BEHAVIORAL AND BRAIN SCIENCES (1998) **21,** 241–299
*Printed in the United States of America*

# Linear correlates in the speech signal: The orderly output constraint

**Harvey M. Sussman**
*Department of Linguistics and Communication Sciences and Disorders, University of Texas at Austn, Austin, TX 78712*
**Electronic mail:** *sussman@mail.utexas.edu*

**David Fruchter**
*Department of Linguistics, University of Texas at Austin, Austin, TX 78712*
**Electronic mail:** *fruchter@mail.utexas.edu*

**Jon Hilbert**
*Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712*

**Joseph Sirosh**
*HNC Software, Inc., San Diego, CA 92121*
**Electronic mail:** *sirosh @hnc.com*

**Abstract:** Neuroethological investigations of mammalian and avian auditory systems have documented species-specific specializations for processing complex acoustic signals that could, if viewed in abstract terms, have an intriguing and striking relevance for human speech sound categorization and representation. Each species forms biologically relevant categories based on combinatorial analysis of information-bearing parameters within the complex input signal. This target article uses known neural models from the mustached bat and barn owl to develop, by analogy, a conceptualization of human processing of consonant plus vowel sequences that offers a partial solution to the noninvariance dilemma – the nontransparent relationship between the acoustic waveform and the phonetic segment. Critical input sound parameters used to establish species-specific categories in the mustached bat and barn owl exhibit high correlation and linearity due to physical laws. A cue long known to be relevant to the perception of stop place of articulation is the second formant (F2) transition. This article describes an empirical phenomenon – the locus equations – that describes the relationship between the F2 of a vowel and the F2 measured at the onset of a consonant-vowel (CV) transition. These variables, F2 onset and F2 vowel within a given place category, are consistently and robustly linearly correlated across diverse speakers and languages, and even under perturbation conditions as imposed by bite blocks. A functional role for this category-level extreme correlation and linearity (the "orderly output constraint") is hypothesized based on the notion of an evolutionarily conserved auditory-processing strategy. High correlation and linearity between critical parameters in the speech signal that help to cue place of articulation categories might have evolved to satisfy a preadaptation by mammalian auditory systems for representing tightly correlated, linearly related components of acoustic signals.

**Keywords:** acoustic; linearity; locus equations; neuroethology; noninvariance; perception; phoneme; place of articulation; sound categories; speech signal

> Scientists do tolerate uncertainty and frustration, because they must. The one thing that they do not and must not tolerate is disorder.
>
> George Gaylord Simpson (1961, p. 5)

The goal of this target article is to provide a conceptualization of human speech sound categorization and representation in the brain that is neurobiologically viable and consistent with basic auditory-processing algorithms known from both avian and mammalian auditory nervous systems. Speech sounds that form contrastive categories in the phonological systems of languages are similar, in principle, to biologically important sounds in other species. The information conveyed by complex acoustic signals can be utilized across species in a wide variety of auditory-based behaviors such as acoustic communication, sound localization, or echolocation via biosonar. Neuroethological investigations of the mustached bat and the barn owl have uncovered species-specific auditory specializations for the processing of complex acoustic signals that, if viewed in sufficiently abstract terms, could have an intriguing and striking relevance for human speech sound categorization and perception.



harvey sussman is the holder of the R. P. Doherty Jr. Centennial Professorship in Communication Sciences and Disorders, as well as a professor in the Department of Linguistics at the University of Texas, Austin. He is the author of over 75 scientific articles in the areas of speech production and perception and cognitive neuroscience.

© 1998 Cambridge University Press 0140-525X/98 $12.50



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

The rationale for using a known neural model from neuroethology as a theoretical springboard to establish by analogy a speculative model for human auditory processing is straightforward because, first, the human brain is a product of evolution with its design and architecture generally conserved; second, overwhelming similarities exist in the structure and function of neural substrates across species possessing common stimulus-processing requirements; and, third, evolution tends to produce similar solutions to similar problems. As neuroethological research advances, it appears to be more and more obvious that each species, under selective pressures, solves its own idiosyncratic problem of "constructing and uniquely organizing combinatorial properties of acoustic attributes that are of clear importance for that animal's perception of its external world" (Pollak et al. 1995, p. 494). This basic combinatorial principle, common across species as a strategy for processing information-bearing acoustic attributes of complex input signals, should also be relevant for human speech processing, at least in the more peripheral stages. This is not to say that auditory substrates or functional properties across species are identical but, rather, that they are likely to utilize many of the same neural processing mechanisms and strategies. As Churchland and Sejnowski (1989, p. 42) stated: "Whatever the basic principles of language representation, they are not likely to be utterly unrelated to the way or ways that the nervous system generates visual representations or auditory representations, or represents spatial maps and motor planning." We will argue, and provide data to support it, that basic operational principles underlying phoneme encoding and category formation in human speech evolved from neural features that first appeared long before early *Homo sapiens* discussed the events of the day around the campfire.

We will first review neuroethological data that reveal three important generalities of auditory processing and representation: (1) a basic processing unit beyond isofrequency coding is the combination-sensitive neuron; (2) combinatorial processing of two acoustic parameters yields a third, higher-order, emergent property of biological significance to the organism; and (3) critical features of the input signal to combination-sensitive neurons are inherently linearly related as a result of basic physical laws. Following this discussion we will describe a specific speech/language phenomenon – the locus equation phenomenon – that presents a simple, robust, and empirically well-supported law governing the form of an acoustic attribute of consonants in various vowel contexts. Similar to the neuroethology examples, locus equation data also take the form of linear relationships with little noise (i.e., the acoustic data are very well-fit by a line). We will consider some alternative explanations for this high correlation and linear relationship between key signal components of the consonant-vowel (CV) unit, particularly arguing the idea that it could be a coevolutionary adaptation of the human speech production system to an evolutionarily conserved auditory processing strategy. This idea will be formulated as the "orderly output constraint" (OOC). According to the OOC, high correlation and linear relationship between critical acoustic elements of a complex signal enhance the processing and eventual representation of those inputs by categorical-feature-extracting two-dimensional (2D) arrays of combination-sensitive auditory neurons. The type of hypothesis that an acoustic pattern from speech data has been optimized via natural selection for a speech-encoding function is difficult to support, and we will not be able to do so in this article. Our purpose is simply to motivate the proposed constraint by marshaling available but, of necessity, indirect evidence from diverse domains of neurophysiological, behavioral, and computational research.

## 1. A neuroethological perspective on the generality of highly correlated and linearly related information-bearing parameters in acoustic signals

The leap from a "lower" mammalian neural system performing echolocation to a human neural system performing, for example, stop consonant place of articulation perception (namely, was it a "ba," "da," or "ga"?) might be thought extreme, but in principle it is not. In auditory areas of the thalamus (e.g., medial and dorsal divisions of the medial geniculate) immunocytochemical differences within certain cell groups are found across mammalian species (Pollak et al. 1995). These differing patterns of neurochemical adaptations have been interpreted as underlying "pivotal evolutionary features subserving some important facet of species-specific signal processing" (Pollak et al. 1995, p. 483). Each species adapts to its own auditory needs, but a fundamental continuity and functional similarity exists across mammalian species. One common theme is that combinatorial response properties of higher-order auditory neurons encode key physical aspects of complex signals underlying a biologically important auditory behavior. Forebrain structures, driven by selective and ecological pressures and characterized by evolutionary plasticity, contain combinatorial neurons possessing neural processing specializations precisely matched to the on-line signals that shape them (Pollak et al. 1995). Peripheral neural processing of human speech may be no different from what has been repeatedly documented in neuroethological studies of species-typical vocalizations.

### *1.1. Combination-sensitive neurons*

The neural unit that serves as the ubiquitous higher-order auditory processor appears to be the combination-sensitive neuron. Combination-sensitive neurons are specifically "tuned to coincidence (synchronization) of impulses from different neurons in the time, frequency and/or amplitude domains" (Suga 1994, p. 143). Combination-sensitive neurons compare ascending information derived from two or more spectral components of the signal.[1] In the mustached bat – the species that has received the most scrutiny – combination-sensitive neurons were initially thought to be created in the medial geniculate of the thalamus by converging tonotopically varied inputs from the inferior colliculus. Mittman and Wenstrup (1995) have recently shown that combination-sensitive neurons are already operative in a midbrain processing area – the central nucleus of the inferior colliculus.

A variety of combination-sensitive neurons have been documented in the mustached bat. Many respond to similar components of the biosonar pulse and its echo. The pulse and returning echo consist of four harmonics (30 kHz to 120 kHz) with each harmonic having a constant frequency (CF) and frequency modulated (FM) component. The



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

echo is time-delayed and Doppler-shifted in frequency from the pulse. CF/CF neurons encode target velocity by sensing Doppler shifts between various CF pairings of harmonic components of the emitted pulse and returning echo, and delay-tuned FM-FM neurons encode target range via echo delays relative to the pulse for FM components of the pulse/echo signal (Olsen & Suga l991a; 1991b). A recently discovered type of combination-sensitive neuron in the auditory cortex of the mustached bat processes signals that are particularly close to the acoustic structure of human speech in that the input components are a dynamic transition followed by a "steady state" (FM plus CF components), the same acoustic pattern produced by humans articulating a consonant and a vowel. These cortical neurons showed maximal facilitative discharges to the FM1 component of the biosonar pulse (≈30 kHz) and the CF2 component (≈60 kHz) of the returning echo (Fitzpatrick et al. 1993). The existence of such delay-tuned combination-sensitive neurons in the mustached bat, sensitive to FM and to CF components, suggests that similar types of auditory neurons could easily have evolved in human auditory substrates to encode the FM and CF components of consonant-vowel utterances.

Combination-sensitive neurons have been documented across a wide range of vertebrates in frogs (Fuzessery & Feng 1983; Mudry et al. 1977), in birds (Margoliash 1983; Margoliash & Fortune 1992; Takahashi & Konishi 1986), and in mammals (mustached bat, Suga et al. 1978; Suga et al. 1983; brown bat, Neuweiler 1983; 1984; mouse, Hoffstetter & Ehret 1992; cat, Sutter & Schreiner 1991; monkey, Olsen 1994; Olsen & Rauschecker 1992). Combination-sensitive neurons in the white-crowned sparrow are specialized for whistle-whistle, whistle-buzz, and buzz-trill combinations (Margoliash 1983). Sutter and Schreiner (1991), investigating response properties of cells in the dorsal region of the cat primary auditory cortex, found certain cell populations that were tuned to two or in some cases three frequencies and noted numerous similarities between these cortical fields in the cat and the CF/CF cortical areas in the mustached bat. In the primate nervous system of the squirrel monkey, Olsen (1994) reported combination-sensitive neurons encoding temporal delays between signal components that served to functionally categorize species-typical calls. In addition, several varieties of combination-sensitive neurons were found in the dorsal division of the medical geniculate body of the squirrel monkey. Among the varied calls of the squirrel monkey are acoustically simple sounds known as "peep," "yap," and "cackle," and a complex call known as a "chuck." The chuck consists of a tightly ordered sequence resembling an initial peep followed by a yap and ending with a cackle. In addition to finding peep-, yap-, and cackle-selective neurons, Olson found a combination-sensitive neuron that showed no response to a simple call (peep, yap, or cackle), but instead showed a maximum facilitative response to the complex chuck call. Eliminating any simple call from the chuck elicited a significant decrease in the neuron's response, and reversing the natural ordering of the three simple calls eliminated the neuron's repsonse altogether. Such multicomponent selectivity of an auditory neuron has striking relevance for human speech that is often characterized by multiple acoustic cues contributing to the identification of contrasting consonant plus vowel sounds such as "ba" versus "da" versus "ga."

### 1.2. Multifunctional processing across auditory behaviors

Kanwal et al. (1994) have described the rich variety of communication sounds ("calls") emitted by mustached bats, including at least 33 different types of sounds ("syllables") that possess both combinatorial properties and an extensive range of variation. These "social" calls also contain constant frequency patterns, frequency modulated patterns, and noise bursts. There is a fundamental frequency with concomitant harmonic structure and resonances shaped by a supralaryngeal filter. Of most importance, however, to our claim that the neural processing of human speech is analogous to auditory processing documented in other species is the recent finding that combination-sensitive neurons engage in multifunctional processing. Ohlemiller et al. (1994) have shown that combination-sensitive neurons in the auditory cortex of the mustached bat that had previously been regarded as exclusively performing echolocation processing actually had a dual function in that the same neuron also performed combinatorial analysis of acoustic parameters making up communication calls. A change in the context of processing from echolocation to communication calls was accompanied by a switch in the preferred temporal delay separating the two input elements that the cell best responded to, which is 2.4 msec for echolocation and 17 msec for analysis of calls. Combination-sensitive processing of species-specific calls by these bats is relatively similar to what would be required in human speech perception because it is performed on two elements from within the same input signal and not on two separable elements (pulse/echo) of an acoustic signal used for navigation and location of prey.

Regardless of the specific auditory behavior, it is readily apparent that combination-sensitive neurons perform the essential processing of stimulus components that possess category-specific attributes. Because human speech contains similar acoustic features to those found in the call repertoire of the mustached bat, there is no reason to suspect novel processing strategies or neuron types to have arisen for basic auditory encoding of the acoustic cues signaling feature contrasts in human speech. Categorical sorting of consonant-vowel syllables based on combinatorial analyses of certain features of the F2 transition, long known as an important cue for stop place of articulation (e.g., "ba"-"da"-"ga") perception (Liberman et al. 1954), seems a distinct and reasonable possibility.

### 1.3. Emergent properties, "category" formation, and linearly related inputs

A basic encoding problem for any perceptual system is to establish species-relevant categories[2] based on "information-bearing elements" (Suga et al. l983, p. 1574). Mustached bats form representations for target velocities, target distances, target sizes, and so on, and barn owls form representations for interaural time and intensity differences that signal azimuth and elevation coordinates for target location in space. In both avian and mammalian auditory processing centers, specific attributes and selected ranges of stimulus variation within a complex input signal are represented or mapped, using 2D arrays. A scheme common to both systems is displayed in Figure 1. A map of two independent stimulus attributes, *x* and *y*, is laid out systematically – not necessarily with linear scales as more important parts of the



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

range of stimulus variation are often overrepresented, for example, the second harmonic of the bat's biosonar signal (Fitzpatrick et al. 1993; Suga & Jen 1976). Documentation of response characteristics of combination-sensitive neurons (Fitzpatrick et al. l993; Olsen & Suga 1991a; 1991b; Suga et al. 1983) has shown that processing of *xi* and *yi* combinations typically yields a derived, emergent property, *z*. *Z* is a "category," an equivalence class of all the ordered pairs, $\langle x_i, y_i \rangle$ , belonging to a function relating *x* and *y*.

One feature common to the neuroethology examples is that the sensory input functions represented are quite linear. This is not always obvious, especially since cortical projections are somewhat distorted versions of Cartesian space. To bring out the linear relationships between input variables mapped by the mustached bat and the barn owl, data on the response characteristics of individual combination-sensitive neurons from these animals were plotted in Cartesian space.

**1.3.1. Isovelocity maps in the mustached bat.** Figure 2A is adapted from Suga et al. (1983), Figures 12 and 13. Their Figure 12 plots pairs of best facilitative frequencies for combination-sensitive neurons in CF1/CF2 and in CF1/CF3 specialization areas of the bat auditory cortex. The interpreted isovelocity functions corresponding to these pairs of pulse and Doppler-shifted harmonic echo frequencies are plotted in Suga et al., Figure 13. To produce our figure, we chose four isovelocity functions that were represented by the highest number of neurons: 0.7 m/sec, 2.8 m/sec, 4.8 m/sec, and 6.7 m/sec. Several harmonic pairs (CF1 of the pulse with either CF2 or CF3 of the echo) that were unambiguously matched to one of these four isovelocity values were then selected. CF1 values were plotted along the abscissa and CF2/2 and CF3/3[3] values along the ordinate. It can be seen that the derived velocity maps are laid out systematically within CF/CF neural space. The isovelocity contours projected onto Cartesian space are of course linear by virtue of the Doppler effect equation. Echolocation processing utilizes pairs of signal components that are linearly related.

