the challenge of speech perception. Here, the well-documented phenomena of experimental phonetics prove irreducible to the simple formulation used by the mechanism, which fails the task of consonant place categorization. In contrast to localization, which is sufficiently described as a mapping of phase differences to azimuth, the relation between second formant (F2) onset and F2 vowel as a correlate of phonetic place is admittedly more complex. The target article describes cases and countercases, and the eventual maps do not resemble an array of the place features of English, at least not according to standard linguistic description (labial, labiodental, linguodental, alveolar, postalveolar, palatal, velar) (Catford 1988). Particular values along this *n*-ary dimension are omitted (Fig. 16), and the detailed findings of the statistical analyses include erroneous assignment of consonants sharing a place feature (such as /s/ and /z/) to different loci. Rather than considering this to falsify the hypothesis that categorization relies on low variance linear mappings of acoustic to phonetic properties, the modelers adapted the model, placing a bat-based processor alongside a more heterogeneous set of feature analyzers. The properties of these additional feature analyzers were not chosen in reference to specific sensory or psychophysical evidence.

The insufficiency of the linear component of the model must be taken to disconfirm not only the perceptual account of phonetic categorization but the evolutionary one as well. If the articulatory repertoire had been shaped by a perceptual insensitivity to all but linear low-variance vocal sound production, should the acoustic variation of English consonants still be so recalcitrant? Does English preserve atavistic features that somehow failed to evolve an optimally linear form and variation? Implicitly, the last model (Fig. 17) concedes by virtue of its composition that speakers abrogate an orderly output constraint each time the categorization of a consonant requires an F3 or a burst analyzer, to say nothing of the other acoustic properties that evoke phonetic impressions despite their dissimilarity from the likely acoustic products of vocalization (Remez et al. 1994).

We have all been impressed by the informative power of frequency variation in F2 (Remez et al. 1997), and the present critique of the reality of the mechanism allegedly producing consonant place maps should not be taken to demote this acoustic attribute. The question of the acoustic-phonetic projection – does the F2 transition bear phonetic information? – is separate from the question provoked by the target article – does a human listener represent F2 frequency transitions of speech sounds the way Figure 16 does? The authors are judicious in noting the speculative nature of their proposal. However, to demonstrate that linear, low-variance phonetotopic maps accomplish the categorization of speech sounds requires a point of evidence that the target article did not deliver: such perceptual or physiological evidence would show that something similar to this neural map of F2 variation exists in the human auditory system and that its function is causally and necessarily involved in the perceptual registration of consonant place. For an alternative, evidence would identify an animal model of the phonology of English and would determine whether the topography of the response properties of auditory neurons conforms to a collection of iso-stop-place territories. Either of these points of evidence would convert an analogy to a proof that chiropterans, strigiforms, and hominids indeed exhibit this allegedly universal form of neural analyzer, and that the analyzer is equal to the task of analyzing consonants. Although evidence from the wet lab is convincing that such neural maps are employed in auditory localization and echolocation, the statistical evidence adduced about locus equations leaves a definite impression that the bat or owl listening to speech in the dark does not hear consonants the way a human listener does.

Patterns of evolution in human speech processing and animal communication

Michael J. Ryan, Nicole M. Kime, and Gil G. Rosenthal
Department of Zoology, University of Texas, Austin TX 78712.
mryan@mail.utexas.edu nmkime@mail.utexas.edu
fishman@mail.utexas.edu uts.cc.utexas.edu/~ryanlab/

Abstract: We consider Sussman et al.'s suggestion that auditory biases for processing low-noise relationships among pairs of acoustic variables is a preadaptation for human speech processing. Data from other animal communication systems, especially those involving sexual selection, also suggest that neural biases in the receiver system can generate strong selection on the form of communication signals.

This commentary provides a perspective from animal behavior that is probably unfamiliar to many linguists and neuroscientists. Specifically, we will address the proposed patterns of evolutionary events that result in human speech, patterns that have parallels to those proposed by some recent studies of animal communication.

One of the basic functions of many animal communication systems is to identify members of the same species for the purpose of mating. To do so, many species are characterized by signals that are species-specific, and perceptual systems whose response properties are biased toward these signals. Evolutionary biologists have been interested in how such congruence between signaler and receiver comes about in the new signaling systems that characterize new species (e.g., Doherty & Hoy 1985).

