but generally not true. Moreover, Studdert-Kennedy, Liberman, Harris, and Cooper (1970) found even Lane's selected data unconvincing: whereas there were peaks in his functions, Lane found few deep troughs and less correspondence between discrimination and identification functions than would be desired. As seen when comparing the perception of vowels and consonants in Fig. 5, troughs are vital. If the distinctive nature of the peaks can be acquired through training, and it is not entirely clear that they can be, the troughs do not appear to be learned: there seems to be no pneumatic tradeoff between the acquisition of peaks in training and the loss of ability to discriminate within a category.

But the Lane and the Pisoni stimuli were not "natural" nonspeech sounds; nor are sine wave tones and other more familiar psychoacoustic stimuli "natural" in any real sense. Are there commonly occurring stimuli in our environment which are perceived categorically and which obey the laws of the Magical Number Two? An obvious candidate here is musical sounds. They are natural at least to the extent that they rely on simple mechanical action of easily fashioned materials. Locke and Kellar (1973) varied the middle component of triadic chords in search of categorical perception, and found some categorical ten-
dencies in musically trained listeners, but few in musically naive listeners. Their results were promising, but had two important drawbacks. First, the discrimination functions for the musically trained listeners were more similar to the vowel function shown in Fig. 5 than the typical stop consonant function beneath it. Second, and more damning, is the fact that extensive musical training seemed to be a requisite for even these vowel-like functions. Again, we are back to “acquired distinctiveness,” and to the lack of sufficient troughs in the discrimination functions.

More recently, categorical perception has been found in a musically relevant dimension and the results meet all the requirements for binary processing according to the laws of the magical number two (Cutting and Rosner, 1974; Cutting, Rosner, & Foard, 1976). The dimension is that of attack, or rise time. Rapidly rising sawtooth waves, for example, sound like the plucking of a stringed instrument, such as a guitar; more slowly rising sawtooth items sound like the bowing of a similar instrument, such as a violin. Oscillograms of token “pluck” and “bow” sounds are shown in Fig. 6. Rise time cannot be systematically varied when playing actual musical instruments, but it can be varied readily on a Moog synthesizer. We chose to vary the rise time from 0 to 80 msec in 10-msec increments for several continua of musical sounds. Note that such

**FIG. 6** Oscillograms of token pluck and bow stimuli, compared against token CHA and SHA items. Both pairs of stimuli differ in rise time.
variation is minor in magnitude compared to the long and tapering offset of the stimuli.

When these stimuli are placed in the same paradigms mentioned previously for the speech syllables /ba/ and /da/, they yield remarkably similar results. Sawtooth wave items with less than 40-msec rise time are identified as pluck nearly 100% of the time, stimuli with rise times greater than 40 msec are identified as bow nearly 100% of the time, and only the 40-msec rise time stimulus is ambiguous (identified as pluck about 40% of the time and as bow 60% of the time). When these stimuli are placed in the sequential discrimination task, items within a category sound identical to the listener: Asked to judge whether pairs of stimuli are the same or different, listeners report that items of 10- and 30-msec rise time and of 50- and 70-msec rise time are the same item more than 75% of the time, well below chance. Only when items with 30- and 50-msec rise times are compared do listeners perform well, and here they make only about 15% errors. These results clearly indicate that there are quantal identification functions for pluck and bow sounds, that there is a peak in the discrimination function lying astride the crossover point of the complementary identification functions, and that the discrimination function falls off into deep troughs at either side of the peak. Thus, the magical number two reigns in nonlinguistic domains as well as in speech.

Six other aspects of our data are important concerning categorical perception of these musical sounds. First, the time interval between the items in a discrimination pair matters not at all: performance on within-category pairs, for example, is no better when the two stimuli are separated by 250 msec than when they are separated by nearly 2 sec. These results strongly suggest that the musical sounds adhere to the magical number two more strictly than do vowels and that the "fudge factor" is likely to be of the trifling rather than the not-so-trifling size.

Second, the results do not appear to be mediated directly by the labels pluck and bow. In our initial study we were careful to administer the test in two conditions. In one we carefully tutored subjects in the use of the terms pluck and bow, playing extreme tokens from the continuum several times before they participated in the identification test. The identification test gave them an additional fifteen minutes of practice using the terms before they listened to their first discrimination pair. In a second condition the subjects started right away with the discrimination test, did not hear practice items, and were not told of the labels pluck and bow. The results for the two groups were essentially identical and suggest that the labels pluck and bow have little to do with the perceptual process.

Third, one might think that our results are robust because the stimuli emulate common musical sounds or because the stimuli have complex spectra. Would such results occur for simpler auditory events varied in the same manner which do not sound like convincing tokens of musical sounds? We varied rise times of
sine wave stimuli and found the same pattern of identification and discrimination results as for the sawtooth items. These results indicate that the perceptual process involved here is more fundamental than just a music-processing system might be. Sine waves with rapid rise times sound vaguely like a flute played staccato style, but sine waves with less rapid rise times are not at all convincing as notes from a flute played in a more legato style.

Fourth, one may be suspicious of our stimuli since we played them over loudspeakers and earphones. Because certain members of the musical arrays have very rapid rise times, they are likely to induce clicks into the transduced signal; that is, the response characteristics of the broadcasting devices may be sluggish enough to produce audible short bursts at the beginning of the items. One fear is that the presence or absence of such clicks might correlate perfectly with the perceptual categories: pluck equals presence of a click, bow equals no click. To insure that this artifact did not account for our results we inspected our stimuli after they were played through a loudspeaker, redisplaying them with high resolution on a computer-controlled oscilloscope. We found such small perturbations in the signals only for the 0- and 10-msec rise time stimuli. However, items with 20- and 30-msec rise time, which were members of the same perceptual category, did not have these irregularities.