**1.3.2. ITD maps in the barn owl.** Figure 2B illustrates *x-y* coordinates for the physical input to the barn owl neural array encoding interaural phase differences as a function of frequency. The phase-locked responses of tonotopically organized delay-line neurons in the nucleus laminaris (Sullivan & Konishi 1986) and their ascending projections to central nucleus neurons of the inferior colliculus are processed by elegant 2D matrices systematically representing frequency/phase relationships within the complex input signal (Wagner et al. 1987). Similarly to the bat, combination-sensitive neurons have been documented in the barn owl (Konishi et al. 1988). To derive Figure 2B, data values were taken directly from Wagner et al. (1987), Figure 13. Their schematic matrix illustrates how a derived variable, interaural time difference (ITD), emerges from variable frequency/phase relationships. Along vertically organized columns, different phase relationships spanning isofrequency laminae invariantly code a given ITD value. As can be seen in Figure 2B, the relationship of phase is plotted as percent of a cycle, and frequency is linear. Each line is a set of coordinates representing information of special behavioral significance to the barn owl – a microsecond time differential that is translated into a spatial

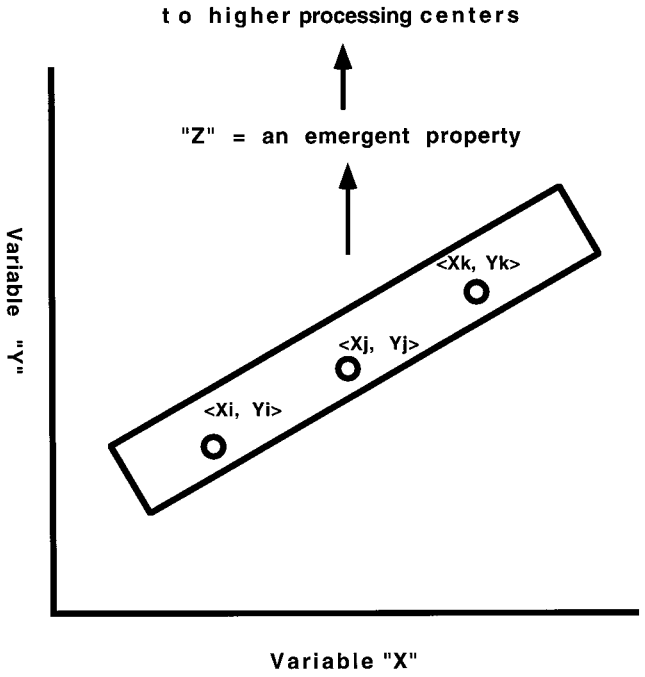



Figure 1. A schematic map of two independent stimulus attributes, *x* and *y*, systematically laid out such that combinations of ordered pairs ($\langle x_i, y_i \rangle$, $\langle x_j, y_j \rangle$; . . . ) yield a derived and emergent property, *z*, that represents a category or equivalence class. In neuroethological models from the mustached bat and barn owl this combinatorial acoustic/neural space has linearly arranged data coordinates that reflect emergent and species-specific biologically relevant categories.

coordinate in the azimuthal plane. These ITD columns have ascending projections to space-specific neurons in the external nucleus of the inferior colliculus that invariantly signal target azimuth.

Auditory maps in the mustached bat and the barn owl represent the best-known examples of how auditory substrates organize, represent, and signal information. In both cases, there are 2D maps of bivariate acoustic space in which there are linear functions that represent categories (or equivalence classes). In the bat, these linear functions are "isovelocity" contours. In the barn owl, they are "iso-ITD" functions. The organizational principles underlying the auditory encoding systems of the mustached bat and the barn owl can offer valuable clues for models of human speech perception. The following quote from Suga expresses well the rationale for using such models:

> The auditory system of humans shares "basically" the same anatomical structure with animals. Therefore, I believe, animals and humans share "basic" neural mechanisms for hearing. However, the mustached bat has developed certain specialized mechanisms for biosonar from the shared mechanisms. Humans have also developed specialized mechanisms for speech from the shared mechanisms. So there must be a difference between them. In bats, frogs, song birds, and, recently, monkeys, it has been found that the basic structure of species-specific complex sounds is processed by combination-sensitive neurons. I think the human auditory system has many combination-sensitive neurons to preprocess the basic structure of speech sounds, and has specialized mechanisms built upon that for speech processing. (Suga, personal communication.)



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

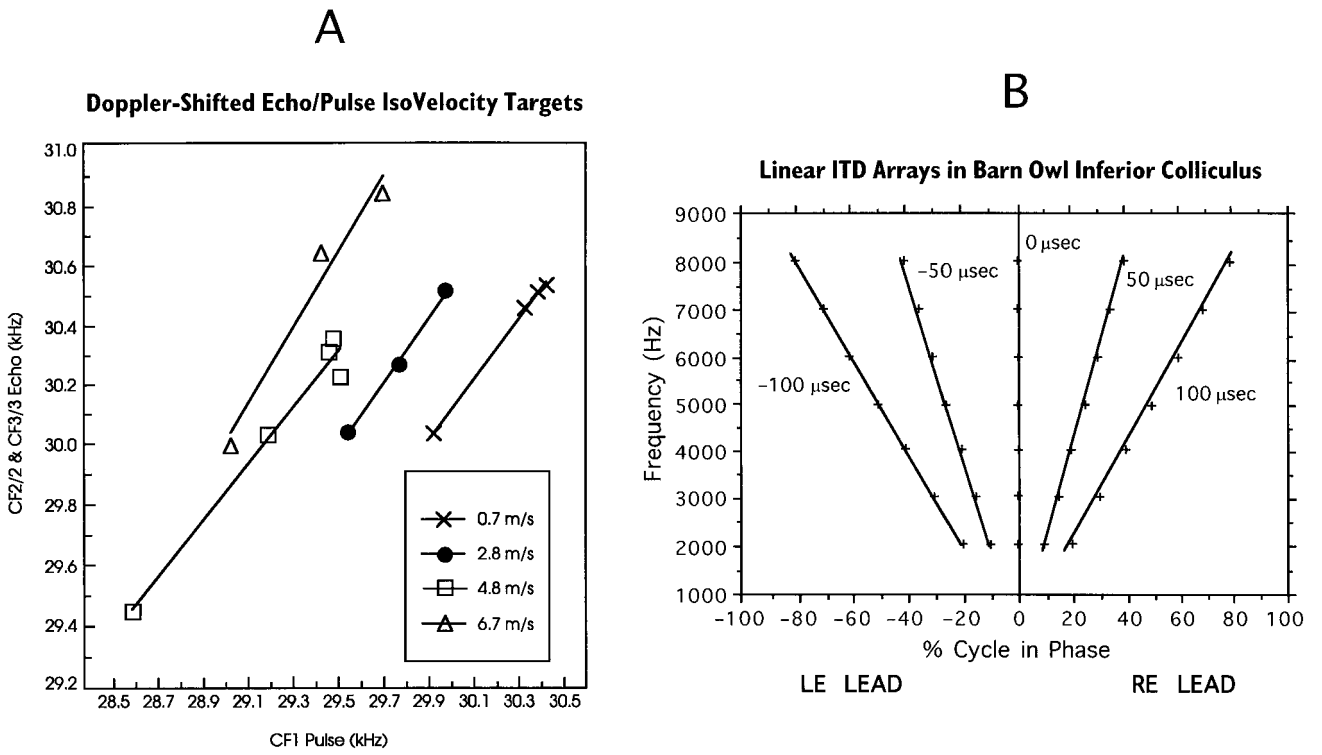



Figure 2. **A.** Examples of four linear isovelocity functions from the mustached bat obtained by plotting CF1 (kHz) of the pulse along the abscissa in relation to Doppler-shifted second (CF2) and third (CF3) harmonics of the echo plotted along the ordinate. (Data from Suga et al. 1983.) **B.** Examples of five linear iso-ITD categories from the barn owl obtained by plotting phase differences (as percent of cycle) along the abscissa in relation to frequency plotted along the ordinate. (Data from Wagner et al. 1987.)

We feel that auditory maps, as found in the mustached bat and in the barn owl, may be reasonable models (at least in a homoplasic sense and at an appropriate level of abstraction) for similar computational problems in the human auditory system (Sussman 1986; 1988; 1989). Stop consonant place perception across vowel contexts, if it involves an auditory map similar to locus equation plots (to be introduced in sect. 3), could utilize a processing strategy abstractly isomorphic to that of the mustached bat and the barn owl. Again, there would be a 2D map of a bivariate acoustic space in which linear functions represent categories. In analogy to the isovelocity contours of the mustached bat and iso-ITD functions of the barn owl, the lines of locus equation plots can be conceptualized as "iso-stop-place" functions.

## 2. The noninvariance problem in speech perception

The physical speech waveform encoding language has long resisted attempts to uncover laws relating the acoustic and symbolic levels of language structure. The speech signal tends to be extremely variable, as examples of the same phoneme (a contrastive speech sound) are often physically different in each context. For example, the initial *b* in "beat," "bit," "bait," "bet," "bat," "bought," "boat," "boot," "but" is categorized by listeners as the phoneme *b*, even though every instance of b is physically different. A seeming lack of order at the acoustic level within certain phonemic categories is one of the fundamental problems of speech perception and has greatly limited progress in machine recognition of speech.

The lack of a straightforward map between the physical signal and a unit of the message (in this case between the acoustic waveform and the phoneme) is known as the noninvariance problem. This issue has dominated theoretical debate in speech research for the last 50 years (e.g., Liberman & Mattingly 1985; Perkell & Klatt 1986). A particular paradigmatic exercise, namely, defining the nature of acoustic cues for stop consonant place of articulation (/b,d,g,p,t,k/) across vowel contexts, has been traditionally emphasized as a challenging test for those who would maintain that there is some level of signal-based invariance within a phoneme class (Blumstein & Stevens 1979; Kewley-Port 1982; 1983; Lahiri et al. 1984; Liberman et al. 1967; Stevens & Blumstein 1978).

In the next section we introduce locus equations, which may represent a partial solution to the noninvariance problem in speech perception, focusing on acoustic cues for perception of stop consonants (/b,d,g,/) across vowel contexts. What is especially appealing and intriguing about locus equations, apart from the much needed sense of order they bring to the noninvariance issue, is their potential parallelism with neuroethological models of combinatorial processing, as presented in section 1.

## 3. Locus equations

A frequency by amplitude display of speech over time (the spectrogram) shows acoustic energy concentrated at specific frequency regions known as formants. Formants represent acoustic resonances of the vocal tract. The specific formant structure of a vowel helps determine its acoustic and hence phonetic quality. During production of isolated



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

vowels, the formants (F1, F2, F3, etc.) are relatively steady. When articulatory movements occur – for example going from a stop consonant such as /d/ to a vowel such as /a/ – the formant frequencies change in response to the changing filter function of the vocal tract. These frequency modulations, known as formant transitions, occur in the vicinity of the consonant-vowel (CV) interface. The second formant (F2) transition is perhaps the single most important cue in speech perception (Liberman et al. 1967), as it best encodes the dynamic consonant-to-vowel gesture from the moment of consonantal release to the vowel nucleus or midpoint. Locus equations are derived by plotting the frequency values of F2 transition onsets and the related F2 vowel midpoint in CV utterances.

More specifically, locus equations are linear regression fits made to scatterplots of coordinates representing, separately for each consonantal category, all F2 transition onsets, plotted on the y-axis, in relation to midvowel frequencies, plotted on the x-axis.[4] Figure 3 illustrates how a locus equation scatterplot is derived from spectrographic measurements. Three sample syllables are shown in spectrographic form – "daught," "dut," and "deet." The arrows on the spectrograms indicate the locations in the F2 where F2 onset and F2 vowel frequencies are measured. These $\langle x,y \rangle$ coordinates are then plotted for the various vowel contexts and, for a given stop consonant category, fitted with a line expressed as F2 onset = k ∗ F2 vowel + *c*, where *k* and *c* are slope and y-intercept, respectively. Note that each data point in a locus equation plot represents an F2 transition. The transitions are thus compactly parameterized via their onsets and offsets (i.e., endpoints).

By displaying all variants of a given phonological category (e.g., initial *d* in a range of vowel contexts, as in "deed," "did," "dade," "dead," "dad," "dode," "dude," "dud") in one scatterplot, a dramatic orderliness, not evident at the level of single speech tokens, emerges for the first time, in the form of tight clustering about the iso-stop regression line. Each line characterizes, in acoustic space, a place of articulation category (e.g., in English – labial /b/, alveolar /d/, and velar /g/). Place of articulation refers to a location along an anterior-to-posterior dimension of the vocal tract, where the articulatory constriction or occlusion is formed (e.g., occlusion of the vocal tract at the lips for /b/, tongue tip against the alveolar ridge behind the incisors for /d/, or tongue body on the velar or soft palate area for /g/). For syllable-initial oral stops (/b,d,g/), the frequency of F2 onset has been found to vary as a linear function of F2 in the midvowel nucleus (see Sussman 1989; 1994; Sussman et al. 1991). In addition, the particular linear function relating these two parameters is itself a function of place of articulation. Labials have been found to have the steepest regression functions, followed by velars, and then alveolars. R2 values usually exceed .90, and standard errors of estimate are very small – 88 Hz, 57 Hz, and 108 Hz for /b,d,g/, respectively (mean standard errors of estimate, SEs, pooled across ten male speakers). Examples of locus equations for a representative English native speaker producing syllable-initial stops /b,d,g/ with 10 vowel contexts are displayed in Figure 4.[5] In contrast to the homogeneous scatterplots for /b/ and /d/, /g/ has two distinct clusters of points, and each cluster is linearly arranged. Phoneticians have long described two allophonic variants of /g/ – a palatal [g] preceding front vowels /i, ɪ, e, ɛ, æ/ (phonetically characterized by being produced with relatively anterior tongue placements

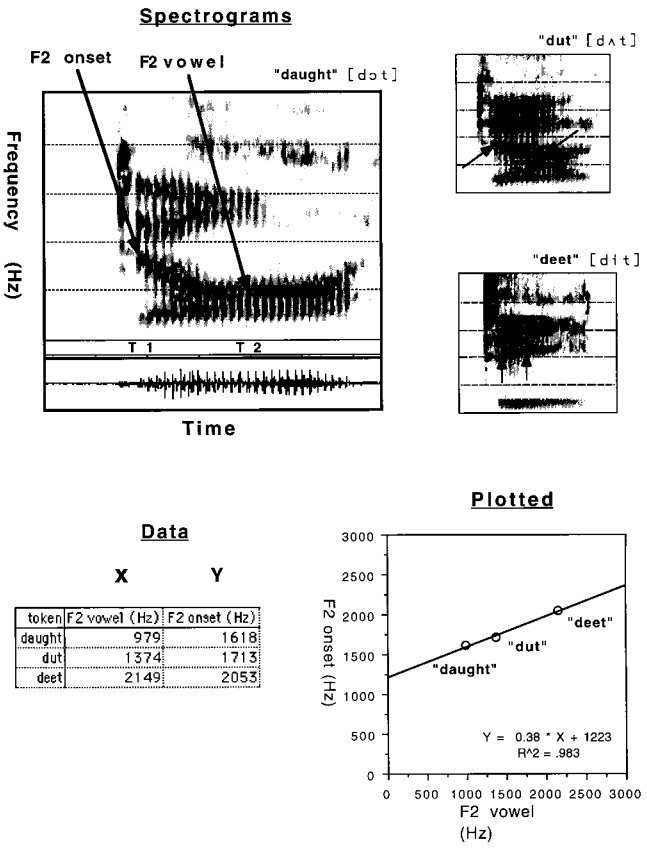



Figure 3. Spectrograms, sample data for F2 onset and F2 vowel, and a plot showing how locus equation regression functions are derived.

as in the vowel sounds in the words "beet," "bit," "bait," "bet," "bat," respectively), and a velar [g] preceding back vowels /a, ɔ, o, u, ʌ/ produced with more posterior tongue positions (as in the vowel sounds in the words "bot," "bought," "boat," "boot," "but," respectively).

The typical locus equation form has been validated cross-linguistically. Sussman et al. (1993) analyzed languages with two- (Thai) and with four- (Cairene Arabic and Urdu) voiced stop place contrasts. Once again, locus equation slope/y-intercept means were found to be significantly different as a function of stop place of articulation; and scatterplots for each category were linear, with little noise for every speaker. Sussman et al. (1992) applied the locus equation metric to children and found linear low-noise scatterplots for /b/, /d/, and /g/ in the acoustic output of 3 to 5 year olds, with slope/y-intercepts reflecting stop place of articulation.