There are several possibilities for matching signals and receivers. A match could be achieved by single genes or tightly linked sets of genes that similarly influence both the signaler and the receiver. One example might be central pattern generators in crickets, in which a neural timing mechanism determines temporal parameters of both call production and recognition (cf. Doherty & Hoy 1985). Signals and receivers can also be brought into congruence when there is sufficient neural developmental plasticity to allow receiver response properties to be biased by experience with the signals, as with song learning in birds (Konishi 1994).

An alternative explanation for signal-receiver congruence is that one system constrains the form of the other. Recent studies of sexual selection suggest that receiver systems can have a strong influence on signal structure, in that males evolve signals that exploit previously unexpressed response biases in the females. For example, there is such a bias for extra syllables added to calls of some frogs and birds (cf. review of sensory exploitation in Ryan 1997).

Therefore, while tightly coincident patterns of coevolution might occur, they are certainly not the only mechanism by which signal-receiver congruence can evolve. The target article suggests that the evolution of human speech signals has been constrained by features of auditory processing:

> . . . linear relationships with low noise are quite general . . . and . . . auditory systems include mechanisms preadapted to process just such acoustic patterns, so that the human speech production system has been constrained to produce acoustic patterns that conform to this preadaptation (the orderly output constraint). (sect. 6)

Bats and barn owls decode spatial information with combination-sensitive neurons that respond to highly predictable (low-noise, linear) covariation of pairs of acoustic parameters; this association is a matter of acoustics and not biology (e.g., frequency and interaural time difference). Sussman et al. suggest that a similar relationship between the onset and offset frequency of second formant (F2) transitions in consonant-vowel sequences helps to resolve the noninvariance problem in human speech. They also suggest that the low noise in this system is not simply a by-product of acoustic constraints, as in sound localization, but of evolution. The acoustic parameters in speech have evolved this tight correlation because these are the kinds of cues that the mammalian (if not vertebrate, see Sussman et al., sect. 1.1) auditory system is biased

toward processing. Because results of vocal-tract area models also result in low-noise locus equations (Fig. 13 in Sussman et al.), we must ask if the human vocal tract has evolved to produce these low-noise relationships, or if this is a result of biophysical constraints on any sound-producing system.

One might expect at least some degree of correlations between onset and offset frequencies due to biomechanics. Whether a frequency sweep (Fig. 3 in Sussman et al.) is generated by changing the volume of resonating chambers as in humans, the tension of the medial tympaniform in birds, or the vocal cord tension in frogs, frequency onset and offset could be constrained if time durations (relative to the dynamics of the mechanism generating the sweep) were short. A correlation could also arise if the shape of the sweep, rather than its onset and offset, were a salient feature in processing. Data from other primates might be helpful in evaluating this claim, but a more global comparison might be rewarding as well. For example, the call of male túngara frogs is a frequency sweep with a statistically significant ($N = 300$, $F = 10.49, p = 0.001$) but high-noise relationship ($r^2 = 0.034$) between frequency onset and offset. Signals in nonhuman animals might not be identical to consonant-vowel transitions in humans, and thus by themselves cannot reject the coarticulatory resistance hypothesis. If, however, a variety of animals also tended to show such a high-noise relationship between frequency onset and offset, this would further suggest that the human speech production system is an adaptation for producing low-noise locus equations.

We end by suggesting a possible scenario for the origin of the "preadaptations" posited by Sussman et al.'s model. Many animals, not just bats and barn owls, need to localize sound in order to detect predators, find food, avoid competitors, or locate mates. Localizing a sound in space is another invariance problem. As we have seen, there are by necessity low-noise relationships of acoustic parameters that can be used in localization. It is possible that natural selection or an ancestral auditory system (i.e., ancestral at least to tetrapod vertebrates) to localize sounds in the environment resulted in the general use of combination-sensitive neurons, and perhaps auditory maps, to process these highly correlated pairs of acoustic variables such as frequency and interaural time of arrival differences. If so, such processing might be a general property of the vertebrate auditory system that was then co-opted for use in systems highly specialized for sound localization, for speech processing, and perhaps for other kinds of signal processing in other animal communication systems.