Fifth, we found that the long tails on the pluck and bow stimuli are necessary for the successful identification and discrimination of the items. When the items are trimmed to 250 msec in duration by simply lopping off the last 750–850 msec of the items, identifiability and discriminability are markedly impaired. At first, this seemed rather bizarre to us: the final three quarters of the stimulus does not appear to carry any information about stimulus onset and would seem unnecessary for maintaining musical integrity of the sounds. This view is clearly incorrect. Moreover, we should not have been surprised at these results. When speech items such as /ba/ and /da/ are severely trimmed from 300 msec to 40 or 50 msec, removing only the steady-state vowel, they often cease to sound like speech stimuli and are unlabelable by many listeners. The integrity of the syllables is thus violated, and in a manner similar to the truncation of our pluck and bow items.

Sixth, and perhaps most interesting of all, is the fact that rise time is not only a cue for the distinction of musical items, but it is also used as a cue in speech: CHA (/tʃa/) as in chop has a very rapid rise time in its fricated (or noise-like) portion, whereas SHA (/ʃa/) as in shop has a much more gradual onset. Tokens of these speech syllables are shown in Fig. 6 next to the pluck and bow items. When an array of these syllables is generated on a computer-driven speech synthesizer, and when they are inserted into the same paradigms as we have discussed thus far, listeners yield patterns of identification and discrimination results which are nearly identical to those for the pluck and bow items. We find it compelling that a single cue, rise time, is used to distinguish categories inside and outside of speech. While speech production and speech perception are
unique to man we should not expect all speech processing mechanisms to be unique as well. In evolutionary terms it would have made sense to build a speech processing system on underlying and already existing auditory faculties. It seems reasonable that at least some of the binary distinctions on which speech is built would be based on binary auditory distinctions. We suggest rise time is one of them (Cutting & Rosner, 1974).

To account for discontinuities in the discrimination functions of stimulus arrays /ba/-to-/da/ and /i/-to-/I/, many speech researchers have thought in terms of phonetic and auditory memories. The peaks in these functions have been taken as a measure of the strength of phonetic memory similar to the more commonly known short-term memory; and the troughs, in relation to the peaks, are taken as a measure of an auditory memory similar to what is often called “echoic memory.” The categorical perceptions of pluck and bow stimuli jar this view somewhat. The notion of an auditory memory accounting for the troughs remains unchallenged. However, the notion that a phonetic memory underlies the peaks in a discrimination function must be cast aside. The peak in the pluck-and-bow discrimination function can in no way be thought of as phonetic. Instead this higher-level memory may be reserved for highly coded decisions about auditory signals; pluck versus bow or /ba/ versus /da/ would both qualify here. It is relatively easy to understand why speech sounds are categorical and coded into a phonetic string rather than left as raw acoustic information. The memory storage capacity required for one second of high-quality speech (such as reproduced on a tape recorder) would be 40,000 bits of information, whereas the storage capacity required for one second of phonetically coded speech would be only about 40 bits of information, plus the necessary subroutines to decode that string (Liberman, Mattingly, & Turvey, 1972). Clearly, in terms of a thousand-to-one savings in storage capacity, it makes sense to code speech phonetically and at a rather rapid rate so that it can be comprehended. The rub, however, is to understand why the same system appears to code and categorize musiclike sounds when such a saving may not be needed. The answer is necessarily indirect, and must first take us back to the notion of “natural” categories and their apparent function.

4  NATURALNESS OF THE MAGICAL NUMBER TWO
IN SPEECH AND MUSIC

Thus far I have presented what should be compelling evidence of discrete categories in speech and music, but I have said nothing of their naturalness. Rosch concluded that certain color and shape categories are natural because they appear to remain largely unmodified by presence or absence of language terms for them. To find exactly parallel results for speech items is difficult. Speech syllables have the unique property of providing their own nonarbitrary labels:
/ba/ is BAH and /da/ is DAH, and they are pronounced and labeled as such by (almost) all peoples of the world. We are, therefore, forced to use a different technique for assessing naturalness of speech categories and boundaries, and we must use the same method for those in music.

The *Oxford English Dictionary* defines *natural* as: "present by nature; innate; not acquired or assumed." If speech categories are natural by this definition, all humans should be born with the ability to use them. One approach for determining whether or not the perception of these categories is innate is to test young infants. For reasons of practicality the subjects were from one to four months old. The assumption here is that these infants will have had little if any opportunity to learn much about their to-be-native language and that any results which they yield are characteristic of those capabilities which are genetically "wired in."

**Speech Categories**

It should be clear that one cannot ask infants to identify /ba/ and /da/. Young children typically cannot produce such differences in a systematic and controlled fashion until they are many months older. Therefore, it is out of the question to try to obtain identification functions. One can determine, however, whether infants can discriminate speech sounds and discriminate them in a manner that approximates that for college-age subjects. This is exactly the approach of Peter Eimas and his colleagues (see, for example, Cutting & Eimas, 1975; Eimas, 1974; Eimas, Siqueland, Jusczyk, & Vigorito, 1971) in a series of pioneering studies.