Figure 5 shows "prototypical" regression functions obtained by averaging F2 onset and F2 offset frequencies for all stop plus vowel contexts across 10 male and 10 female speakers (data from Sussman et al. 1991). There are two areas of overlap among the lines – /d/ and /g/ in back vowel contexts (F2 vowel in the vicinity of 1300 Hz) and all three stops in high front vowel space (F2 vowel > 2,500 Hz) – therefore, in terms of F2 transition endpoints, the stops are perfectly confusable in those regions of overlap. However, the F2 transition is but one component of a redundant cue set signaling stop place (the stop release burst preceding the F2 transition is another crucial cue), so locus equations



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

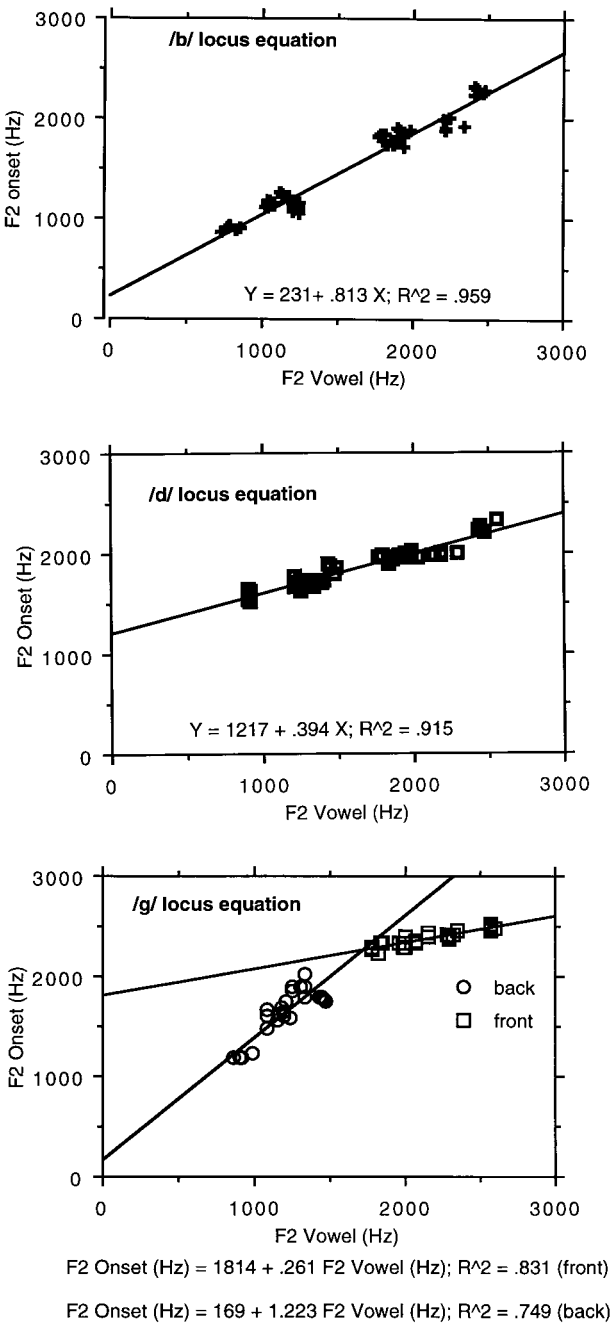



Figure 4. Representative locus equations for syllable-initial labial /b/, alveolar /d/, and velar /g/, each across 10 vowel contexts.

need not by themselves solve the vowel context nonin-variance problem. Nevertheless, as was very plainly shown 43 years ago (Delattre et al. 1955), the F2 transition is an important cue for stop place of articulation. The question ever since has been what parameters of the F2 transition are encoded, and how can the diverse transitions characteristic of a particular stop consonant in its various vowel contexts be organized into a single perceptual entity by the auditory system. The particular role of locus equations in a theory of stop consonant place perception is addressed fully in section 6.1. There we suggest that locus equations represent rules for computing a feature we are calling "vowel-normalized F2 transitions," which then contributes,

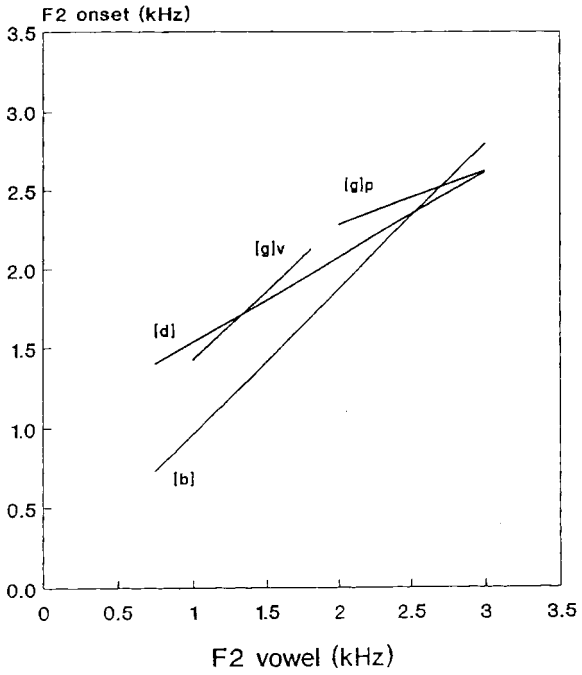



Figure 5. Prototypical locus equations derived by pooling frequency coordinates for 20 speakers. Velar /g/ has been divided into two allophonic groupings – [g]v(elar) for /g/ preceding back vowels and [g]p(alatal) for /g/ preceding front vowels.

along with other cues, to stop consonant place of articulation perception.

### 3.1. The parameterization of stop consonants in terms of locus equation regression coefficients

Locus equations are derived, for a given stop consonant, over an entire set of vowel contexts. The lawful variability seen at this level is enhanced when we proceed to cluster the functions themselves, as derived for different speakers. This was statistically verified by comparing classification results from discriminant analyses using two different sets of predictor variables across a speaker population of 10 male and 10 female adults (Sussman et al. 1991). When token-level predictors, F2 onset and offset frequencies, were used for each gender group, correct token classification rates for labial, alveolar, and velar stop place categories were 82%, 78%, and 67%, respectively, for female speakers, and 84%, 81%, and 69% for male speakers (chance = 33%). When category-level variables, locus equation slopes and y-intercepts, were used as predictors (for /b,d,g/ functions pooled across gender groups), a perfect (100%) classification rate of the 60 functions into labial, alveolar, and velar stop place categories was achieved. Celdran and Villalba (1995) – using 5 female and 5 male adult speakers – have recently replicated this result of 100% correct classification of stop place, using locus equation slopes as predictors for place categories in Spanish stops (/b, p/, /d, t/, /g, k/). Figure 6 illustrates how speaker functions are successfully segregated by place of articulation in a regression coefficients space (data from Sussman et al. 1991). Each point represents a single speaker. Though speakers vary within a given stop place cluster, the categories are for the most part nonoverlapping. The lack of overlap between stop places of



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

articulation at this level of abstraction does not solve the problem of overlap in the transition endpoint space, namely, we do not interpret the distinctness of the b, d, and g clusters in Figure 6 to mean that slope and y-intercept could be invariant specifiers (in the sense of Fowler 1994) for the place of articulation of single tokens, which do not have function-level characteristics.

Another important attribute of locus equation slopes is that they quantify, for each speaker, the overall degree of coarticulation, or articulatory overlap of the following vowel with the preceding stop consonant. This aspect of locus equations was initially described by Krull (1988). No coarticulation between vowel and consonant is reflected by a slope = 0 (k = 0, F2 onset = c); maximal coarticulation (k = 1.0, F2 onset = F2 vowel) occurs when F2 onsets are identical to each different vowel steady state (see Sussman et al. 1993, Fig. 10). Speakers evidence slope values varying within these two hypothetical limits. Prior to locus equations, the degree of coarticulation being used by a speaker had never been quantified. Historically, coarticulation has always been viewed as the culprit responsible for context-induced variation and hence the noninvariance problem (Liberman & Mattingly 1985; Liberman et al. 1967). Locus equations, however, present the opposite view: a lawful variance in the acoustic manifestation of coarticulation that is consistent within a stop place category and distinctive across categories, so that degree of coarticulation, as indexed by locus equation slope, becomes a parameter of the categories.

### *3.2. Robustness of the locus equation phenomenon*

It is important to show that methodological factors and parameters contributing to normal speaker variation, such as gender, speaking style, and speech rate, do not disrupt the lawful form of locus equation functions and their ability to serve as acoustic indices of place of articulation. In addition, application of the locus equation metric to consonant classes beyond voiced oral stops (/b/, /d/, /g/) would support a more phonetically universal role for locus equations as place of articulation descriptors.

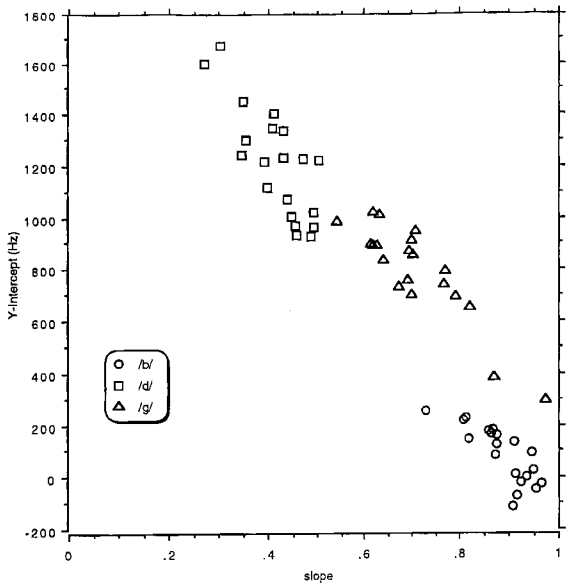



Figure 6. A plot of regression coefficient space for locus equation functions from 20 speakers (10 male and 10 female). Slope and y-intercept coordinates do not overlap for /b/ versus /d/ versus /g/ functions across a varied group of speakers. These derived and higher-order locus equation abstractions of CV categories reflect a lawful variability not seen at the level of individual speech sounds.

**3.2.1. Methodological consistency.** The exact time at which F2 vowel frequencies are sampled does not seem to be too important. In the locus equation studies described above, the measurement point for F2 vowel frequency was the subjectively determined midpoint of the F2 resonance (when the resonance pattern was steady-state, or diagonally rising/falling). If the F2 pattern was parabolic, a maximum/minimum point was chosen. In contrast, Nearey and Shammass (1987) measured F2 vowel frequency at a constant interval (60 msec) after stop release. Analyses of these frequency coordinates for 10 speakers of Canadian English showed a strong correlation (mean $R^2 > .90$) with slope/y-intercept of the regression functions systematically varying as a function of stop place of articulation.

**3.2.2. Effects of gender, speaking style, and speech rate on locus equations.** In comparing locus equation coefficients for 10 male and 10 female speakers, Sussman et al. (1991) found no significant difference in slope for corresponding consonants as a function of gender. In general, mean frequency coordinates pooled across gender groups tightly clustered around the single regression function with female coordinates lying slightly above the gender-pooled line and male coordinates slightly below the line. Locus equation coefficients also remain stable across alterations in speaking style. Krull (1989) compared locus equations obtained from citation-style formal speech to those from more spontaneous informal speech. Five male speakers producing syllable-initial /d/, /n/, /l/, /b/, and /m/, followed by a varied set of Swedish vowels, were analyzed to derive locus equation functions. In general, the reduced form of spontaneous speech was characterized by slightly steeper slopes reflecting a small increase in coarticulation compared to the more formal "laboratory" speech (mean slope difference between speaking styles across all consonants was only .06). Most important, speaking style variation did not perturb locus equation slopes in their role as phonetic descriptors of consonant place. The dentals-alveolars /d, n, l/ had a mean slope across speaking styles of .35, and the labials /b, m/ had a mean slope of .71.

Speaking rate is another aspect of speaker-induced variation that appears to exert a limited effect on locus equation parameters. Kugel et al. (1995) analyzed locus equation slopes obtained from 10 male and 10 female speakers, for fast versus slow speaking rates. Significant effects as a result of altered speech rates were not found, only a significant effect based on the consonant place (/b,d,g/).

**3.2.3. Extending locus equations across manner classes.** Of considerable interest to speech theorists is the ability of the locus equation metric to be extended beyond voiced oral stops /b, d, g/ to other consonant manner classes, such as fricatives, nasals, and voiceless stops. Figure 7 shows locus equations from a representative speaker producing fricative (/v, s, z, ʃ/) plus vowel tokens (data taken from Sussman 1994). Note that all functions are characterized by unique slopes, extremely high R2 values, and tight clustering of coordinates about the regression lines. The ability of



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

locus equation coefficients to reflect systematically place of articulation within a fricative series was also shown by Fowler (1994). The progression of place of articulation from labiodental /v/ to interdental /ð/ to alveolar /z/ to palatal /ʒ/ was nicely captured by decreasing slopes and increasing y-intercepts – .73/337 Hz, .50/903 Hz, .41/1,078 Hz, and .34/1,408 Hz, respectively. However, when testing two consonants from different manner classes that shared the same alveolar place of articulation, a significant slope difference was reported between voiced stop /d/ (slope = .47) and voiced fricative /z/ (slope = .42; Fowler 1994). Sussman and Shore (1996) recently explored this issue by analyzing a diverse set of consonants varying across several manner classes but all sharing the same "alveolar" place feature – voiced stop /d/, voiceless aspirated stop /t/, nasal /n/, voiced fricative /z/, and voiceless fricative /s/. Locus equations were derived from 50 tokens (10 vowels × 5 repetitions) for each phonetic category, for 22 speakers. Slope and y-intercept values were entered into a doubly dependent multivariate analysis that yielded a significant effect for manner class ($F[10, 208] = 68.31$, $p < .001$). Post-hoc tests, however, showed that slope and y-intercept means among /d/, /z/, and /n/ were not significantly different, nor for /d/ versus /t/ when F2 onset measure ment points were equated (by taking into account the lengthy aspiration interval following stop release for /t/). In an additional test of whether or not locus equation coefficients could serve as general descriptors of consonantal place across manner classes, a discriminant analysis was conducted using slopes and y-intercepts obtained from labial /b/, alveolar /d, t, n, z, s/, and velar /g/ functions as predictor variables for assignment to one of three place-of-articulation categories. Of interest was whether the diverse consonants from the various manner classes would be similarly categorized as alveolars and kept apart from labials and velars. Correct classification as "alveolars" was 87.1% (115 of 132 total cases) despite the wide variety of manner and voice conditions of the five alveolar consonants.[6]

**3.2.4. Locus equations derived from compensatory articulation.** In another study (Sussman et al. 1995) locus equations were found to be extremely robust under conditions of articulatory perturbation, that is, speaking with bite blocks inserted between molar teeth (behaviorally similar to speaking while clenching a pipe stem). Individual speakers served as their own controls, as normal productions by the same speakers were compared to bite block productions of identical utterances. Examples of locus equations for /b, d, g/ in normal versus bite block conditions for a representative speaker are shown in Figure 8. It can be seen that for all functions, normal and bite block data are virtually indistinguishable.

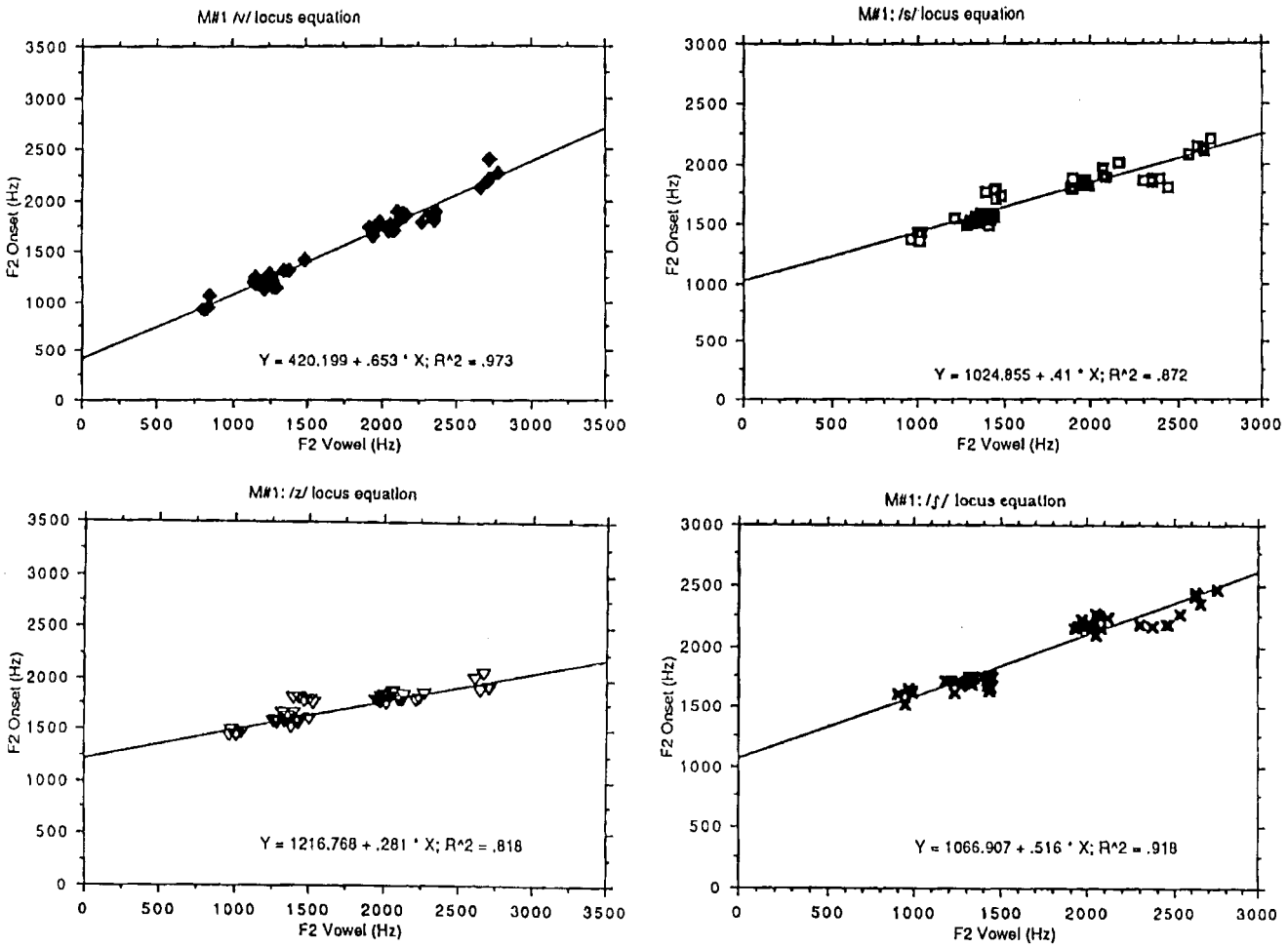



Figure 7. Representative locus equation plots for a speaker producing fricative plus vowel tokens. Initial consonants are /v/, /s/, /z/, and /ʃ/ (as in *sh*oe), each with 10 vowel contexts.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

Thus, it was found that altering the specific kinematics of articulation while maintaining perceptual equivalence had no effect on the degree of correlation between the locus equation acoustic variables nor on the particular linear relationship between them. The bite block results suggest that the specific articulatory commands used to produce stop closures and vowel shapes do not affect the nature of the F2 transition endpoint relationship. It has long been known that speakers operate within a motor equivalence framework (Hebb 1949) to achieve quasi-constant goals via a multitude of movement trajectories and strategies. The results of this experiment suggest that the articulatory system's quasi-constant goal in this case might be to maintain the integrity of the F2 transition endpoint relationship, presumably for purposes of perceptual equivalence.

