Categorization and separability. In bats, owls, and humans, feature maps could be interpreted differently by the next system in the processing chain: Sussman et al. consider the representation of velocity in the bat (sect. 1.3.1 and Fig. 2A) and interaural time difference (ITD) in the owl (sect. 1.3.2 and Fig. 2B) as categorical, which could be misleading. Both are represented continuously, and so they are perceived. Clearly, one can look on continuity as the limit of categorization as the number of classes goes to infinity.

What seems to work with bats and owls (Fig. 2) does not work with human consonant identification (Fig. 16B), namely, uniquely associating a position in the space of input features (the decision space) with a definite class. In F2 onset-F2 vowel space, representatives of different consonants occupy overlapping regions. In both bats and owls, however, separability is provided by the physics of signal generation. Generally, with input features x1 and x2 the following type of equation holds:

$$\mathrm{x2} = k * x1, \text{ i.e., } \mathrm{CF2} = k * \mathrm{CF1},\ k = 2(a + \Delta \mathrm{v})/(a - \Delta \mathrm{v})$$

where CF1 and CF2 are the constant frequencies of the first and second formant of the pulse and its echo, respectively; Δv is the velocity of the target with respect to the bat; *a* is the sound speed in air.

$$\mathrm{F} = k * \mathrm{P}$$

and

$$k = 1/(2\mathrm{PI} * \mathrm{ITD}) * \mathrm{P},$$

where F is frequency and P is phase. Consonant locus equations, however, are of the form $2 = k * x1 + c, c \neq 0$, which, by itself, does not provide separability.

Linearity recognition, emergent properties, and higher-order feature detectors. The neural realizations of decision spaces are topologies of combination-sensitive neurons. The receptive field of each of these neurons covers a certain part of the input space; that is, there exist best values of the input features to which a neuron responds maximally. If neurons are arranged in such a way that neighboring neurons respond to similar points in input space, a pair of input features is identified by the position of the most active neuron in the map. The question then arises whether, in separable decision spaces, mechanisms will be necessary to project this position information to neurons further up in the hierarchy that can detect higher order features, or emergent properties, such as slopes (k) and y-intercepts (c) of the regression lines. Neurons in the separable afferent map could be connected directly (mapped) to neurons in an efferent map continuously coding the appropriate behavior in response to the input situation; for example, in bats, to speed up, or slow down, or change the frequency of the emitted sonar in order to catch the prey.

Human phoneme categorization based solely on F2 onset and F2 vowel, however, does require such higher-order feature detectors. Sussman et al.'s results (sect. 3.2.3) might indicate that in *k, c* space, one can discriminate between most consonants from different manner classes, at least between the voiced stop consonants /*b*/, /*d*/, and /*g*/ (Fig. 6). But how could this decision space be realized neurally; that is, how could linearity be recognized? In order to derive *k* and *c*, at least two different F2 onset-F2 vowel pairs representing the same consonant would be needed. These are not available at a single instant in time, and there are no temporal correlations between consonant-vowel articulations of the same consonant that could be exploited.

If these higher-order features cannot be determined, consonants can only be identified by introducing one or more additional features, as Sussman et al. suggest in their Figure 17. Adding a third dimension in the decision space by an appropriately chosen feature or combination of features, consonants could be separated by a plane. The choice of F3 and burst descriptors as possible candidates is in agreement with suggestions from other authors. We suppose that voice onset time as an evolutionarily old percept could be an additional cue (Ehret 1992).

So what is linearity good for? The input to any auditory system is a time course of a physical entity. There are always multiple ways of defining features that describe the same relevant correlations in the input signal. Linearity, however, could simplify the form of the decision boundary; that is, make it easier to implement by whatever neural mechanisms are used.

Self-organizing maps and mappable inputs. The question of whether there are computational reasons for the existence of strongly correlated components in speech signals (sect. 7) seems to confuse cause with effect. The right question was asked in section 4: Why has the human articulatory system developed to fulfill the orderly output constraint?