It is one thing to ask an infant to discriminate two speech sounds, but it is another to pose that question in a manner for which he or she can give a suitable and measurable response. Eimas has used a conditioned nonnutritive sucking procedure; others have used heart rate (Morse, 1972). In the Eimas and Siqueland procedure the infant is given a hand-held nipple upon which to suck. Instead of transducing nutrients, it transduces pressure to a pressure-sensitive apparatus which, in turn, triggers, for example, the speech sound /ba/. The more frequent the high suction responses of the infants, the louder (or, in another procedure, the more frequent) the speech sound is presented against background noise. The infant quickly learns this association and is quite willing to make several hundred sucking responses over the course of about 10 min merely to hear the same sound repeated. The time–frequency course of the infant's responses is of particular interest. Over the course of about three minutes after the initial learning of the association, the infant increases her responses to peak of as much as 50 or 60 per minute, well above a preassociation nonnutritive baseline. Shortly thereafter the infant seems to tire of the situation and responses taper off rather dramatically in the following 2 min.
After a drop in responses of at least 30%, one of three things happens to the infant. In one control condition the infant continues to hear the same stimulus, Stimulus 3 in Fig. 2, say, over and over again. Responses here continue to approach asymptote at or below the baseline rate. In the experimental condition, however, the stimulus is shifted to /da/ (Stimulus 5) and the infant’s responses begin to increase again, only starting to fall the third or fourth minute after the stimulus change. Most important is the second control condition. Here the stimulus shifts from Stimulus 3 to Stimulus 1, but both are identified as /ba/ by adults. In other words, this change is physically just as great as the across-boundary shift, but both stimuli lie within the same category. As in the first condition, the infant’s responses continue to approach asymptote. All these trends are shown in Fig. 7.

Three aspects of these infant data are interesting when compared with the adult data discussed in previous sections. First, the across category stimulus shift in the experimental condition here corresponds to the peak in the discrimination functions seen in the top panel of Fig. 4. In the infant’s case, the breaking of the sucking response habit is taken as evidence that she perceives that a new stimulus has been presented, one which deserves more attention and subsequently more sucking responses. Hence, like adults, infants as young as one month can perceive phonetically relevant features. Second, the within-category stimulus shift in the crucial, second control condition corresponds to the troughs in the adult discrimination function. Continued habituation of the sucking response is taken as evidence that the infant did not perceive that a “new”

![Graph](image)

**FIG. 7** Schematic data display of the time course of infant sucking responses to auditory stimuli in the Eimas paradigm.
stimulus had been presented. Just as for the adult, the infant may have merely regarded the second stimulus as identical to the preshift stimulus. Thus, like adults, infants as young as one month cannot perceive phonetically irrelevant changes in acoustic features even when they are identical in magnitude to the across-category, phonetically relevant change. Third, and more speculative than the previous two points, is that the difference between the functions for the no-shift and within-category-shift conditions suggests that even for very young infants there is a trifling-sized fudge factor that modifies the magical number two. Although the difference between these two groups has never been significant in a single study, the trend is unmistakable: the infants in the within-category-shift condition attenuate their habituation rate slightly more than the no-shift group.

Do infants perceive categorically the same speech continua that adults perceive? It appears that they probably do, and maybe even more so. For example, they yield results functionally identical to those in Fig. 7 when discriminating a voice-onset time continuum from /ba/ to /pa/, and when discriminating /ra/ from /la/ cued only by changes in the third formant transition (Eimas, 1975). This second result is important since the difference is one which native Japanese-speaking adults cannot perceive and do not have in their language. The notion is a tantalizing one that infants may be superior to adults in perceiving certain speech-relevant dimensions. It suggests that, while it is true that the distinctiveness of speech categories is not acquired by a learning process, it may also be true that certain potential distinctions are lost when unused by the developing child. One might consider this process "acquired indistinctiveness." It might even appear to support some of Lane’s (1965) original contentions. It does not, however, since it is not the troughs of the discrimination functions which get deeper (indistinctive) but rather the peaks themselves which disappear.

Music Categories

Are pluck and bow sounds discriminated categorically by infants? They are, as we have recently discovered (Jusczyk, Rosner, Cutting, Foard, & Smith, 1975). The stimuli used in this study were selected from those used in the adult studies (0-, 30-, and 60-msec rise-time items) plus an additional stimulus from the original set (with 90-msec rise time). Each of 18 two-month-old infants was studied under two conditions: all participated in a condition involving the cross-boundary, 30- and 60-msec items, and six infants each were involved in three control conditions. One group performed the no-shift control. That is, each infant continued to listen to the same stimulus throughout the experimental session. A second group listened to the 0- and 30-msec rise-time items, and the third group listened to the 60- and 90-msec rise-time items. Counterbalancing of pre- and postshift stimuli was observed, as well as counterbalancing the
order of the experimental (30–60 msec) and control conditions (no shift, 0–30 msec, or 60–90 msec).

Results were compelling. Seventeen out of the 18 infants demonstrated a higher sucking response rate in the cross-category shift condition (30–60 msec) than in the control condition, regardless to which of the three groups she belonged. The general patterns of responses plotted over time was functionally identical to those shown in Fig. 7: Habituation of the response continued in the no-shift condition and in both within-category shift conditions, while dishabitu-ation occurred for the across-category shift condition.

The adult boundary for these sawtooth wave pluck and bow stimuli is at about 35 to 40 msec rise time; it is clear that the infant boundary is near this same mark. It should also be clear that both speech categories and "nonspeech" (music-like) categories and boundaries are innate to humans: one can think of little evidence supporting the notion that these young infants could have acquired the distinctive categories demonstrated here. Thus, these categories and boundaries appear to be natural according to the strictest possible psychological interpretation of that word.