### 3.3. Limits to the robustness of the locus equation phenomenon

We should note that the high correlation and linear relationship between transition onset and offset are not proper ties generalized across all formant transitions but rather are

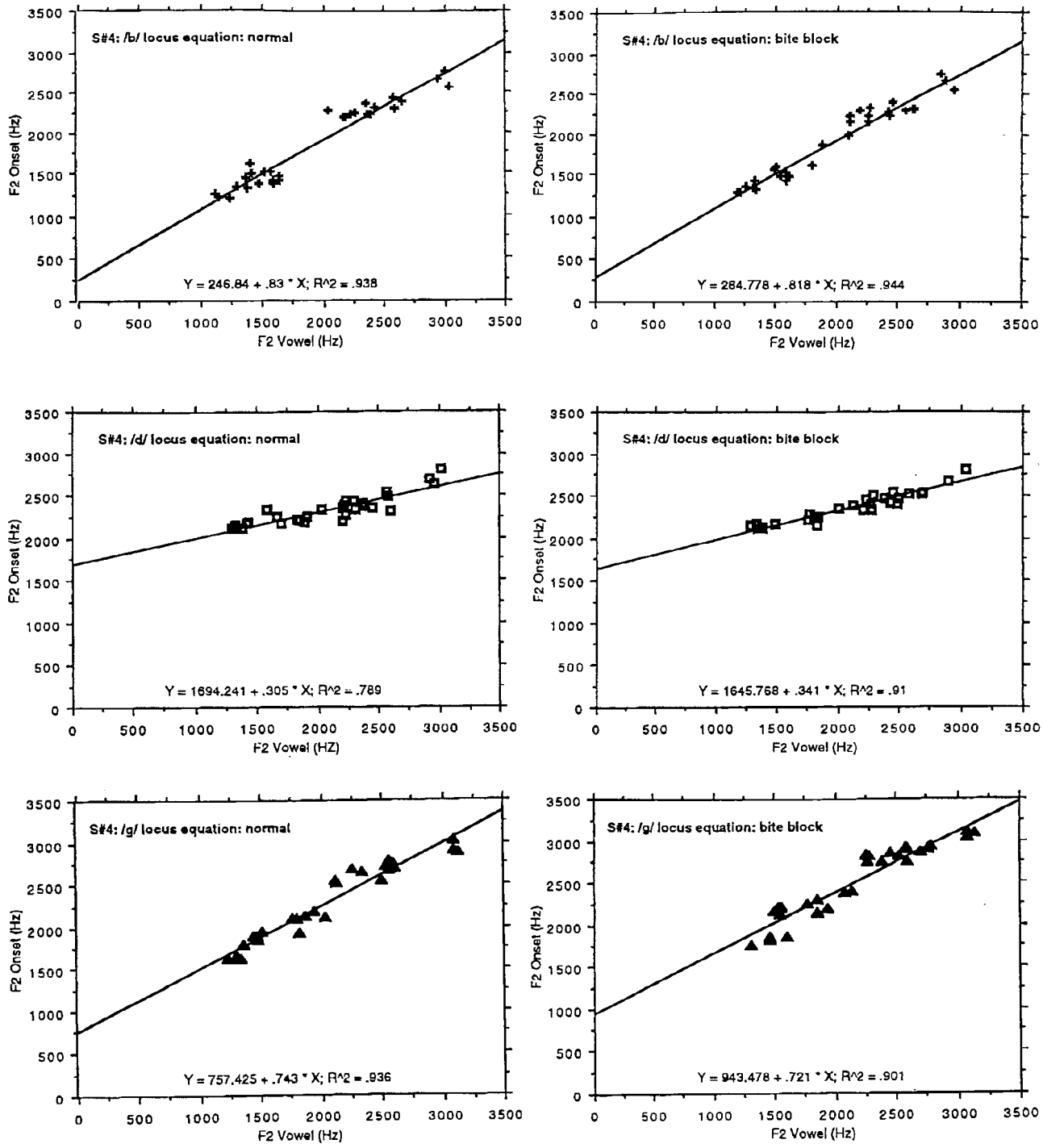



Figure 8. Representative locus equation plots comparing normal speech to speaking with a bite block for a single speaker. Slopes and y-intercepts are nearly identical in the two conditions, and linearity is preserved despite the perturbation condition.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

exclusive properties of F2. Figure 9 shows representative locus equation scatterplots for F3 onset (Hz) in relation to F3 offset (Hz). F3 locus equation data do not resemble the characteristic scatterplots of F2 data. Correlations are reduced, and standard errors of estimate are increased, over F2 plots.

**3.3.1. Locus equations in canonical infant babbling.** Adult and child (aged 3 to 5) speakers produce the stereotyped locus equation plots (Sussman et al. 1991; 1992; 1993). Are these high correlations and linear relationships physically unavoidable and hence present in the earliest output of the prelinguistic child? One segment of our research program is aimed at investigating canonical babbling in infants. At around six to eight months, normal hearing infants initiate a vocal babbling stage where consonant-vowel syllablelike utterances are produced in a reduplicated fashion (Oller 1978). Investigating the acoustic structure of infant babbling permits us to ascertain whether linear trends are present in the "primordial CVs" produced at this earliest stage of articulatory and phonological development. If so, we can conclude that whatever articulatory parameters are responsible for the linear trend, they are manifested very early, well before phonemic categories develop. Representative locus equations derived from babbling samples recorded from one infant, spanning ages 7 to 9 months, are shown in Figure 10. There were 98 "bV," 118 "dV," and 79 "gV" tokens obtained from the 12 hours of recordings spanning this 3-month interval. The relationship between F2 transition onset and offset is somewhat noisy, as can be seen by the large standard errors of estimate (264 Hz, 330 Hz, and 357 Hz for "bV," "dV," and "gV" utterances, respectively). Thus, the prelinguistic CV utterances of this infant, as well as data from another child measured at 12 months (Sussman et al. 1996), do not conform to the typical locus equation pattern observed by the age of 3 years (Sussman et al. 1992).

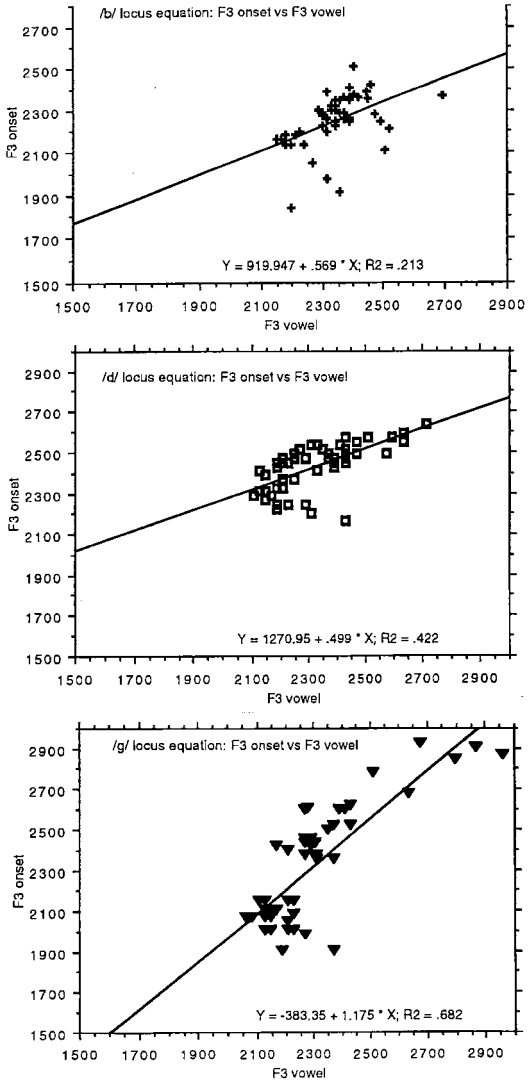



Figure 9. Representative F3 locus equations for syllable-initial /b/, /d/, and /g/ across 10 vowel contexts. Scatterplots are noisier than is consistently observed for F2 locus equations.

**3.3.2. Locus equations in developmentally apractic speakers.** The rationale for investigating communicatively disordered speakers is to determine whether speech motor control factors can, independently of the filtering properties of the human vocal tract, affect locus equation linearity or noisiness. If speakers with severe articulatory problems but intact vocal tracts manage to produce distinctive and linear scatterplots, it would most likely suggest that highly correlated, linearly related F2 onsets and offsets are a highly buffered outcome depending primarily on vocal tract filtering properties. If the locus equation plots are nonlinear and/or noisy, or undifferentiated as a function of stop place, this would support the contention that normal motor control strategies contribute significantly to the typical form of locus equations.

Developmental apraxia of speech (DAS) is a congenital disorder in the ability to program speech movements in the absence of neuromuscular pathology. The phonological output difficulties of those affected lead to poor intelligibility of their speech. Acoustic measures were obtained from two children (DL and MG) clinically diagnosed with DAS, but both having /b/, /d/, and /g/ target consonants in their phonological repertoires. MG was $4\frac{1}{2}$ years old and DL was 5 at the time of recording. Each child was asked to repeat /bVt/, /dVt/, and /gVt/ syllables in an imitation task with the 10 different vowel targets used by Sussman et al. (1991). In terms of acoustically analyzable productions, MG produced 26 /b/ tokens, 26 /d/ tokens, and 21 /g/ tokens; DL produced 28 /b/, 28 /d/, and 28 /g/ tokens. Figure 11 shows locus equation plots for DL and MG. Slope values were poor descriptors of stop place, and the scatterplots showed only moderate degrees of correlation, as data points did not cluster tightly around the regression line. R2 values range from a low of .25 (MG, /d/) to a high of only .70 (DL, /d/). Standard errors of estimate were large compared to values obtained for age-matched normal children. SEs for DL by place of articulation were 188 Hz, 272 Hz, and 232 Hz, for /b,d,g/, respectively, and corresponding values for MG were 287 Hz, 199 Hz, and 257 Hz. The poor intelligibility of DAS children is a predictable correlate of their atypical locus equations.

We have found that the continued refinement of locus equation scatterplots (quantified by decreasing SEs) parallels the maturation level of the speaker. Figure 12 shows a plot of standard errors of estimate versus age from a wide assortment of speakers. From left to right along the abscissa



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

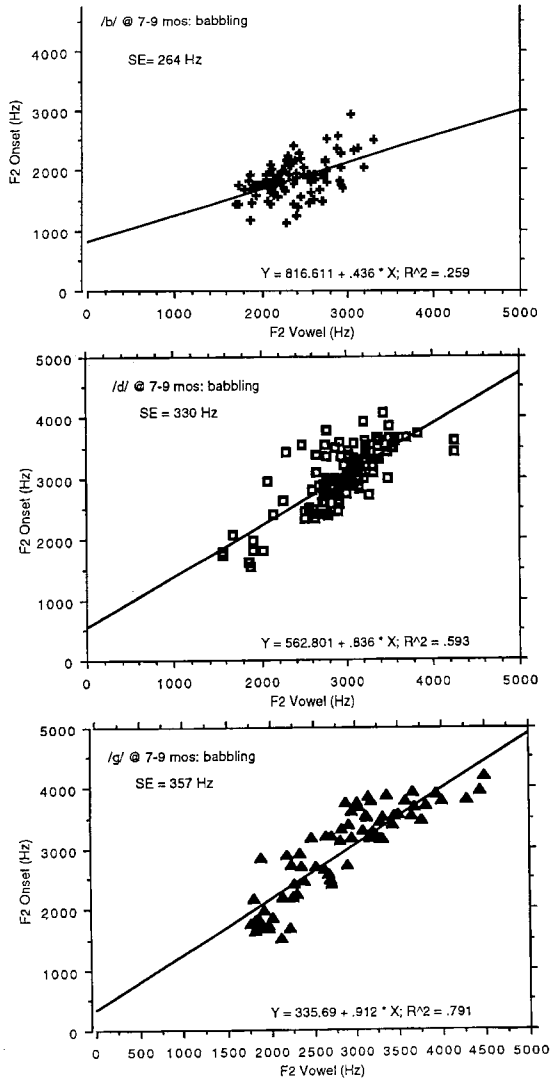



Figure 10. Locus equations for canonical babbling utterances beginning with "b," "d," and "g" obtained from a female infant at the age of 7 to 9 months. Standard errors of estimate reflect considerable variation of F2 onset F2 offset coordinates about the regression lines, and slope values do not correspond to normal adult values.

the SEs, averaged across stop places of articulation (/b,d,g/) and speakers, are shown for infant babbles at 7 to 9 months, two DAS speakers at 3 years old (N = 2), 4 years old (N = 7), 5 years old (N = 7) (Sussman et al. 1992), and adults (N = 20) (Sussman et al. 1991). Infant CV babbles had the highest SE at 317 Hz. The DAS children had, at around 5 years of age, the second highest SE (239 Hz), followed by normal 3 to 5 year olds in a systematically decreasing order, and least for adults (97.5 Hz). It appears that deficits in articulatory motor control affect the noise level in locus equation data, as well as attainment of appropriate slope values for stop place contrasts that approach adult standards. The greater scatter of points shown by DAS children and the clear increase in definition of the linear trend with maturation of normal speakers strongly suggest that development of precise motor programming skills contributes appreciably to the "prototypical" form of locus equation plots.

## 4. Question: Why are F2 onset and F2 vowel normally so highly correlated and linearly related?

So far we have established that the high correlation and linearity typical of F2 locus equation data is an extremely robust feature of consonant-vowel output by the human vocal tract, both reproducible and general. The high correlation and linearity are preserved across languages, across consonantal manner classes, across speakers of various ages and both genders, and across speaking conditions (informal vs. formal, fast vs. slow, bite block vs. normal). The linear trend, however, appears to be incompletely developed in a prelinguistic infant and in older children with developmental apraxia of speech. Having confirmed that the locus equation phenomenon is bona fide, we are ready to concentrate on a more theoretical question with which we will be concerned for the balance of this article – Why are F2 onset and F2 vowel normally so highly correlated and linearly related?

It should be noted at the outset that this could actually be two separate questions, namely, that the normally high correlation of these two variables could conceivably have a separate explanation from the linear relationship between them.[7] On the other hand, the hypothesis we will be pursuing most seriously addresses the correlation and linear relationship together, hence we have posed these potentially separate questions in this combined form. In our view, they may have the same answer.

## 5. Articulatory explanations of high correlation and linear relationship between components of the speech signal

Perhaps the simplest sort of explanation would be that the acoustic patterns have no function but rather arise as a byproduct or epiphenomenon of the speech production system. It may forestall some confusion if we bear in mind that in a limited sense there must be an articulatory explanation for locus equations, since the locus equation relationship is an acoustic phenomenon produced by an articulatory system. However, we argue that the articulatory system may actually be going to considerable trouble to achieve a uniform locus equation slope, or constant ratio of F2 onset to F2 vowel within a consonant across vowel contexts. It seems highly unlikely that the speech motor system would be doing this if it were simply a nonfunctional epiphenomenon. If it can be confirmed that the acoustic pattern is indeed being optimized, the articulatory account would be of how it is optimized, not why.