If mapping is defined as a function *f:* R^m to R^n, which uniquely assignes to each input vector *x* ⟨ELEMENT⟩ R^m a vector *u* ⟨ELEMENT⟩ R^n, then, combinations of arbitrary variables or features are always mappable. Another question is how useful this mapping actually is. In self-organizing maps, the components of *x* are the features extracted from the sound signal, and *u* describes the position of the neuron that is excited maximally in response to *x*. For further processing, whether there exists a mapping from a neuron's position to the category it should be assigned to is important. Here, again, we have the separability problem. The mappings in Sussman et al.'s Figures 18A–C are of the type R^2 to R^2. Because they do not involve a dimension reduction, topology can be perfectly preserved, and the receptive fields of the neurons mirror the distribution of the input vectors *x*; that is, Figure 18A resembles the situation in Figure 16B. Is such a mapping useful at all?

A phonological perspective on locus equations

William J. Idsardi
Department of Linguistics, University of Delaware, Newark, DE 19716-2551.
idsardi@udel.edu www.ling.udel.edu/idsardi/

Abstract: Locus equations fail to provide adequate abstraction to capture the English phoneme /g/. They also cannot characterize final consonants or their relation to pre-vocalic consonants. However, locus equations are approximately abstract enough to define the upper limit on phonological distinctions for place of articulation. Hence, locus equations seem to mediate phonetic and phonological perceptual abilities.

To listen to speech is to be fooled much of the time. Physically different sounds are heard as the same sound, and physically identical sounds are heard as different sounds. This description is reminiscent of that of visual illusions. What is different in human language is that the grouping of speech sounds (indicated with [ ]) into mental equivalence classes (*phonemes,* indicated with / /) is different in different languages, and children must learn the phonemes used in their particular language. This problem is simplified somewhat by the fact that phonemes are not the basic units of speech sounds. Speech sounds are made up of phonological *features,* much as chemical compounds are composed of chemical elements; see Halle (1991). Sussman et al. suggest that locus equations can explain human speech sound categorization in a neurobiologically plausible way. This is a laudable goal, and locus equations do better than previous measures. But do locus equations adequately characterize the mental equivalence classes (the phonemes)? That is, do the phonemes of a language emerge out of the locus equations derived from pronunciation?

Whole phonemes certainly do not emerge out of locus equations. The data regarding different manner classes (sect. 3.2.3) show that locus equations provide cues not to phonemes, but to one of their featural components: the place of articulation. That is, locus equations provide cues to the *major articulator* of the sound, in Halle's (1991) terms. This interpretation explains the results of Sussman et al. (1993), who found no significant difference in locus equations for Arabic [d] and [dʕ] or for Urdu [d] and [ɖ]. All these sounds share the same major articulator: the front portion of the tongue; they differ in their secondary articulations. Hence, locus equations do group together sounds that share this major articulator.

Let us now consider English. English has a phoneme /g/, which has several different pronunciations, depending on the neighboring sounds. Look into a mirror and say the words *goose* and *geese.* You will notice that the lips are rounded in *goose* even as you prepare to speak, but not in *geese.* This is a coarticulation effect, whereby the /g/ takes on some characteristics of the following vowel, in this case lip-rounding. It is not as easy to observe, but the position of the body of the tongue is also different in the production of /g/ in these two words, again anticipating aspects of the following vowel. In *geese* the tongue body is more toward the front of the mouth, in contact with the hard palate, [g^{j}] (palatal-g), whereas in *goose* the tongue is in contact with the velum, [gɣ] (velar-g). However, what every speaker of English knows is that none of this matters. The words *goose* and *geese* begin with "the same sound," /g/. Sussman et al.'s Figure 4 (sect. 3) shows that /g/ does not emerge out of the locus equations. The best fit is with two equations, separating /g/ into two categories – palatal-g and velar-g. There is no question that these categories exist in *pronunciation.* Indeed, as Sussman et al. indicate "phoneticians have long described two *allophonic* variants of /g/ . . ." (sect. 3, para. 3; emphasis added). However, splitting /g/ into two categories contradicts what every speaker knows about the memorized form of these words: *goose* and *geese* both start with the same sound (this is the meaning of the term *allophonic*). Thus, in the case of English /g/, locus equations still hug the physical ground too closely. Locus equations do not provide sufficient abstraction to capture the phonological invariant of English /g/ – its major articulator, the body of the tongue. However, there are languages (e.g., Russian) that do distinguish between palatal-g and velar-g; we will return to this point, below.