An interesting question now arises. These categories and boundaries seem innate to humans, but are they innate to other animals (primates) as well? In other words, is there any evolutionary continuity in the development of the mechanisms behind these effects? Morse and Snowdon (1975), for example, have tested rhesus monkeys in the discrimination of speech syllables such as those discussed here. Their results do not support a strong form of categorical perception in infrahuman primates, but there are some categorical tendencies. Would rhesus monkeys discriminate pluck and bow sounds in a manner as striking as adult and infant humans? Alas, we do not know as yet. Results here will be interesting regardless of the outcome. If rhesus monkeys (or higher nonhuman primates) do discriminate these musiclike sounds, we will have evidence that the perception of rise time, which is used to cue certain speech categories, has an evolutionary history older than that of speech. This would support the view that certain categories in speech were built on preexisting nonlinguistic categories (Cutting & Rosner, 1974). If the nonhuman primates do not discriminate pluck and bow sounds, we will have evidence that the perception of rise time, and the categories and boundaries in speech and music that it cues, evolved relatively late. Moreover, waxing slightly toward the philosophical, music perception (and subsequent music appreciability) has a recent evolution of its own.

In this first two-thirds of this chapter I have presented considerable evidence to demonstrate the existence, and the naturalness, of certain perceptual categories and boundaries in speech and music. I have not attempted to be overwhelming in breadth: few stimulus dimensions in speech were discussed and only one in music, partly because of the lack of knowledge in this relatively new field.
In the rest of the paper I will consider the nature of the mechanisms behind these manifestations of the magical number two.

5 MECHANISMS BEHIND THE MAGICAL NUMBER TWO

Very recently a new paradigm has taken the world of speech perception by storm. The technique is known as selective adaptation. Its roots are in vision research and in the mapping of the architecture of brain-cell function during perception. However, it is logically similar in many ways to the much older and better known phenomena associated with visual afterimages. I will use those photochemical data as a basic framework for discussing the perceptual effect in speech perception.

It is commonly known that if a person stares at a patch of blue color for about 15–30 sec, and then stares at an illuminated white wall, she will see a patch of yellow color with the same contour as the original blue patch. This effect is known as a chromatic afterimage. It is best explainable in terms of opponent process mechanisms, as first championed by Hering. Blue is thus viewed as the opposite color to yellow, at least at some stage of photochromatic analysis subsequent to excitation of the cones on the retina. Blue and yellow appear to synapse, if you will, onto the same cell bodies, but one color in an excitatory fashion and the other in an inhibitory fashion. Staring at a blue patch for many seconds fatigues the visual system so that when given a neutral stimulus (white), the viewer will perceive for a brief period of time that stimulus as being the opposite color (in this case, yellow). In a reciprocal fashion, staring at a yellow patch will yield a sensation of blue when the viewer is presented with a postadaptation neutral field.

One cannot stare at a speech syllable. Thus, the experimental adaptation situation must be changed to a degree for such auditory sounds. Since the auditory signal fades rapidly (which Hockett, 1960, for one views as a blessing) it must be presented over and over, perhaps as many as 100 or 200 times, to continually refresh the perceptual “image.” If this adapting stimulus is /da/, for example, and one is presented with a neutral stimulus near the /ba/-/da/ boundary (Stimulus 4 shown in Fig. 2), the listener will perceive that neutral stimulus as being a good exemplar of /ba/. One may view the stimuli /ba/ and /da/ as being “opposites” of one another along the dimension of change in the second-formant transition. Hence, just as yellow is the opposite of blue, /ba/ is the opposite of /da/, and after adaptation a neutral stimulus between the two prototypes will be perceived as being a member of the opposite class. Moreover, after adaptation the physical domain of the category of the adapting stimulus shrinks, while the physical domain of the unadapted stimulus category expands to fill the void.
The adaptation paradigm as used in speech perception is an exhaustive one. Typically, the listener is presented with more than 100 tokens of one of the end-point stimuli in a speech continuum (in this case the Stimulus 1 /ba/ or the Stimulus 7 /da/) and then given a brief identification test of the array of stimuli. This postadaptation test may consist only of one token of each of the seven stimuli in the array presented in random order. The subject identifies each of these items as /ba/ or /da/. Then another long series of adaptation presentations begins using the same stimulus as before. After this sequence a second seven-item series of stimuli is presented for the listener to identify. This cadence may be repeated a dozen times or more before the experimental session is completed.

Typical results are shown in the top panel of Fig. 8. Given an adapting item such as the Stimulus 7 /da/, the number of /ba/ responses to all members of the array tends to increase. Plotting only the number of /ba/ responses for each stimulus in the array (unlike the complementary plots in Figs. 3 and 4) one sees that the crossover point or 50% response level has shifted toward the /da/ end of

![Graph showing adaptation](image)

**FIG. 8** Schematic identification and discrimination data before and after adaptation with the Stimulus 7 /da/.
the continuum. In particular, Stimulus 4, which is normally identified as /ba/ on only about 40% of all trials in a preadaptation identification sequence, is not identified as /ba/ on more than 95% of all postadaptation identification trials.