### *5.1. Simulated locus equations using a vocal tract area function model*

Does the locus equation pattern arise as an inherent characteristic of the filtering properties of human vocal tracts? If so, then simulations of consonant-vowel syllables using an accurate vocal tract area function model should yield the typical locus equation plots. This hypothesis can be directly tested by using a computer-implemented model of the human vocal tract and obtaining simulated F2 onset and F2 vowel frequencies for stop plus vowel sequences beginning with /b,d,g/ and followed by a wide assortment of vowels. The model used here, the distinctive regions model (DRM)



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

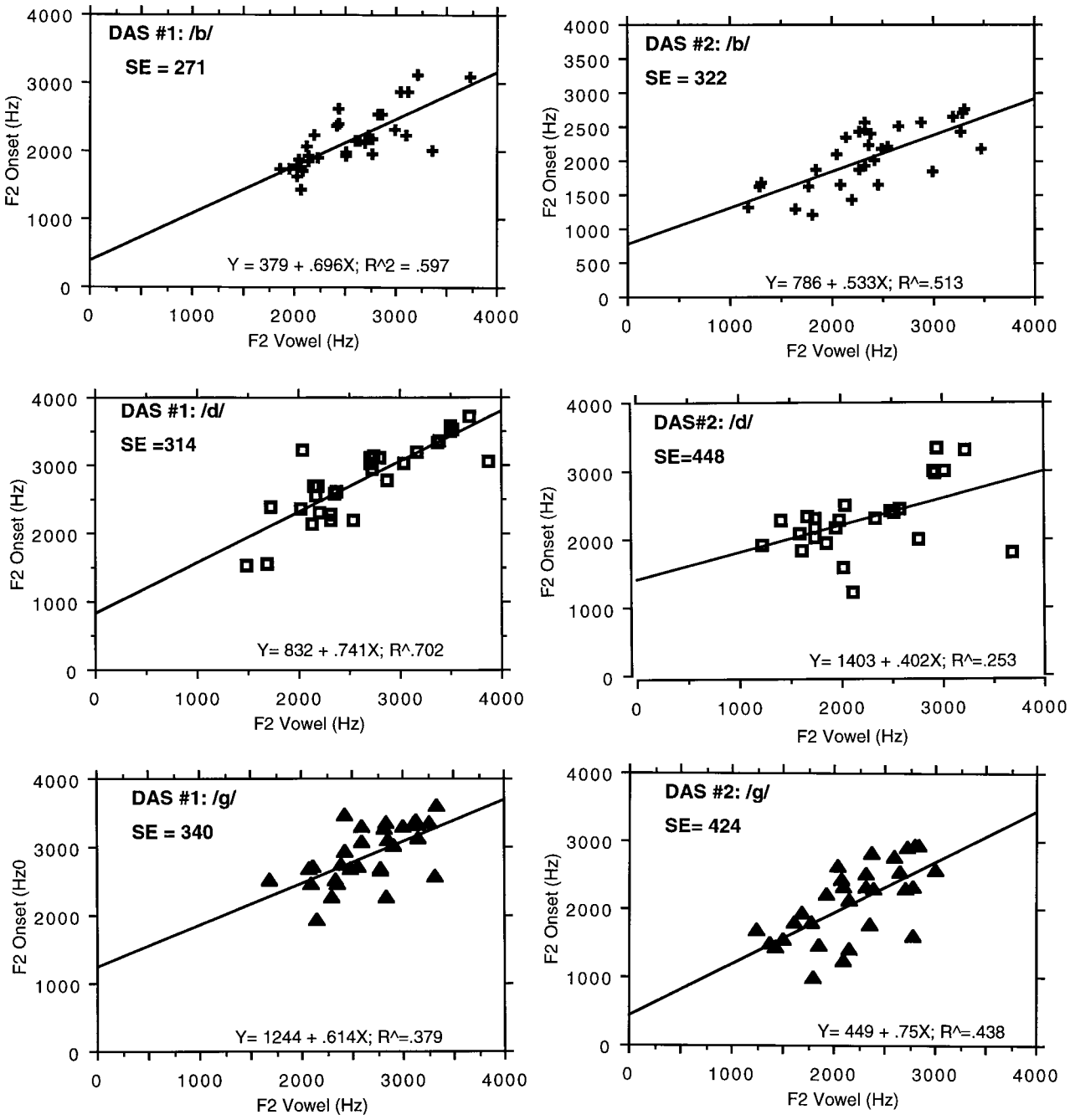



Figure 11. Locus equations for "b," "d," and "g" obtained from CV utterances produced by two children, aged 5 (DL) and 4½ (MG), diagnosed with developmental apraxia of speech.

(Carré & Mrayati 1992), is based on an acoustic tube segmented lengthwise into eight distinctive regions.

Formant frequencies are altered in this model by modifications in the cross-sectional areas of specific regions, from the glottis to the lips. In effect, the "pinches" on the tube configuration simulate tongue constrictions or vocal tract occlusions for stop consonants superimposed on vowel-to-vowel gestures. Figure 13 illustrates simulated locus equations for /b,d,g/ preceding 11 French vowels. All three functions are extremely linear. Using the standard errors of estimate as an index of clustering along the regression function, the following values were obtained: /b/ = 177 Hz, /d/ = 89 Hz, and /g/ = 196 Hz. Thus, a model derived from the acoustics of tubes effectively produces a linear relationship between F2 transition onset and offset frequencies.[8] These models do confirm that the human vocal tract is configured to produce these patterns, but they say little about why human vocal tracts are so configured – whether it might be accidental, part of some non-speech-related adaptation, part of a speech-related adaptation having nothing to do with perception, or part of a speech production system coadaptation to speech perception. At this point, the articulatory modelers are somewhat mystified (R. Carré, personal communication, 1995; B. Lindblom, personal communication, 1993).

The question then becomes – What is crucial about these configurations that produces the locus equation acoustic pattern, and what, if anything, might be enforcing this



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

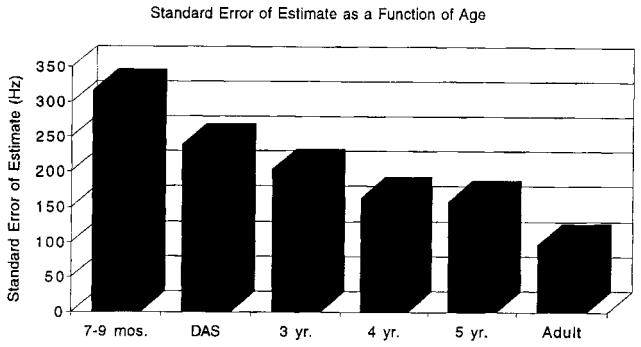



Figure 12. Bar graph showing the mean standard errors of estimate of locus equation functions averaged across stop place of articulation categories ([b,d,g]) obtained from several speaker groups – an infant's babbling over the 3-month interval of 7 to 9 months, two children diagnosed with developmental apraxia of speech (DAS), two 3-year-old children, seven 4-year olds, seven 5-year-olds, and 20 adults.

pattern? It is known, for example, that the hominid two-tube architecture (pharynx and oral tract) has the effect of expanding the vowel acoustic space compared to the single-tube architecture of ancestral primates (Lieberman 1984). In locus equation terms, the two-tube plan extends the range of the independent variable, F2 vowel. However, this does not explain the extremely constrained relationship between F2 vowel and F2 onset.

### *5.2. The uniform coarticulatory resistance hypothesis*

Fowler (1994, p. 600) provides the following coarticulation-based account of F2 transition onset-offset correlation:

> The functions have a positive slope, because talkers coarticulate – that is, they overlap the production of serially ordered consonants and vowels. Accordingly, if a vowel has a high F2, F2 will also be relatively high at the acoustic onset of the syllable, because vowel production began before consonant release, and vowel production affects the acoustic signal at release. If a vowel has a low F2, F2 will be low at acoustic-syllable onset for the same reason. Therefore, F2v, F2o points tend to fall on a line with positive slope.

This account is sufficient to yield a monotonic relationship between F2 onset and F2 vowel, a general tendency for them to be correlated. Yet what is striking about the locus equation phenomenon is that the degree of correlation and linearity is unusually high. There is almost perfect linearity, and it is stable across many speaking conditions. Fowler goes on to suggest that phonetic segments (e.g., /b, d, g/) have variable levels of resistance to overlap with neighboring segments, but within a place of articulation category there will be a uniform level of coarticulatory resistance, as reflected by the locus equation slope. This idea is flawed on two counts, one empirical in nature and the other deductive.

On the empirical side, the premise that coarticulatory resistance has a uniform value within a consonant is dubious, in view of articulatory studies that have observed variable and vowel-specific degrees of coarticulation. Amerman (1970), in a cinefluorographic analysis of tongue body-tongue tip coarticulation, showed differences in the extent of anticipatory coarticulatory movements as a function of vowel context. Similarly, Sussman et al. (1973) showed unequal degrees of anticipatory mandibular coarticulation (elevation for a medial stop) in vowel-consonant-vowel (VCV) tokens as a direct function of the height of the second vowel. Lindblom (1983) has also demonstrated, in an articulatory model, differential effects of vowel context on synergy constraints for tongue tip–tongue body co-articulation within a given stop place category.

The deductive failing of Fowler's (1994) explanation of locus equation linearity is that, even if coarticulatory resistance were uniform within a consonant and across vowel contexts, the articulatory-to-acoustic transform will not yield a uniform slope, because vocal tract tube resonances do not automatically yield such an acoustic end product (Lindblom, personal communication, 1996). Fowler's hypothesis fails to distinguish between coarticulatory resis-

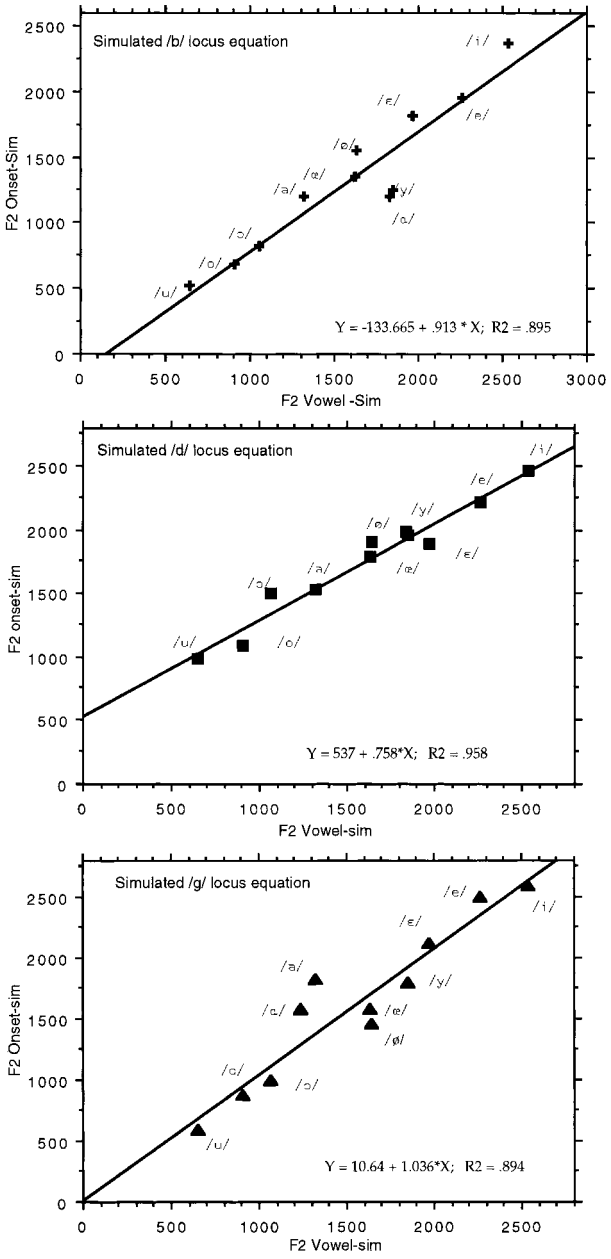



Figure 13. Simulated locus equations using a vocal tract area function model to generate values of F2 onset and F2 vowel for vowel-consonant-vowel utterances with medial /b/, /d/, and /g/, and 11 French vowels. Functions are linear, but slopes do not conform to those of human speakers as realistic coarticulatory variations are not as yet able to be incorporated into the model.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

tance, which is a property of the articulatory domain, and the acoustic ramifications of articulatory events, the realm of locus equation data. Uniform locus equation slopes (acoustic domain) have been interpreted by Fowler as implying uniform coarticulatory resistances (articulatory domain), but the many-to-one, quantal, and nonlinear nature of the mapping from articulatory events to acoustics does not allow a simple conflation of the two levels.

### 5.3. A model incorporating vowel-specific coarticulatory effects

We can accommodate the known vowel-specific coarticulation effects if we conceive of them as adjustments of the articulatory system made in order to achieve a desired acoustic result. Figure 14 illustrates this idea schematically. If a house were to be built on uneven terrain, support pilings of different heights would naturally be used to achieve a level flooring. In terms of locus equations, the level flooring is a uniform F2 onset/F2 vowel ratio, that is, a uniform slope for the locus equation, within a stop place of articulation and across all vowel contexts. This is the desired acoustic result, a vowel normalization of the variable F2 transitions. Support pilings correspond to the mapping of vowel-specific vocal tract area functions to their output resonances (the F2 in this case). The pilings/mappings connect the two levels, articulatory and acoustic. The uneven terrain corresponds to vowel-specific motoric adjustments in consonant-vowel coarticulation (mostly changes in tongue body contour as a result of the effect of the vowel) that alter the vocal tract area functions. By appropriately contouring the ground, one can achieve level flooring, that is, by tailoring degree of coarticulation to each vowel context one can achieve a uniform F2 onset/F2 vowel ratio. Each place-specific locus equation function has a uniform slope, but there are different slopes for different place categories. Thus, to be more complete, Figure 14 should show three separate level floors, one per each place of articulation category. The overall picture that is being presented here is that the articulatory system, across diverse articulators (tongue, lips, jaw, velum), adjusts consonant-vowel coarticulation with respect to the acoustic output in order to fine-tune a feature of that output, the F2 onset/F2 vowel ratio.

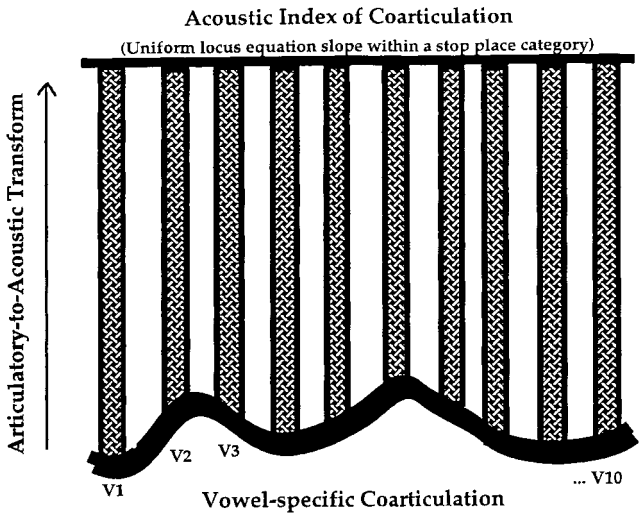



Figure 14. Schematic showing an architectural analogy to vowel-specific coarticulation that is capable of yielding, at the acoustic level, a uniform locus equation slope.

## 6. A perception-based explanation of high correlation and linear relationship between components of the speech signal

Articulatory explanations of the typical form of locus equation data appear at this time to be inadequate and/or incomplete. Moreover, there is evidence both from studies of coarticulation and from our bite block (compensatory articulation) study that the articulatory system adjusts its output in order to preserve the relationship between F2 onset and F2 vowel. A plausible interpretation of this would be that the relationship is normally optimized for some function, probably a communicative one. Could an explanation for this very stable, highly constrained acoustic pattern be forthcoming from speech perception? We will now make that argument. Several diverse but convergent sets of data will be presented in an attempt to support our hypothesis – the orderly output constraint (OOC) – which claims that the high correlation and linear relationship between F2 onset and F2 vowel are functional, satisfying constraints on category representation by auditory neurons that map acoustic features encoding speech. First, we argue for a theory of stop consonant place of articulation perception that includes an auditory system representation of the acoustic information summarized by locus equations. Next, we suggest a formal and evolutionary relationship between the neural computation implied by the aforementioned perception theory and the examples from neuroethology discussed in section 1. On this basis we conjecture that linear relationships with low noise are quite general in the acoustic world of species that do complex sound processing, and that vertebrate 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 OOC). Finally, we explore the pertinence of correlated, linearly related inputs to the "mappability" of those inputs by a type of neural computational system. Our proposed constraint equates orderly output to mappable input, so that indeed "orderly" is defined in terms of "mappable." Thus it is desirable to begin to examine what exactly mappabilility might be.