Acoustic correlates and perceptual cues in speech

James R. Sawusch

Department of Psychology and Center for Cognitive Science, State University of New York at Buffalo, Buffalo, NY 14260.
jsawusch@acsu.buffalo.edu
wings.buffalo.edu./soc-sci/psychology/labs/srlsawusch.htm

Abstract: Locus equations are supposed to capture a perceptual invariant of place of articulation in consonants. Synthetic speech data show that human classification deviates systematically from the predictions of locus equations. The few studies that have contrasted predictions from competing theories yield mixed results, indicating that no current theory adequately characterizes the perceptual mapping from sound to phonetic symbol.

When one listens to someone speak, one hears a string of words. However, this simplistic observation hides the considerable computation involved in the mapping of sounds to segments to words. The locus equations described by Sussman et al. are one attempt to specify part of this mapping from sound to segment. This commentary will focus on two aspects of locus equations. First, how general are these equations as a description of the acoustic correlates of place of articulation in consonants? Second, is the acoustic correlate described by the locus equations also the effective perceptual cue in the processing of speech by humans?

Some limits on locus equations as an acoustic correlate of perception. In studies with synthetic speech, the direction and extent of the second formant (F2) transition has been consistently shown to influence the perception of place of articulation in consonants. However, the labels used by adult listeners for synthetic speech syllables do not always coincide with the predictions of the locus equations. Sawusch (1986) described a relevant study using synthetic two-formant syllables. In a voiced stop-vowel series in which the second formant transition went from rising through steady-state to falling, listeners reported hearing /ba/, then /da/, and finally /ga/. In a second series, the voiced excitation of the formants was replaced by aspiration for the first 60 msec of each syllable. Listeners labeled the stimuli with a rising F2 transition as /pa/ and the rest of the stimuli in the series as /ka/. That is, syllables that had been labeled as /da/ with a voiced source were labeled as /ka/ with a voiceless source. Because all other synthesis parameters except for the voicing difference were the same, the F2 transitions for comparable stimuli in the two series were also the same. Thus, if the locus equations indicate that a stimulus in the voiced series was /d/, then the corresponding stimulus in the voiceless series should have been identified as /t/. However, for all of the voiced stimuli that listeners identified as /d/, their identification of the corresponding voiceless stimuli was as /k/ (a different place of articulation). Consequently, something other than the locus equation is governing perception of one or both sets of stimuli. These data indicate that the locus equation is not a true invariant. It may, however, be one of a set of acoustic correlates used by listeners (see Sussman et al., sect. 6.1).

Alternative perceptual cues. The second step in understanding the role of locus equations in speech is to elucidate their role in perception. The question here is not whether locus equations correlate with perception. Rather, it is whether the processing model described by Sussman et al. is an accurate characterization of the perceptual processing of consonant place of articulation information. Testing this model involves creating stimuli that contrast predictions of Sussman et al. with alternative computational descriptions of consonant place perception. Lahiri et al. (1984) proposed that stop consonant place is cued by the change in the tilt of the spectrum from stop release to the onset of voicing. Forrest et al. (1988) described the perception of consonant place in terms of the shape of the spectrum as captured by the mean and the first three moments about the mean of the spectrum. Each of these computational descriptions has been shown to correlate with listeners' perception of consonant place of articulation. That is, like the locus equations, these descriptions have been shown to capture an acoustic correlate of perception.

Richardson (1992) created sets of synthetic stop-vowel syllables. In one set, synthetic /b/, /d/, and /g/ were modified so that the formant transitions remained the same but the shape of the spectrum at stop release was altered. In another set, the shape of the spectrum at release was maintained, but the formant transitions (including F2) were changed. The results showed that both changes to the formant transitions and the shape of the spectrum altered perception. One interpretation of these data is that the formant transitions (including F2) and the shape of the spectrum at stop release are cues that are jointly sufficient, but individually unnecessary in perception. Alternatively, all of these descriptions of the stimulus are incorrect characterizations of perceptual processing and some alternative is needed. Results such as these indicate that the F2 transition and locus equations are not a perceptual invariant (but see Dorman & Loizou 1997 for additional data). They also raise the possibility that the model proposed by Sussman et al. is not an accurate characterization of the perceptual processing of consonant place information, even