At least two important aspects distinguish the adaptation effect with speech stimuli from the chromatic afterimage. First, it lasts a much longer time: the chromatic afterimage may last for only about 30 sec or so, whereas the shift in the /ba/-/da/ identification function may last up to a few hours or even longer. This result is directly linked to the second difference between the phenomena. The chromatic afterimage does not transfer from one eye to another. That is, if one looks at a patch of blue with only the left eye open and then stares at a white wall with only the right eye open, there will be no afterimage. This simple demonstration indicates that the locus of the chromatic effect is peripheral, or very near the retina and certainly before the neural pathways of the two eyes first converge in the lateral geniculate body. The adaptation effect with speech stimuli does transfer from one ear to the other and generally maintains its magnitude. This result indicates that the locus of the effect is central, and occurs after the pathways of the two ears converge (as low in the system as the superior olivary complex or as high as the cortex). These two factors—the duration of the effect and its locus—make it more similar to the visual work done in the 1960s by McCollough (1965) and by Blakemore and Campbell (1969), than to work with chromatic afterimages done originally in the nineteenth century. The adaptation work done in the field of speech perception, like the infant work discussed in Section 4, was pioneered by Eimas (Eimas & Corbit, 1973; Eimas, Cooper, & Corbit, 1973) and has been reviewed recently by Cooper (1975).

Several other aspects of speech adaptation are important and are closely related. First, the speech results have been interpreted in terms of feature adaptation. Features are thought to be processed by perceptual decision mechanisms. Adaptation shifts here contrast with the similar response shifts resulting from changes in adaptation level (Helson, 1964), since the latter may be accountable in terms of cognitive decision mechanisms. Second, these features are binary: that is, they are neurological correlates of the Magical Number Two. Third, these features are often thought of as phonetic in nature, that is, as linguistic rather than as auditory. By extension they have been thought unique to language. These three points deserve elaboration.

Feature Analyzers as Perceptual Mechanisms

A major thrust of the first speech adaptation study (Eimas & Corbit, 1973) was that the apparent shifts in the identification functions were not attributable to response bias or other "conscious" shifts in decision criteria. As proof of this position one must find that not only do identification functions shift, but that corresponding discrimination functions also shift and by the same extent.
Indeed, Eimas and Corbit (1973; and later Cooper, 1974) found that the discrimination functions do shift and by the anticipated amount.

Although their data did not deal with /ba/ and /da/, I may legitimately generalize as follows. In a preadaptation identification condition, Stimuli 1–3 are identified as /ba/ and Stimulus 5–7 as /da/. After adaptation with the Stimulus 7 /da/, Stimuli 1–4 are now identified as /ba/ and only Stimuli 6 and 7 as /da/. In other words, whereas Stimulus 4 is the most ambiguous item in the preadaptation condition, Stimulus 5 is most ambiguous in the postadaptation condition. If this change in identifiability is perceptual in nature, a preadaptation discrimination peak should be at the Stimulus 3–5 comparison, just as we have seen in Fig. 4, whereas the postadaptation peak should be at the Stimulus 4–Stimulus 6 comparison. This is exactly the type of result found by Eimas and Corbit, and is shown in the lower panel of Fig. 8. These are schematic, not actual, data since these authors used different stimuli, but they accurately reflect their results. The categories of the Magical Number Two change their locus with regard to the physical continuum, but they do not appear to change in any other manner.

It should be noted that postadaptation discrimination data are extremely difficult to gather. Only one or two discrimination trials are given after each long sequence of adapting stimuli. Thus, the task is very time consuming and only the most dedicated subjects will listen to the many hours of nonsense syllables, tediously and incessantly presented over and over and over again.

Feature Analyzers as Neural Mechanisms

Underlying the shifts in identification and discrimination functions are some allegedly quite simple mechanisms. A scheme of how they might work is shown in Fig. 9. Imagine two detectors in the perceptual system, one whose primary job it is to respond to the phoneme /b/ and the other to respond to /d/. Each of these is maximally sensitive to a prototypic stimulus (perhaps Stimulus 1 for the /b/ detector and Stimulus 7 for the /d/ detector). In addition, each will respond to other neighboring stimuli as well, but at a somewhat reduced rate. The /b/ and /d/ detectors are relatively "close" to one another in that they can respond to the same stimulus provided that the stimulus is roughly midway between the two stimulus prototypes along the physical continuum most relevant to the phonetic distinction, in this case the second-formant transition. Normally, the boundary between /b/ and /d/ is at the crossover point of the sensitivity curves, as shown in the top panel. The figure is drawn to be reminiscent of hypothetical signal detection functions and of a simplified one-dimensional rendition of Selfridge and Neisser's (1960) Pandemonium model of pattern recognition: At some neural level subsequent to the detectors themselves a decision demon will decide which feature demon, that for /b/ or /d/, has yelled the loudest (which
neuron has fired the most rapidly) and deserves to be recognized and identified over the pandemonium of screams (neural firings) of all the other demons (feature analyzers). This ultimate decision determines the psychological identity of the stimulus.

During extensive adaptation to the same stimulus, repeated over and over, the particular feature analyzer may fatigue. The precise nature of the fatiguing process is not known, but one possibility is shown in the bottom panel of Fig. 9. After adaptation with /da/, /d/ analyzer may become less and less sensitive to stimulation by the /d/ prototype and to all similar stimuli which would normally trigger it. The decrease in sensitivity may manifest itself in several ways. First the height of the sensitivity curve may decrease, then the shoulders of the function would slump inward. The effect would be that the new crossover point of the two feature analyzers would be moved slightly toward /d/, away from the old boundary. The new crossover would mark the locus of the postadaptation boundary between /b/ and /d/. A refractory period of a considerable amount of time would be necessary to restore potency to the /d/ analyzer. Once restored,
however, it would resume the sensitivity function shown in the top panel and consequently the old phoneme boundary would be restored as well.