### 6.1. The perceptual relevance of locus equations

Could the relationship between F2 onset and F2 vowel be of use during speech perception? More specifically, might there be an auditory feature map utilizing F2 onset and F2 vowel to help derive stop place of articulation categories during speech perception? There are several arguments in support of this idea. First, the F2 locus equation phenomenon could reasonably be claimed to be a linguistic universal in the speech of normal adults, as would be expected of an important cue for an important phonemic contrast. Second, when the typical locus equation form is preserved in the face of articulatory perturbation, as with bite block speech, there is perceptual equivalence, but when it breaks down, as with DAS speakers, intelligibility suffers. Third, there is linkage between the cue value for stop consonant place and



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

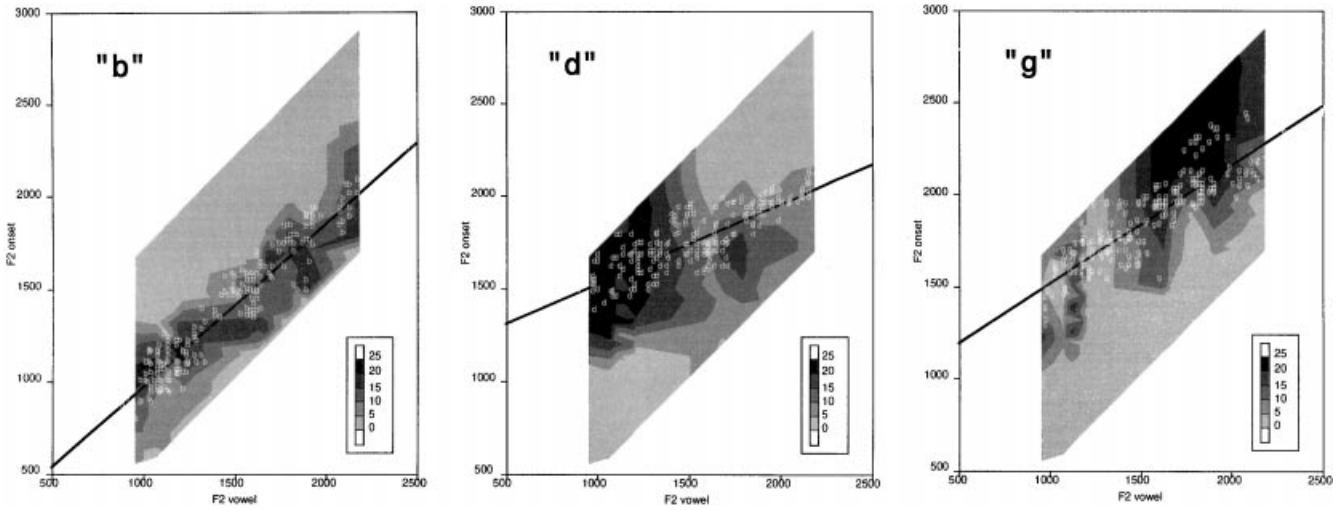



Figure 15. Identification surfaces for "b," "d," and "g" showing correspondences between perceptual results (gray scales) and acoustic speech data (white "b-d-g" letter overlays) in locus equation space. Darkest regions reflect unequivocal identification of a consonant in response to synthesized CV stimuli varying across a full range of F2 onsets for each of 10 vowels.

the degree of relationship between the transition onset and offset: the F2 transition is known to encode important cues for stop place and shows the locus equation phenomenon, whereas the F3 transition is a much weaker cue for stop place and does not show the locus equation phenomenon. Fourth, computational experiments in which time-delayed neural networks were fed sampled spoken consonant-vowel waveforms, tasked to classify the consonants by place of articulation, and then analyzed to determine which parts of the input were most effective for the task, showed that the parts of the signal most informative about consonant place of articulation were the F2 onset and F2 vowel frequencies (Hinton & Lang 1988; McDermott & Katagiri 1988; Unnikrishnan et al. 1988; Waibel et al. 1987; Watrous 1988). Fifth, neural substrates suitable for the task of processing a frequency-modulated signal onset and offset in combination have already been demonstrated in animal models (Fitzpatrick et al. 1993). Finally, the most direct type of experiment, in which F2 onset and F2 vowel are varied orthogonally in synthesized consonant-vowel tokens submitted to human subjects for identification, has been carried out twice, both times indicating strong cue value of the F2 transition onset-offset combination for consonant place of articulation. The first of these studies (Liberman et al. 1954) was interpreted without regard to locus equations, which were not discovered until afterwards (Lindblom 1963a). A more recent study (Fruchter 1994) was able to relate the acoustic phonetic space of locus equations to the corresponding perceptual space of human listeners.

Fruchter (1994) orthogonally varied F2 onset frequencies across 10 vowel contexts in synthesized (5 formants) consonant-vowel syllables with no burst. Stimuli were presented to listeners for identification, and identification frequencies were then tabulated (maximum = 24) and pooled across subjects (N = 3) to yield "identification surfaces" for each place of articulation ("b," "d," or "g"). The identification data, as shown in Figure 15, are rendered as a stepped gray scale in a manner similar to the amplitude ($z$) axis of a spectrogram (the $x$ and $y$ dimensions are simply locus equation space – F2 vowel × F2 onset). Superimposed over the perceptual results are token-level acoustic data (in white) from five native English-speaking males producing "beat, bit, bait, . . . , deet, dit, date, . . . , geet, git, gate, . . . " (data from Sussman et al. 1991). The overlays allow appraisal of the correspondences between the distribution of the acoustic data and features of the perception data. There are clear "peaks" in the three surfaces where a given stop consonant perception dominates – "b" dominates at low F2 onsets across the entire vowel space, but especially at F2 vowel = 1,000 Hz to 1,250 Hz, "d" dominates at F2 onsets spanning 1,250 Hz to 2,000 Hz for back vowel space (1,000 Hz to 1,500 Hz), and "g" dominates at high F2 onsets above 2,000 Hz for F2 vowel > 1,500 Hz. The way in which the sampled acoustic space is partitioned among the three stops is schematized in the "territory map" of Figure 16A. The results of this preliminary perception study closely resemble identification results obtained with only two-formant synthesis in the earlier perception study (Liberman et al. 1954).

The notion that a map of locus equation space somewhere in the auditory system could contribute significantly to consonant place identification is supported by the good match between the locus equation acoustic data and corresponding peaks of the identification surfaces. The darkest areas, indicating unequivocal identification of particular stops, can be thought of as analogous to partial "phonological homunculi" (at least as can be envisioned in these acoustic dimensions), whereas the overlaid acoustic data could represent the sensory inputs that organize the homunculi. Recall, as indicated in Figure 16B, where the acoustic data for all three stops are combined, that there are regions of overlap or competition between the stops in locus equation space. Essentially, [d] and [b] data overlap in front vowel contexts, whereas [d] and [g] data overlap in the back vowel region. A dominance hierarchy hypothesis, schematized in Figure 16C, is offered to help conceptualize the relationship between the token-level acoustic data (Fig. 16B) and the identification patterns for the burstless stimuli used in this study (summarized in Fig. 16A). It can be seen that front vowel [d] and back vowel [g] are in a sense missing from the identification surfaces. In Figure 16C each outline represents a particular stop consonant's cloud



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

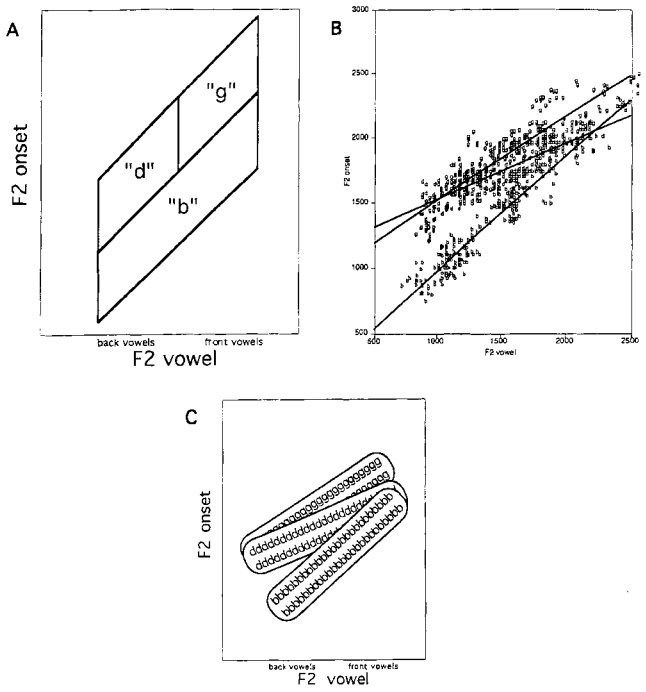



Figure 16. **A.** Schematic of "b-d-g" identification territory map in locus equation acoustic space.
**B.** Acoustic scatterplots of [b, d, g] locus equation data across 10 vowels showing areas of overlap in production space.
**C.** Schematic of a "perceptual dominance" hypothesis for burstless stimuli with [b] > [d] in front vowel space and [d] > [g] in back vowel space.

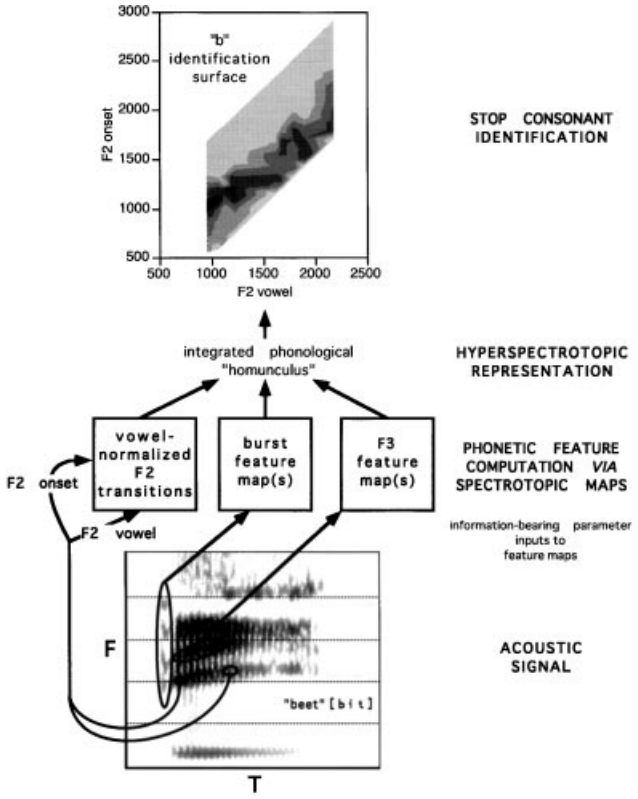



Figure 17. A bottom-up model of stop consonant place identification, including spectrotopic phonetic feature maps that combinatorially process select/critical features of the acoustic signal (e.g., F2 onset and F2 vowel yield a vowel-normalized F2 transition). Higher-order integration of multiple cues establish a phonological "homunculus."

of points in locus equation acoustic space. It is an abstract rendition of 16B, except that the opacity of the "clouds" models the postulated dominance effect for perception in regions of acoustic overlap. The proposed dominance hierarchy would be b > d, d > g. The idea is that a *b* identification will tend to prevail when tokens fall in the region of overlap between [b] and [d] (in the front vowel region), while, likewise, a *d* identification will tend to prevail when tokens fall in the region of overlap between [d] and [g] (in the back vowel region). The cues that allow normal identification of [d] in front vowel contexts and [g] in back vowel contexts are not to be found in this acoustic space. The stops [b] and [g] do not overlap, so their dominance relation is irrelevant.

Of course, other information, such as the release burst, shape of the onset spectra, and voice onset time will also contribute to stop place identification during normal speech perception. Figure 17 presents a summary of some types of representation thought to participate in the transformation of an acoustic input into an identification response, including a contribution by locus equations. Working up from the bottom, the acoustic signal (the word "beet") is shown as a spectrogram; three candidate cues for stop place and their ascending codes are indicated. The stop release burst is circled and shown to be abstracted in a burst feature map or maps; F3 information is separately represented in an F3 feature map or maps; F2 onset and F2 vowel frequencies are shown as inputs to a map computing the feature "vowel-normalized F2 transition" (a locus equation representation). Information-bearing parameters from the speech signal are separately encoded as feature-extracting spectrotopic maps. At the present time, the specific information-bearing parameters from the burst and F3 are not known. Perhaps they would be combined at an early stage with F2 information. The contribution of this article is to suggest that locus equation variables, F2 onset and F2 vowel, are information-bearing parameters from the second formant that contribute significantly to the encoding of stop place of articulation. The next hierarchical level is envisioned as a composite hyperspectrotopic representation of a phonological entity fed by lower maps with feature-specific coding. It is this higher level of phonological encoding that is thought to bind together all the partial and/or redundant cues that combine to allow for a unitary phonemic perception. The *b* identification surface shown at the top of Figure 17 is identical to that shown in Figure 15. It is meant to represent the behavioral level of the identification process.

In summary, there is strong evidence that F2 transition onset and offset, in combination, are major cues for stop consonant place of articulation. These components of the speech signal are likely to be mapped together and extracted as a feature, the vowel-normalized F2 transition, during speech perception.

### *6.2. The orderly output constraint*

It is striking that in the two best known neuroethological models of auditory processing there are shared computational strategies and mechanisms, some of which could be easily adapted to process F2 onset and F2 vowel in combi-



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

nation as information-bearing parameters for consonant place of articulation across vowel context, if those two variables were highly correlated and linearly related as inputs processed in combination appear to be in the animal models. Evolutionarily speaking, language is the "late-comer." If lineages ancestral to ours had already evolved auditory processors (combination-sensitive neurons) and algorithms (2D maps yielding emergent properties) that compute critical features of acoustic input signals using physically inherent linear relationships with little noise, it would make good evolutionary sense for humans to evolve speech signals that the auditory system could map using its old strategies. The question would then become – How do you ensure linear relationships with little noise in the input signal? One obvious solution is to adapt the system that generates these inputs, which are that system's outputs. We suggest that that is exactly what the human vocal tract and articulatory system have evolved to do in producing consonant-vowel sequences. The orderly output constraint hypothesis asserts that the speech production system has adapted to a mapping property of the auditory system by producing a signal with extremely high correlation and linear relationship between two of its most important information-bearing parameters. Elements of the articulatory system are viewed as coevolving with the auditory system – the latter imposing a neurobiological constraint on the former – to produce an output signal that can be reliably and efficiently processed by auditory processors. This line of reasoning is entirely in accord with current thought in auditory neuroethology: "The vocal and auditory systems have evolved together for acoustic communication. In other words, the vocal system has adapted to produce sounds suitable for detection and processing by the auditory system, and the auditory system has evolved to detect and process these sounds" (Suga l988, p. 684).

In coevolved systems it is common for one of the parts to be more constrained in its adaptation. For example, it seems likely that the pigments of certain flowers have adapted specifically to features of bee vision rather than bee vision becoming adapted to a wide range of floral pigments. Similarly, it is plausible that the human speech production system has had to adapt to an auditory system constrained to represent only linear functions with little noise in the 2D, bivariate, category-deriving map domain. We conjecture that the auditory-processing strategy commonly documented in mustached bats and barn owls, and no doubt existing across many other animal species, has been evolutionarily conserved; possibly, in the sense of Stebbins (1974), it is an evolutionarily canalized trait. Humans have inherited this conserved processing strategy, and the high correlation and linear relationships between certain information-bearing parameters important for cuing phonemic categories have been determined by it.

## 7. Computational rationales for orderly outputs – mappable inputs: Self-organizing maps

Are there general computational reasons for the existence of strongly correlated components in speech signals? Assuming that the perceptual system relies upon a mechanism that learns and categorizes speech sounds, there is indeed a powerful reason. Any learning system (even purely statistical) must rely upon correlations between the inputs to identify and organize them into categories. If related inputs

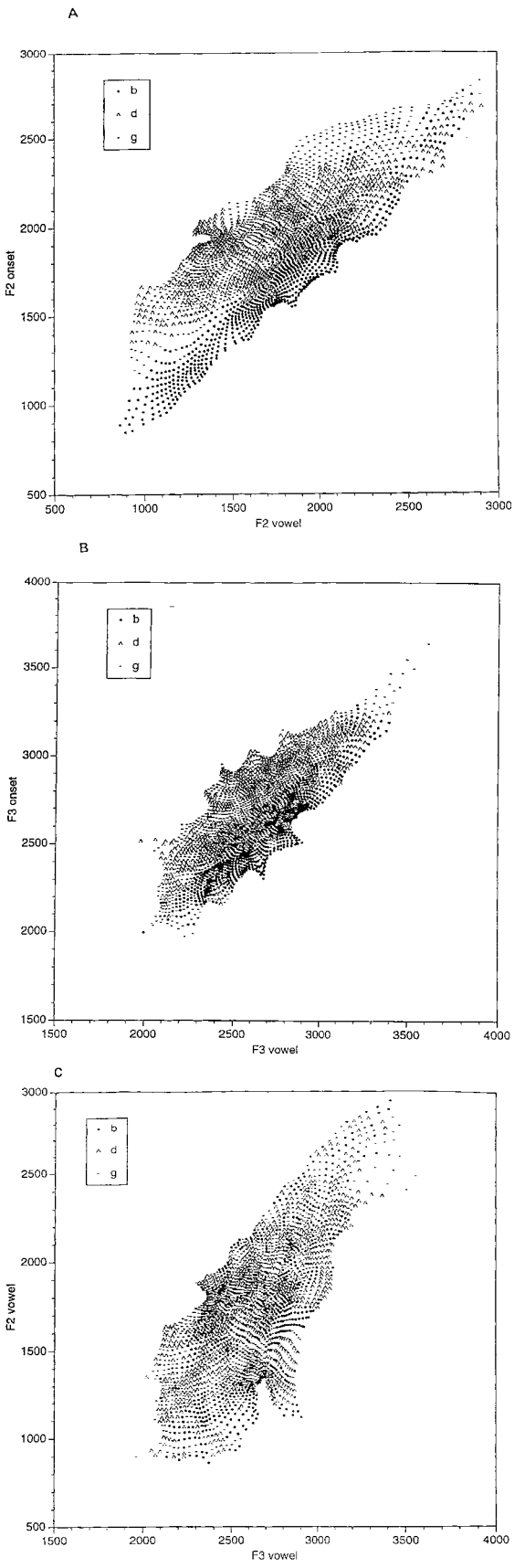



Figure 18. A. Self-organizing map formed by training on F2 onset and F2 vowel frequencies. B. Self-organizing map formed by training on F3 onset and F3 vowel frequencies. C. Self-organizing map formed by training on F2 vowel and F3 vowel frequencies.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

have a common feature that correlates strongly, the learning system can use this feature as the key discriminant to learn and organize categories. Conversely, to establish auditory communication the speech production system must introduce such strong correlations so that the perceptual system can learn and subsequently encode contrastive categorical inputs in an efficient manner. In speech, when a transition between two sounds is the important perceptual cue to be learned and categorized, a simple way to introduce a discriminant feature is to tightly correlate the frequencies at the transition for each cue. It is conceivable that the linear correlations in the F2 transitions are introduced by the vocal system for exactly this purpose.