Another problem faced by locus equations is that English words can end in various consonants and still remain distinct in speech. For example, *bib, bid,* and *big* are all different English words, but in isolation there is no vowel following the final consonant, and by definition there is no locus equation for the final consonants. Therefore locus equations can neither characterize final consonants nor provide the basis for their categorization. Moreover, every speaker knows that the /g/ at the end of *big* is "the same sound" as that in the middle of *biggest.* A locus equation is available for *biggest,* but locus equations cannot be the source of the perceptual equivalence of the /g/ in *big* and *biggest.*

Sussman et al. also claim that the slope of a locus equation measures the degree of coarticulation, in the range [0, 1] (sect. 3.1, para. 2). However, five speakers in Sussman et al. (1991, p. 1317, Table II) have slopes greater than 1. How are we to interpret such hypercoarticulation values?

So what do locus equations accomplish? Phonemes do not emerge directly from them. Even the place of the major articulator does not adequately emerge, as English /g/ shows. But locus equations seem to provide about the right abstraction for the set of *potential* phonological differences of the major articulator in consonant-vowel contexts. By this I mean that locus equations provide just enough detail to categorize as different two sounds that could be classified as having different major articulators in some human language. If this is correct, then locus equations would define the upper limit on phonemic place categorization and thus mediate phonetic and phonological perceptual abilities. This would be a significant achievement even though it would not explain language-specific phonemic perception, or how children tune their perceptual abilities to their language.

Are locus equations sufficient or necessary for obstruent perception?

Allard Jongman

Department of Modern Languages, Cornell Phonetics Laboratory, Cornell University, Ithaca, NY 14850. **aj12@cornell.edu**
www.phonetics.cornell.edu/allard/aj.html

Abstract: Two issues are addressed in this commentary: the universality and the "psychological reality" of locus equations as cues to place of articulation. Preliminary data collected in our laboratory suggest that locus equations do not reliably distinguish place of articulation for fricatives. Additionally, perception studies show that listeners can identify place of articulation based on much less temporal information than that required for deriving locus equations.

Sussman et al. make a compelling case for locus equations as derived invariant cues to place of articulation in stop consonants. The reported high correlation and linearity between the second formant (F2) at vowel onset and at vowel midpoint for consonant-vowel (CV) syllables constitutes a very significant finding, given the long and largely unsuccessful quest for invariance in this domain.

I am currently exploring the role of locus equations as invariant cues to place of articulation in fricatives. English fricatives are produced at four distinct places of articulation: labiodental /f,v/, dental /ɵ,ð/, alveolar /s,z/, and palato-alveolar /ʃ,ʒ/. Acoustically, it is notoriously difficult to distinguish labiodental /f,v/ from dental /ɵ,ð/. Perception experiments (Harris 1958; but see Jongman 1989) have suggested that cues to this distinction may reside in the transition between fricative noise and the following vowel. The fact that locus equations explicitly encode this transition information may therefore make them appropriate candidates for distinguishing fricatives.

Data have been collected from 20 speakers (10 females, 10 males), each of whom produced three repetitions of each fricative followed by six different vowels (/i, e, æ, ɑ, o, u/). This is, to my knowledge, the largest database of fricatives for which locus equations have been derived (for a preliminary report of a subset of the data, see Jongman & Sereno 1995). Mean slope and intercept values for each place of articulation across all speakers are shown in Table 1.

Separate analyses of variance on the slope and intercept values revealed main effects for both slope ([F (3, 76) = 32.25, $p < 0.0001$]) and intercept ([F (3, 76) = 40.27, $p < 0.0001$]). Post-hoc tests showed that only the slope value of labiodental /f,v/ was significantly different from that of the other three places of articulation. In addition, y-intercept values were distinct for labiodental /f,v/ and for palato-alveolar /ʃ,ʒ/, but did not distinguish among dentals and alveolars. These preliminary data suggest that neither slope nor y-intercept serve to distinguish place of articulation in fricatives. Although discriminant analyses have yet to be conducted, the fricative data appear to be less clear-cut than stop data.

Instead of reliance on a single cue for distinction of fricatives at four different places of articulation, a simple binary model in which different cues are considered in parallel may be more

Table 1 (Jongman). *Mean slope and intercept values for each fricative place of articulation across 20 speakers and 6 vowel contexts.*

	Labiodental	Dental	Alveolar	Palato-alveolar
Slope	0.768	0.530	0.517	0.505
y-intercept (Hz)	356	879	914	1065