This simple account seems to serve well in explaining shifts in identification and discrimination functions. Nevertheless, it is not without its problems. Consider one of the central aspects of this model: During adaptation the feature analyzer becomes more and more insensitive to the prototype of the stimulus category. If this were true, we would expect to find some way to measure the decrement in sensitivity to the stimulus prototype. The data in identification functions provide no help. Each function necessarily asymptotes at zero and at 100% by the time it reaches the prototype at Stimulus 1 or Stimulus 7. Hence we are restricted by floor and ceiling effects, and can make no inferences about any alleged attenuation in the sensitivity function. In a pilot study, Michael Posner and I supposed that a reliably more sensitive measure, that of reaction time, might serve to demonstrate the anticipated effect. Thus we adapted listeners to one of the endpoint stimuli in this /ba/-to-/da/ array, and measured their reaction time in responding to (identifying) each member of the array including the endpoint stimuli. The results seemed very clear. Indeed, there was a shift in the identification functions, just as we had anticipated, but there was no significant change in reaction time for the identification of stimulus prototype used in adaptation. An increase would have reflected, we thought, the decreased sensitivity to the category prototype; a decrease in reaction time might have reflected something else. No change is more difficult to interpret.

Perhaps, then, the sensitivity function shown in the bottom panel of Fig. 9 is incorrect. Perhaps rather than decreasing in “height” the function merely becomes more leptokurtic—that is, just as tall, but much slimmer. This would have the same effect in shifting the identification and discrimination functions. Regardless of the final form of the model, it probably does not differ greatly from the one presented here and should serve to account for all data. I return to this model later and discuss it in some detail from a different perspective. What must be considered now is the role of these feature analyzers in language, and in music.

Feature Analyzers as Linguistic Mechanisms, and as Auditory Mechanisms

Linguists have talked of the binary features of language for a long time and the influence of Jakobson, Fant, and Halle’s little book Preliminaries to Speech Analysis (1951) has been very influential in investigating the acoustic basis for these features. The notion of features was so well developed in the 1950s and 1960s that this term for various stimulus aspects of speech may have been borrowed by neurophysiologists when they discovered neural devices in the visual cortex which responded solely to edges moving in certain directions. The psychologist interested in speech and in effects of adaptation might look to
linguists and see one use of the term *features*, look to neurophysiologists and find another use, and then yearn to close the gap between the two. Eimas and Corbit (1973) found this link and described it in their seminal paper "Selective Adaptation of Linguistic Feature Detectors." But were these detectors linguistic in nature, or not? There is a difference between a linguistic feature detector and a linguistically relevant feature detector. A linguistic feature detector would be one which responds only to speech sounds and does not respond to similar nonspeech sounds. A linguistically relevant feature detector, on the other hand, would respond to certain speech sounds but would also respond to relevant nonlinguistic sounds as well. As yet it is too soon to determine conclusively whether or not certain of these mechanisms are linguistic or merely linguistically relevant. Let me first present some evidence which supports the latter view, then evidence for the former.

After the establishment and replication of the phenomenon, the question arose of the extent to which these underlying unknowns were linguistic. That is to say, while there may be specific phonetic feature analyzers, are there general phonetic feature analyzers? The difference between specific and general is crucial. Specific detectors would be sensitive only to those aspects of a particular speech sound (/d/ for example) which occur in particular speech contexts (as in /da/). General detectors, however, would be sensitive to broader aspects of that speech sound /d/ as it occurs, for example, not only in /da/ but in /di/, /du/, and /ad/ as well (as in deep, dupe, and odd). In other words, will the effect, as measured by shifts in identification functions pre- and postadaptation, transfer across different vowel environments? Will it transfer across different syllable positions? The answers are yes and no, respectively. Ades (1974) found that adaptation with /de/ (as in date) shifted the identification function of a /bae/-/dae/ continuum (as in bad, dad), but less readily than did adaptation with the endpoint /dae/ stimulus. He also found that adaptation with /dae/ had no effect on the identification of an /aeb/-/aed/ continuum. From such results the adaptation effect seems phonetic enough to transfer across small acoustic differences such as those seen when changing between similar vowel environments, but not phonetic enough to transfer across widely discrepant acoustic forms such as those found when shifting a phoneme from initial to final syllable position. Restated, the effect is general enough to transfer across small acoustic differences, but not general enough to transfer across large differences.

The fact that acoustic differences matter at all is an important issue: the postadaptation function for /bae/-/dae/ is shifted considerably by adaptation with /dae/, less by adaptation with /de/, and even less (that is, not at all) by adaptation with /aed/. Rather than account for this array of results by phonetic feature adaptation one might account for the results in terms of acoustic or auditory feature adaptation. (I will use the term auditory so as to be more general than the term acoustic might allow. I will consider not only those acoustic features that are analyzed logically prior to the labeling of speech
sounds but also entertain the possibility of a higher-level feature analysis of an auditory signal, which would appear to occur beyond the registration of the raw acoustic signal but would not necessarily involve language. Auditory seems the best term here, implying nonlinguistic as well as postsensorily. Before asserting that auditory feature analyzers are possible alternatives to phonetic feature analyzers, one must demonstrate that the notion of auditory analyzers is viable. Here, we have only to look to the pluck and bow sounds again.