It is possible to demonstrate computationally the benefits of the F2 correlations to phoneme category formation, using simple neurally plausible algorithms such as the self-organizing map (SOM) algorithm (Kohonen 1982; 1990). The algorithm simulates a two-dimensional network of neurons (as a model of the cortical sheet) and adapts their synaptic weights to represent various features of the input signals. Using only the correlations in the input data, the algorithm orders the synaptic weights of the two-dimensional sheet of neurons so that similar inputs are represented by nearby neurons. Various researchers have shown that such a mechanism can account for the development and structure of topographic maps in the brain, such as somatosensory maps (Obermayer et al. 1991; Ritter 1990) and visuocortical maps (Obermayer et al. 1992).

The self-organizing map algorithm has the property that it maps the "topography" of the input space (defined by correlations) onto the topography of the neural network. The inputs that are strongly correlated will be grouped and represented in clearly defined, contiguous areas of neurons, and clear categories will emerge. To demonstrate this idea we simulated self-organizing maps with the stop consonant transition frequencies as input. The input data for each map were pairs of numbers from actual speech tokens, for example, F2 onset/F2 vowel pairings for one map, F3 onset/F3 vowel pairs for another, and so on. We then displayed the organized maps by plotting the weight vectors of each unit (i.e., the portion of the input space represented by each unit) in input space coordinates (namely, frequencies). The OOC hypothesis predicts that the organization of the F2 onset/F2 vowel map should be superior to the organization of any of the other maps.

Figure 18 shows the self-organized maps that resulted when the following pairs of inputs were used for training: (A) F2 onset/F2 vowel; (B) F3 onset/F3 vowel; (C) F2 vowel/F3 vowel. These three input sets vary in the degree of correlation between the input variables.[9] Comparing the three maps in the figure, it can be seen that the clearest topographic organization, that is, the clearest spatial segregation of the stop consonants, occurred with the inputs F2 onset and F2 vowel (Fig. 18A), the most correlated inputs.

ACKNOWLEDGMENTS

Support for portions of this research was provided by the National Science Foundation (Grant No. BNS-8919221) and the National Institutes of Health (NIDCD R01 DC2014-01A1) to the first author. We would like to thank René Carré for providing the simulations using the DRM model and Jadine Shore for analyzing the DAS speakers. We would also like to extend our appreciation to Peter MacNeilage and Barbara Davis for their generosity in allowing us access to their infant speech recordings. The helpful comments of Björn Lindblom, Randy Diehl, Jeffrey Wenstrup, and several anonymous reviewers are also greatly appreciated. We also gratefully acknowledge the kind words and support provided by Nobuo Suga, whose elegant work with the mustached bat provided much of the theoretical impetus for this manuscript.

NOTES

**1.** By cascading these neurons, a form of "binding" can occur, whereby "multi-combination-sensitive" neurons are created that can be tuned to up to four elements of a complex signal (see Suga et al. 1978).

**2.** We will use quotation marks around the term "category" when discussing the neuroethology data as the term is not, strictly speaking, appropriate for the seemingly continuous and nonquantal nature of velocity or ITD (interaural time difference) functions in the bat and in the barn owl.

**3.** On the ordinate of Figure 2A the second and third harmonics of the CF portion of pulse and echo components were divided by 2 and 3, respectively (CF2/2 and CF3/3). This operation yields the appropriate magnitude of Doppler shift in Hz in relation to CF1.

**4.** Second formant offsets are generally measured in the vicinity of F2 vowel "midpoint"; thus, we will often be using the terms *F2 vowel* and *F2 offset* synonymously.

**5.** A complete locus equation account of stop place of production would necessarily entail systematic analyses of stops in varied syllable positions (initial, medial, and final) and perhaps in consonant clusters. Some of these studies are in progress. However, for our immediate purpose of relating the locus equation phenomenon to analogous neuroethological data, the classic noninvariance problem of accounting for syllable-initial stop place across vowel contexts is quite sufficient.

**6.** There were 17 cases (12.9%) incorrectly classified as velars, and they were all [s] tokens, as locus equation slopes for [s] were significantly higher than found for other alveolars. The F2 transition from the /s/ constriction to the vowel could not be reliably observed during the fricative noise interval as was possible during aspirated [th]. Thus, the F2 onset measurement point was necessarily the first glottal pulse of the vowel following the lengthy noise frication of /s/. This vowel onset frequency was very similar to the F2 vowel midpoint frequency, and hence steeper scatterplots were the spurious result.

**7.** Linearity per se is distinct from high correlation as quadratic and ogive functions can be highly correlated but are obviously nonlinear.

**8.** Modeling of the appropriate degree of consonant-vowel coarticulation as a function of stop place of articulation has not yet been sufficiently accurate to provide a close match to the slopes and y-intercepts of locus equations obtained from real speakers. Nevertheless, it is clear that the human vocal tract will tend to produce correlated F2 transition onsets and offsets, given a consistent place of constriction, simply by virtue of its configuration.

**9.** This can be quantified for each input set in terms of the mean R2 and standard error of estimate (SE) across the three consonants /b, d, g/. The most correlated inputs are the F2 onset and F2 vowel (A), for which the mean R2 is .85 and the mean SE is 133. An intermediate case is F3 onset and F3 vowel (B), for which the mean R2 is .74 and the mean SE is 181. The least correlated inputs, F2 vowel and F3 vowel (C), have a mean R2 of .62 and a mean SE of 294. Thus, these three examples form a series along a correlation of inputs dimension.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

# Open Peer Commentary

*Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.*

## The mapping from acoustic structure to the phonetic categories of speech: The invariance problem

Sheila E. Blumstein

*Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912.* **sheila_blumstein@brown.edu**
**www.cog.brown.edu/people.htm primaryfaculty**

**Abstract:** This commentary focuses on the nature of combinatorial properties for speech and the locus equation. The presence of some overlap in locus equation space suggests that this higher order property may not be strictly invariant and may require other cues or properties for the perception of place of articulation. Moreover, combinatorial analysis in two-dimensional space and the resultant linearity appear to have a "special" status in the development of this theoretical framework. However, place of articulation is only one of many phonetic dimensions in language. It is suggested that a multidimensional space including patterns derived in the frequency, amplitude, and time domains will be needed to characterize the phonetic categories of speech, and that although the derived properties ultimately may not meet the conditions of linearity, they will reflect a higher order acoustic invariance.

The search for invariant acoustic properties that correspond to the phonetic dimensions of speech has been one of the major challenges in speech research. The difficulty has been identifying acoustic properties associated with the phonetic categories of speech that remain constant across the large numbers of sources of variability that occur in speech production. As a consequence, the dominant view in the field of speech research today has rejected the very principles and framework that underlie the work reported by Sussman and colleagues, namely, that there are higher order invariants that can characterize the phonetic dimensions of speech; that these dimensions remain stable across various sources of variability such as speaker, vowel, phonetic class, speaking rate, language, and articulatory perturbations; that these properties are used by the listener in speech perception; and that speech processing is based on more generalized auditory processing principles. Sussman et al. are to be applauded for their efforts, and more importantly, for identifying a higher order combinatorial property related to place of articulation, the phonetic dimension that has provided perhaps the most serious challenge to this point of view in the past. Their findings are consistent with a number of theories of the sound structure of language including the quantal nature of speech (Stevens 1989), a theory of acoustic invariance (Stevens & Blumstein 1981), and the acoustic basis of distinctive (phonetic) features (Jakobson et al. 1963); but importantly, they have provided empirical data and a theoretical framework that intersects the higher order invariance for speech with more generalized principles related to auditory processing and to neuroethological investigations of mammalian and avian communication systems.

Having said this, a number of questions remain related to the combinatorial properties for speech in general and the locus equation specifically. It is troubling that although the locus equation successfully categorizes place of articulation across different vowel contexts, there are regions of overlap in locus equation space as a function of vowel context, with overlap between [d] and [g] in the back vowel environment and [b] and [d] in the front vowel environment (see Figs. 5 and 16). The back vowel [u] and the front vowel [i] are considered particularly critical in delimiting the vowel space of languages, and are proposed to play a critical role not only in the evolution of speech (Lieberman 1975) but also in the perception of speech by infants (Kuhl et al. 1997). These findings suggest then that the locus equation alone cannot be used for the perception of place of articulation. Sussman et al. address this issue in section 6.1 by introducing the notion of a dominance hierarchy, where there is a perceptual preference for [b] in front of front vowels and [d] in front of back vowels. However, they have to resort to other cues in the speech signal to ultimately provide a means for perceiving place of articulation in these contexts. What is not clear is what the nature of these cues may be. Are they context-dependent cues of the type that have been described in the speech literature (Liberman et al. 1967), or are they context-independent properties built from the same general principles used to derive the locus equation (e.g., combinatorial properties and linearity)? How does the listener "weight" these cues? How do they "learn" to weight them? Do the invariant cues based on the locus equation have perceptual prominence?

Consistent with Sussman et al.'s proposal, perceptual investigations have shown that listeners can perceive place of articulation in stop consonants in the absence of the burst. However, they can also perceive place of articulation with just the burst and some 20 msec of transitions (Blumstein & Stevens 1980). In this case, the transitions have not reached the steady state and there is no vowel steady-state present in the stimulus. Moreover, 4- to 5-day-old neonates are perceptually sensitive to these onset characteristics (Bertoncini et al. 1987). Thus, in these situations, listeners *cannot* be using the locus equation in making their perceptual identifications.

Although Sussman et al. focus on the locus equation as an invariant for place of articulation, there have been other proposed invariant acoustic properties for place of articulation (Stevens & Blumstein 1978). These properties are also higher order invariants, integrating spectral properties across the time domain. Can there be *several* invariants for a particular phonetic dimension?

The possibility that the sound structure of language is defined in terms of higher order invariance built from combinatorial properties in two-dimensional space is of great interest and importance. But how important is it that the space be only two-dimensional and not three- or even *n*-dimensional? Place of articulation is only one of *many* phonetic dimensions in language, and other acoustic properties are surely needed to characterize these phonetic categories. Such multidimensional space includes patterns derived in the frequency, amplitude, and time domains, patterns to which the auditory system is most assuredly sensitive. For example, manner of articulation contrasts between stop consonants and glides, nasal consonants and stops, or fricatives and affricates all display a higher order invariance related to the nature of amplitude change in certain frequency bands in the vicinity of the consonant release. Although such properties are combinatorial and display higher order invariants, as is the case with the locus equation, it is not clear that they display linearity. Is this crucial? Why? Would a failure to show linearity render the acoustic invariance captured less relevant or important as a potentially biologically significant emergent property?

ACKNOWLEDGMENT

This work was supported in part by NIH Grant DC00142 to Brown University.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

## Does locus-equation linearity really matter in consonant perception?

Lawrence Brancazio

*Department of Psychology, University of Connecticut, Storrs, CT 06269 and Haskins Laboratories, New Haven, CT 06511.*
**lab93006@uconnvm.uconn.edu**

**Abstract:** This commentary focuses on the claim that perceptual demands have caused the linearity exhibited by locus equations. I discuss results of an experiment demonstrating that, contrary to Sussman et al.'s claims, locus equations do not have relevance for the perception of stop consonants. I therefore argue against the plausibility of the orderly output constraint.

Sussman et al. have outlined an orderly output constraint, according to which the linearity in stop consonant production captured by locus equations is perceptually driven. The authors argue that human perceptual systems capitalize on this linearity in discriminating stop consonants because it facilitates auditory mapping. As supporting evidence, they cite Fruchter's (1994) finding that regions of perceptual dominance for different consonants in second formant (F2) onset-F2 vowel regions overlap with their respective locus equation lines.

This experiment, however, does not, in fact, provide distinct evidence supporting the perceptual significance of locus equations. It has long been established (Liberman et al. 1954) that F2 transitions are used in discriminating /b/, /d/, and /g/; furthermore, the locus equation literature contains ample demonstrations that F2 onset and F2 vowel have a robust linear relationship. Thus, a demonstration that perceptual space tends to overlap with locus equation space only serves to underscore that there is some parity between production and perception with regard to informative portions of the speech signal. Support for the view that "a map of locus equation space somewhere in the auditory system could contribute significantly to consonant place identification" (sect. 6.1, para. 3) would require some demonstration that the linearity itself has some significance for perception. Sussman et al. however, do not provide any quantitative measure of the degree of fit between the locus equation lines and perceptual space. In fact, visual inspection of their Figure 15 reveals that this relationship is quite coarse: the regions of consonant "domination" do, for the most part, cover their respective locus equation lines; however, the topographies of the regions themselves (particularly for /d/ and /g/) could hardly be described as linear.

Fowler and I have recently reported on an experiment (Brancazio & Fowler, in press) that provided a test of the perceptual relevance of locus equations. We presented natural tokens of stop-consonant vowel syllables (/b/, /d/, and /g/ with eight vowels) with their release bursts removed, and had subjects identify the consonant of each. We then devised a model of consonant perception incorporating locus equation space: each token's Euclidean distance to the /b/, /d/, and /g/ lines was computed, and the consonant whose line had the smallest distance was the predicted response. We also devised a model using the same F2 onset-F2 vowel space, but with reference to the coordinates of individual tokens rather than to the locus equation lines computed over them. We were concerned with how accurately subjects would classify the tokens with only transitions available and how well the locus equation-referential model would predict performance compared to the alternative model, indicating the relevance of the linearity for perception. We found that subjects correctly classified the tokens only 66% of the time. This indicates that, modeling aside, F2 (with F3, which was present in the stimuli) was not sufficient for highly accurate identification of the consonants. Furthermore, we found that both models performed very similarly in predicting subject performance, and that they only accounted for modest proportions of the variability in subject classifications. Overall, they correctly predicted approximately 57% of subject responses (correct or incorrect), and distance regression analyses using the Euclidean distances to predict response patterns had $R^2$s of around 0.4. Thus, while there was a significantly greater-than-chance relationship between the performance of the models and the subjects, to a large extent the models were unable to account for the patterns of human responding.

The fact the the locus equation-based model did not outperform the alternative model indicates that locus-equation linearity does not have a bearing on stop consonant identification. Furthermore, given our knowledge of the importance of F2 transitions for perception, the fact that the models left so much variability in identification patterns unexplained (especially when one considers that the bursts, another useful cue, had been removed) suggests that reducing the transitions to two static variables and mapping them together does not capture the way that perceivers actually treat the signal. Together, these points call into question the model of consonant perception outlined in the target article.

In fairness to Sussman et al., they are clear in stating that they believe that the F2 system is only one component of the stop consonant perception system. However, fairly successful models have been devised in which F2 cues are integrally processed with other cues such as F3 and the burst (e.g., Krull 1990), instead of having separate processing systems for F2 and for the other cues (see Fig. 17 of the target article). The latter approach is only necessary insofar as the linearity of F2 transitions has some special significance for perception.

Finally, the question must be raised of why the evolving speech perception system would have imposed the F2-linearity constraint on speech production at all. Consider that locus equation lines only correctly classify approximately 80% of tokens in a discriminant analysis (Sussman et al.), and when subjects have only F2 and F3 to identify consonants, they are correct only 66% of the time. F2 transitions, when viewed from the linearity perspective, simply do not approach the inputs to bat and barn owl auditory system in terms of their perceptual utility.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by NIH grants HD-01994 and DC-20717 to Haskins Laboratories.

## Linear correlates in the speech signal: Consequences of the specific use of an acoustic tube?

René Carré

*Département Signal, Unité Associée au CNRS, 75634 Paris cedex 13, France.* **carre@sig.enst.fr**

**Abstract:** The debate on the origin of the locus equation is circular. In this commentary the locus equation is obtained by way of a theoretical model based on acoustics without recourse to articulatory knowledge or perceptual constraints. The proposed model is driven by criteria of minimum energy and maximum simplicity.

The debate on the origin of the locus equation based on constraints imposed either by a perceptual apparatus (orderly output constraint, as proposed by Sussman et al.) or by a production mechanism is unquestionably circular. What is the origin of what (organ of perception, organ of production)? How did speech develop? On a biological level, it is impossible to escape circularity because the mutual adaptation of the organs of production and reception is definitely permanent. This coevolution, however, may have been driven by the task of human communication, which had to be performed using an acoustic tube. In order to communicate in diverse environmental conditions by exploiting modulated vibrations transmitted by air, humans may have discovered that they could carry out this communication task by deforming, as simply and efficiently as possible, an acoustic tube that was there primarily to help them breathe and feed. These two criteria, simplicity and efficiency (criteria of minimum energy applied to obtain a maximum acoustic contrast), correspond to an appropri-



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

ate adaptation of humans to their environment, exemplified here by the acoustic tube obeying physical laws that are permanent. Such a deductive approach makes the problems of circularity vanish, for the mechanisms of production and perception are mere consequences of an efficient exploitation of the acoustic tube's physical characteristics.