We (Cutting, Rosner, & Foard, in press) selected a 440-Hz (Concert A) sawtooth wave continuum to use for postadaptation identification. Items differed in rise time in 10-msec steps from 0 to 80 msec. Rather than just two, there were eight adaptation conditions: adaptation within the same continuum of sounds using the 0- and the 80-msec rise time 440-Hz sawtooth items, which we take to be our “prototype” pluck and bow stimuli, plus six other conditions. There was adaptation across different frequencies using 0- and 80-msec sawtooth items at 294 Hz, adaptation across different waveforms using 0 and 80 msec sine wave items at 440 Hz, and adaptation across both frequency and waveform using 0- and 80-msec sine wave items at 294 Hz. The same very diligent listeners served for all conditions but on eight separate days, one per condition. A preadaptation identification test was given before adaptation tests, and comparisons were always made between pre- and postadaptation identification functions for a particular day.

Results were very clear in support of auditory analyzers of pluck and bow. Adaptation within the 440-Hz sawtooth wave continuum was considerable. The normal boundary of about 40-msec rise time shifted to about 37-msec rise time in the pluck-adaptation condition, and to about 50-msec rise time in the bow-adaptation condition. Both shifts were highly significant and the difference in their size is common in adaptation findings. Such differences are often difficult to interpret, but here we think that they relate to inherent limitations in a continuum such as rise time. For example, a stimulus can be no more abrupt than 0-msec rise time, but can be infinitely less abrupt than 80 msec.

Significant postadaptation boundary shifts occurred for nearly all other conditions as well, but their magnitudes tended to be smaller. For example, pluck-adaptation shifts across one dimension (frequency or rise time) averaged less than 1 msec, and complementary bow-adaptation shifts averaged only 3 msec. Adaptation boundary shifts across both dimensions (frequency and rise time) were smaller still: only 2 msec for both pluck and bow. These results are regular enough to allow a simple interpretation: The more stimulus dimensions are held in common between adapting and test stimuli the larger is the effect. These results for pluck and bow items are very similar, then, to the array of results for speech items.

Let us suppose that the somewhat smaller adaptation effect of /de/ than /dae/ on a /bae/-/dae/ test continuum is due to the fact that fewer dimensions are shared between /de/ and the test items. These dimensions could be linguistic (the
vowels differ between adapting and test stimuli) or they could be auditory (the spectra differ as well). The fact that there is no adaptation effect of /aɛd/ on /baɛ/−/daɛ/ is difficult to account for in linguistic terms without relying on some allophonic or syllabic level of processing, but is relatively easy to account for in auditory terms because the stimuli are very different. Ignoring the common vowel nuclei, the comparable transitions in all formants go the wrong way, that is, in opposite directions. For /daɛ/ for the first formant ascends in frequency into the following vowel whereas for /aɛd/ it descends following it. The reverse is true for the second formants. Perhaps, then, all shifts due to adaptation are actually auditory in nature rather than linguistic: Perhaps they are due to adaptation of linguistically relevant features, not linguistic features.

How, then, does adaptation actually work? Perhaps it is not the feature analyzer itself which is fatigued. Instead, it may be the neural pathway leading to the analyzer which suffers fatigue. The more the adapting stimulus differs from the test-item array the fewer may be the number of intervening processing stages (from the registration of the acoustic signal to the binary feature analyzers) that are shared between adapting and test stimuli. Thus seen, fatigue during adaptation might be an inhibitory process that builds up throughout the many neurons and synapses of a particular pathway for a particular signal. If this notion is correct, then when the test array differs from the adapting stimulus each item in that array will travel a somewhat different neural path from that of the adapting stimulus. The sections of the path for the test stimuli which are not held in common with the adapting item will not be fatigued, while those portions held in common will be fatigued. Roughly speaking, if only half of the pathway neurons are held in common perhaps the adaptation effect might be only half as great. It should be obvious that the use of the term pathway here is at least partly metaphorical, but I do not mean it to be exclusively so (see Posner, in press).

If the pathway-fatigue account of selective adaptation is viable, the sensitivity functions shown in Fig. 9 and the decrease shown in one function in the bottom panel of that figure, may not mark the sensitivity of the actual analyzers themselves. Instead they may mark, for a given stimulus, the rate of firing of the sequence of particular neurons in the pathway which leads to the analyzer: Adaptation leads to pathway fatigue and less neural activity. The end result is exactly the same. Rather than sensitivity plotted on the ordinate of that function, one might substitute the term neural activity. Boundary shifts would occur in the same manner.

What about the notion of linguistic feature analyzers as opposed to linguistically relevant analyzers? The answer can come only after exploring further the notion of pathways. It seems to make no sense to think of a linguistic pathway; that is, a neural route which is traveled exclusively by speech stimuli. Such a pathway would necessarily have to have an early gating device which has already decided that a particular stimulus is speech. If the speech--nonspeech decision
were already made at this early level subsequent analyses would appear to be unnecessary: To determine whether or not an item is speech, the system would surely have analyzed the signal for speechlike features. Instead, then, the pathways leading to the binary analyzers are auditory; that is, general enough to handle both linguistic and nonlinguistic events. Since seemingly all the variance in boundary shifts across different types of adaptation situations might be accountable in terms of pathways, the existence of linguistic analyzers would be unimpeached. It is clear that there are auditory analyzers of similar nature, at least for pluck and for bow, but they may exist side by side, as it were, with linguistic analyzers. The only logical argument I can offer against this possibility is an appeal to parsimony: why have two kinds of binary feature analyzers, linguistic and nonlinguistic, when one set of nonlinguistic analyzers might do? It may be that the assignment of particular speech labels such as /b/ and /d/ occur subsequent to their categorization. Alas, I am getting too far away from the available data to be comfortable with such speculation.

As an internal summary then, postadaptation boundary shifts in identification functions of certain speech and musiclike stimuli seem to be explainable in terms of neural fatigue. Exactly which pathways fatigue remains an important question. I have suggested that the fatigue takes place in the neural pathways prior to arrival at the feature analyzers. Such an account would easily allow for differential magnitudes of boundary shifting according to the number of differences between adapting and test stimuli. The analyzers themselves lie considerably beyond the point at which the two ears converge, and they may not suffer from adaptation “fatigue.” They appear to be binary, at least with regard to any one stimulus, and they appear to function according to signal detection criteria and a simplified Pandemonium model. It seems clear that they can be either linguistic or nonlinguistic in nature. Most may be solely linguistic and a few others may be linguistically relevant. Remember that rise time cues not only the difference between pluck and bow, but also the difference between /fa/ and /tfa/. At present we have found only one set of analyzers which overlaps the domains of speech and music. It would be of considerable interest if others could be found which perform this apparent dual function. It would also be of interest to find possible binary music analyzers which are not relevant to speech.

6 SUMMARIZING REMARKS

What of the magical number two? Part of the answer is the same as for the magical number seven. George Miller noted seven wonders of the world, seven points on a psychological rating scale, seven seas, seven categories of absolute judgment, seven deadly sins, seven objects in the span of attention, seven days of the week, and seven digits in the span of immediate memory. He suspected that all these sevens were merely “pernicious, Pythagorean coincidence [p. 96].”
Such coincidence is found even more easily in twos: the two faces of Janus; the
two types of learning, operant and respondent; the two cosmic forces, Yin and
Yang; the two minds, the conscious and the unconscious; the two sexes, male
and female; the two searches through memory, self-terminating and exhaustive;
the two diurnal segments, day and night; the two locales for research, laboratory
and field; and many other twos, such as up and down, self and others, and
indeed /ba/ and /da/, and pluck and bow. Jakobson, Fant, and Halle suggested
that the dichotomous scale is the pivotal principle of linguistic structure; a quick
glance at twoness elsewhere suggests that it may be the pivotal principle
concerning the manner in which we parse the world (see Ogden, 1932).

Beyond any “perniciousness” of the Magical Number Two, humans appear to
be predisposed to perceive certain auditory events in a dichotomous manner. I
first discussed such discrete perception in terms of identification functions for
an acoustic continuum of speech sounds from /ba/ to /da/ generated by a
computer-driven speech synthesizer. This first expression of the magical number
two proves not to be crucial. Lines slanted at various angles yield equally quantal
identification functions. The crux of the Magical Number Two is revealed in the
discrimination functions, where for the stop consonants one can discriminate
only as well as one can identify, but for the slanted lines one can discriminate
almost infinitely better. A continuum of vowel sounds from /i/ to /I/ yields
intermediate results, and these results appear to be interpretable in terms of a
small perceptual deviation from the number two. The peculiar nonlinearity
found in the discrimination of stop consonant sounds is not unique to speech
items. In fact, evidence for the dichotomous perception of music-like sounds is
just as striking as that for speech.

Our perceptual predisposition toward the magical number two appears to stem
from biological endowment. Infants as young as one and two months parse
certain speech and music sounds in a manner functionally identical to that for
adults. Such results indicate that these categories and boundaries are natural
according to the most stringent criterion: indeed, they appear to be innate.

The neural mechanisms underlying our perception by twos can be thought of
as yoked pairs of feature analyzers lying well beyond the cochlea. Some may be
unique to speech analysis; others, like those for pluck and bow musical sounds,
may be used in both speech and nonspeech analysis. Continued presentation of a
particular stimulus prototype appears to fatigue selectively the feature-analysis
system, and temporary shifts in the locus of category boundaries are obtained.
These mechanisms appear to allow for high-speed speech perception, the rapid
categorization of a particular speech sound and the discarding of its nonproto-
typic vagaries, and also allow for great savings as the speech sounds are coded
into tight bundles of phonetic features suitable for memory and storage.

In music, however, the role of feature analyzers is less clear. Pluck and bow
categories obviously discriminate among modes of playing certain musical instru-
ments, and may relate to the reason these various stringed instruments were
invented, but beyond this minor role and in the absence of other binary musical features (which may yet be discovered), we simply do not know why they exist. It may be that they were auditory precursors to phonetic feature detectors, as if nature were experimenting with the feasibility of such devices. It may be that they are related to orienting and startle mechanisms: sounds with rapid onsets often forebode danger, whereas sounds with more gradual onsets are more likely to be associated with “safer” events. Beyond these speculations it is too early to say what their purpose may be. The present situation has caused some of our colleagues in speech perception to view the mere existence of pluck and bow categories as “pernicious, Pythagorean coincidence.” Indeed, they may be, but I prefer to think that they will eventually be tied to a theoretical framework relating the structures of speech and music.

Whorf (1956, p. 249) suggested that “speech is the best show man puts on. It is his own ‘act’ on the stage of evolution, in which he comes before the cosmic backdrop and really ‘does his stuff.’ But we suspect the watching Gods perceive that the order in which his amazing set of tricks builds up to a great climax has been stolen”—from bits and pieces of underlying mechanical and perceptual processes, like those used in music.

REFERENCES


Lane, H. *The hue and cry concerning categorical perception (and conversely)*. Paper presented at the 23rd symposium of the 28th International Congress of Psychology, Moscow, 1966.


Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, 63, 81–97.