As a first step, when using an acoustic tube for communication, the criteria of simplicity and efficiency allow us to deduce the locus equation. A speech production model (the distinctive region model, or DRM) was obtained by examining acoustic variations of the tube (i.e., formant frequency changes) caused by deformations of the tube's area function, the variation of the cross-sectional area (in $cm^2$) of the vocal tract from the source to the output (in cm). This area is generally between 0 (closure) and 10 $cm^2$. The total length of the tube is around 17 cm. This examination helped us identify regions of the tube that are acoustically the most sensitive to deformations (Mrayati et al. 1988). Any deformation over these regions is consistent with the criterion of efficiency – or minimum energy. We also have been able to note that these regions, defined theoretically (i.e., without any articulatory knowledge), in fact correspond to places of articulation of consonants and vowels in speech production (Carré & Mody 1997).

Furthermore, by superposing a consonant gesture on vowel-to-vowel transition, we were also able to reproduce Ohman's (1966) vowel-consonant-vowel formant patterns (Carré & Chennoukh 1995). In this research, the consonant gesture is strictly in phase with the vowel gesture and thus the degree of co-production (or coarticulation) of the two gestures is maximum. It was using this condition of synchrony that we measured the second formant (F2) onset and F2 offset values for different vowels that were actually discussed in section 5.1 of the target article; the linear relationship between these two measures is, in fact, the locus equation. Hence the locus equation is predicted by our theory starting from a minimum energy criterion paired with co-production in which the consonant and vowel gestures are in phase.

In a further study of the model, we investigated the role of the phase between the consonant and the vowel trajectories by delaying the onset of the vowel gesture with respect to the consonant gesture – in this case, the degree of co-production of the two gestures is also reduced. We observed that the linearity of the F2 onset-F2 offset relationship is preserved and that the slope of the locus equation is correlated with the degree of co-production (Chennoukh et al. 1997). The consonants corresponding to the same place of articulation can be correctly identified by listeners over a wide range of degrees of co-production; instances of incorrect identification occurred mainly for low degrees of co-production. We are thus inclined to conclude that, for a given consonant in different vowel contexts and for a given degree of co-production (same phasing), it is possible to obtain a locus equation with a particular set of parameters. In contrast, when phasing is random, F2 onset-F2 offset data points are no longer on a straight line.

Based on the work just described, we would like to propose that, when the speaker controls the degree of coarticulation, his goal is not to obtain a linear relationship between F2 onset-F2 offset, as Fig. 14 of the target article suggests. Rather, the F2 onset-F2 offset points fall on the same line as a result of the speaker applying a given strategy of co-production to a given consonant (with constant phasing between consonant and vowel gestures). We believe that the objective of the speaker was, during evolution, and still is, during acquisition, to develop the simplest strategy to produce a given consonant, and the simplest strategy consists of using the same phasing between consonant and vowel gestures. Obviously, a strategy resulting in a given phasing will be speaker-dependent. The reader will recognize similarities between these ideas and the uniform coarticulatory resistance hypothesis proposed by Fowler (1994).

Generally speaking, the importance of the locus equation for the organ of perception is debatable. A listener who hears an unknown speaker's consonant-vowel for the first time has no difficulty identifying the consonant. It is thus unnecessary to have a prior knowledge of the speaker's locus equation. Why not take into account the totality of the transition that contains maximum information instead of only two arbitrarily selected, discrete points? The perceptual mechanism must be able to grasp the whole dynamic of information and the consonant gesture must be perceived as such.

ACKNOWLEDGMENT

The author thanks Pierre Divenyi for stimulating discussions and comments.

## Self-learning and self-organization as tools for speech research

R. I. Damper

*Cognitive Sciences Centre, and Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, England.*
**rid@ecs.soton.ac.uk isis.ecs.soton.ac.uk/**

**Abstract:** Locus equations offer promise for an understanding of at least some aspects of perceptual invariance in speech, but they were discovered almost fortuitously. With the present availability of powerful machine learning algorithms, ignorance-based automatic discovery procedures are starting to supplant knowledge-based scientific inquiry. Principles of self-learning and self-organization are powerful tools for speech research but remain somewhat under-utilized.

Locus equations were first discovered by Lindblom (1963a) but have since been more thoroughly investigated by Sussman and colleagues. They offer as much promise for understanding the vexed question of invariance, whereby speech sounds are physically modified by their context but are still perceived as members of the same equivalence class (phoneme category), as any proposal yet advanced. The basic notion is that, while features of the speech signal may vary as a result of coarticulation, the relation between certain key features may exhibit a consistent and lawful (invariant) form.

In the target article, in addition to reviewing the utility of locus equations, Sussman et al. argue for their neurobiological plausibility based on the potential to build relational, higher-order feature detectors (and thereby category detectors) from the combination-sensitive neurons found in a variety of mammalian and avian auditory systems. The essential argument is that "there is no reason to suspect novel processing strategies or neuron types to have arisen for basic auditory encoding of the acoustic cues signaling feature contrasts in human speech" (sect. 1.2). This is at variance with the early "speech is special" hypothesis, which still has its adherents (e.g., Liberman 1996; Liberman & Mattingly 1989). In my view, the consensus of informed opinion is now firmly on the side of Sussman et al.: human speech perceptual mechanisms are thought to be based on general auditory processing principles, common to a range of species, with specialization occurring only at a relatively high level (See also sect. 1.3.2., para. 2, the personal communication from Suga.) Indeed, using a computational modelling approach, we have recently shown (Damper et al., submitted) that the placement of phoneme category boundary in human and animal listeners between initial stops (/b/, /p/; /d/, /t/; /g/, /k/) distinguished by their voice onset time can be replicated by a trivially simple neural processing scheme that needs only to integrate activity over certain time-frequency regions of auditory nerve activity. This aspect of speech perception has attracted enormous attention over decades, yet can be simply explained.

The traditional approach to the speech invariance problem can be characterized as "manual search." That is, using knowledge and ingenuity, the experimenter tries to generate some hypothesis about possible invariant features, which is then tested for consistency with available data. Lindblom's discovery of locus equations is very much in this vein. The approach is inherently



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

unsatisfactory, however, both because knowledge and ingenuity are always in short supply and because the consistency check with the data is a posteriori. With the advent of connectionism between 10 and 15 years ago, and the greatly-increased availability of powerful learning algorithms such as error back-propagation (Rumelhart et al. 1986), it is becoming ever more common to employ data-driven rather than knowledge-driven strategies in virtually all areas of scientific inquiry.

Such automatic search, which exploits the self-learning and self-organizing capabilities of neural networks, ensures that (provided training is successful, and the network convergences onto the desired behavior) only hypotheses consistent with the training data are (implicitly) generated. Furthermore, the search is guided by a general optimization principle (i.e., it is "ignorance-based"). Together, these two considerations mean that features in the widest sense, which are important to categorization (but may not be obvious; e.g., because they are relational), naturally emerge as determiners of network behavior. Of course, it remains to uncover – by an appropriate analysis of the trained network(s) – the implicit hypotheses that have been automatically discovered. Contrary to the wide-spread belief that neural networks are "black boxes" whose operating principles cannot be sensibly determined, techniques for doing just this are improving all the time.

Not only was this approach adopted by Damper et al. (see above), but Sussman and colleagues also mention several neural network studies (e.g., Waibel et al. 1989) that – after analysis "to determine which parts of the input were most effective for the task" (sect. 6.1, para. 1) – confirm the importance of second formant (F2) onset and F2 vowel to the identification of stop-consonant place of articulation. Sussman et al. employ self-organizing Kohonen maps (Kohonen 1990) to confirm the clear emergence of topologically-ordered regions reflecting the consonant categories on the F2 onset/F2 vowel map but not, for instance, on the F2 vowel/F3 vowel map. Of course, the kind of competitive-learning principle embodied in Kohonen-style self-organization could as well have been employed to discover this relation a priori rather than merely to confirm it a posteriori.

In conclusion, Sussman et al. present a clear and convincing case for the emergence of higher-order features (exploiting general neural processing, rather than speech-specific, principles) as the basis of at least some of the category invariance observed in speech perception. There is great scope, however, for such features and principles to be discovered automatically in the future. Self-learning and self-organizing systems offer a valuable and currently under-used tool for speech research.

## Locus equation and hidden parameters of speech

Li Deng

*Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.* **deng@crg6.uwaterloo.ca**

**Abstract:** Locus equations contain an economical set of hidden (i.e., not directly observable in the data) parameters of speech that provide an elegant way of characterizing the ubiquitous context-dependent behaviors exhibited in speech acoustics. These hidden parameters can be effectively exploited to constrain the huge set of context-dependent speech model parameters currently in use in modern, mainstream speech recognition technology.

Sussman et al.'s target article successfully synthesizes a good deal of previously published work and presents a comprehensive set of data demonstrating the consistency of locus equations across diverse speakers, languages, and perturbation conditions. The main purpose of this commentary is to show that the regularity of or relational invariance contained in the speech pattern as exhibited by the locus equations can be exploited to effectively constrain the structure of statistical models of speech for speech recognition applications.

***A statistical model of speech constrained by locus equations.*** I will describe a statistical model that utilizes the locus equations as a basis for parametric modeling of phonetic contexts. The model, called Locus-HMM, is based on hidden Markov model (HMM) representation of formant-transition microsegments of speech. Automatic estimation of the model parameters, which include the slope and intercept parameters in the locus equations, can be accomplished via statistical optimization techniques. The model is capable of generalizing consonant characteristics from a small training set in which the contextual information is only sparsely represented, and is hence applicable to large vocabulary speech recognition problems that would traditionally require exhaustively enumerating all possible contextual factors with no or at best heuristically derived constraints on a large set of model parameters.

The locus equation describes a linear relationship between the onset frequencies of the second formant (F2) transitions and the corresponding midvowel frequencies:

$$m_{onset} = k_c * m_{midvowel} + b_c, \quad (1)$$

where $m_{onset}$ and $m_{midvowel}$ are the F2 values (or other acoustic parameters related to F2 such as spectral centers of gravity within appropriate frequency bounds) measured at onset and at steady state in a consonant-vowel (CV) syllable; $k_c$ and $b_c$ are slope and intercept of the locus equation, which is considered as an "invariant" property for a constant, independent of the vowel context. The parameters $k_c$ and $b_c$, one pair for each consonant, control the degree of contextual dependence but are not directly observable in the acoustic data. In this sense, these parameters are said to be hidden, and can be inferred only by analysis (manually, as described in the target article, or automatically by computer algorithms discussed here) of the acoustic data over a time span in the order of one-syllable length.

We here consider use of a $Q$-state Gaussian HMM, constrained by the locus equations, to represent a formant-transition microsegment in a CV environment. In this Locus-HMM, the means associated with various HMM states are not independent of each other. Rather, the locus equation (1) and the fact that within a CV syllable F2 transition is monotonic impose constraints among the Gaussian mean parameters ($m$'s in equation 1) in the model. Incorporating these constraints on the otherwise conventional HMM, powerful maximum-likelihood based statistical techniques can be effectively used to automatically estimate all the conventional HMM parameters and the locus equation parameters $k_c$ and $b_c$ (see details in equation 1).

The reason the constraints provided by the locus equations are important is that the context-dependent behavior (in the CV context discussed here, but which can be generalized to other contexts easily equation 2) of speech can be succinctly parametrized by vowel-independent, consonant-specific parameters $k_c$ and $b_c$. This eliminates the need to model the context dependence in a traditionally nonparametric manner that creates numerous practical difficulties in speech recognition (especially where rare adaptation data are available to tune model parameters).

***Speech recognition using Locus-HMM.*** Two separate attempts were made to use the Locus-HMM to improve the current speech recognition technology, one in the task of large vocabulary word recognition (2) and the other in the task of phonetic classification defined in the timit database (3). Using a number of engineering considerations and implementation techniques, up to 15% error rate reduction was achieved in comparison with the state-of-the-art speech recognition methods under identical training and testing conditions.

***Locus equations measured from fluent speech utterances.*** In (3), an attempt was also made to examine the validity of the locus equation using fluent speech utterances from timit data. Although the general trend of linearity holds, the degree of linearity is significantly less than that described in the target article. This may be correlated with the limited recognition-performance improvement (at most 15%) despite the substantial engineering efforts made.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

Caution is advised in interpreting the strict linearity imposed by the locus equation as a universal brain mechanism related to "evolutionarily conserved auditory processing strategy." For one thing, the lower degree of linearity found in fluent speech data (Figs. 2–4 of equation 3) compared with that reported in the target article (Figs. 4–8) appears to be easier to account for by the production-oriented interpretation of the locus equations. It would be interesting to examine whether a vocal tract simulation similar to the one described in section 5.1 of the target article would show systematic disparity in the degree of linearity for read-style speech (with little or no formant undershoot) and casual/fluent speech (with strong formant undershoots). Furthermore, if the formant-target undershoot model and the locus-equation model can be shown to have the same origin in production-oriented strategies by the speaker, then one may not need to invoke the auditory strategies to account for the fairly straightforward speech acoustic phenomenon illustrated by the locus equations.

## Locus equations: A partial solution to the problem of consonant place perception

Randy L. Diehl

*Department of Psychology, University of Texas, Austin, TX 78712.*
**diehl@psy.utexas.edu www.psy.utexas.edu**

**Abstract:** In their important work on locus equations, Sussman and his colleagues have helped to simplify the theoretical problem of how human listeners identify place of articulation contrasts among consonants, but much work remains before this problem is solved.

Sussman and his coauthors have described a truly impressive body of work aimed at characterizing important regularities (viz., the "locus equations") in the production and perception of place of articulation contrasts in spoken language. As the authors correctly point out, the place dimension has traditionally been viewed as a critical test case for evaluating theories of speech perception. Because certain place cues are highly context-dependent, some investigators have claimed that the perception of place categories requires reference to underlying motor events, which are assumed to be more nearly invariant than the acoustic consequences of those events (Liberman et al. 1967). The authors have demonstrated that there are acoustic correlates of place that are highly regular and thus potentially mappable onto stable neural representations without the benefit of motor reference. If the seemingly intractable place dimension can be handled in so straightforward a manner, we surely have grounds to be more optimistic about the prospects of a general theory of speech perception.

As the authors readily acknowledge, locus equations alone do not provide a sufficient basis for identifying a consonant's place category. One reason is that the locus equations for /b/, /d/, and /g/ intersect in nonempty regions of the second formant (F2) onset/F2 vowel space, resulting in ambiguity of place category membership for some consonant tokens. In particular, the equations for /b/ and /d/ intersect in the front vowel region, whereas those for /d/ and /g/ intersect in the back vowel region. This means that acoustic correlates of place in addition to F2 onset and F2 vowel are required by listeners in order to identify place reliably. The authors suggest quite reasonably that correlates such as the burst (attributable to transient excitation of the vocal tract upon release of the articulators) may serve this disambiguating role.

A second reason why locus equations do not yield a sufficient basis for place perception is that each equation is an aggregate description of a consonant category (e.g., the category /b/ across all vowel contexts). The equation parameters of slope and y-intercept are clearly not recoverable from any single consonant token (although, by hypothesis, these parameters are part of the permanent neural representation of the category to which individual consonant tokens must be referred). Parameters that presumably *are* recovered during "on-line" perception include F2 onset and F2 vowel. However, as Sussman et al. show, discriminant analyses based on the latter parameters yield only partial separation among place categories. Again, one is led to conclude that other correlates (in addition to F2 onset and F2 vowel) must play a significant role in place perception.

All of this is explicitly noted by Sussman et al. However, in light of these considerations, it is reasonable to ask whether the neuroethological examples cited in the target article are as closely analogous to the case of human speech perception as the authors suggest. The linear functions displayed in Figure 2 of the article ("isovelocity contours" in the mustached bat and "iso-interaural-time-difference" contours in the barn owl) differ from locus equations in two respects. First, they do not intersect anywhere in the effective stimulus space. Second, the data points are more tightly clustered about the linear contours than in the case of locus equations. Thus, in the neuroethological examples there appears to be no potential ambiguity in the mapping between stimulus categories and neural representations. That is, the combination-sensitive neurons described are alone sufficient to identify the stimulus category or value.

In the case of human speech perception, the neural representation that, by hypothesis, corresponds to a locus equation must be supplemented by an indeterminate number of additional neural representations (e.g., the burst characteristics and F3) in order to yield an unambiguous identification of the place category. These various neural correlates of place presumably must be weighted and combined in forming a judgment, and it appears likely that the weights will vary according to phonetic context and other factors. (For example, neural correlates of the burst might be given more weight in just those regions where locus equations intersect.) In other words, beyond the mapping of stimulus parameters onto the neural analogues of locus equations, a good deal of computational work must be performed in order to complete the perceptual task. (For an elaboration of this general point, see Diehl 1981 and Diehl & Kluender 1987). Moreover, there is no guarantee that any of the acoustic correlates of place besides F2 onset and F2 vowel will turn out to satisfy some version of the orderly output constraint.

These comments do not in any way undermine the main thrust of Sussman et al.'s argument. They are intended only as a gentle reminder that much work remains to be done before we have a fully adequate account of how human listeners identify consonant place.

ACKNOWLEDGMENT

Preparation of this commentary was supported by research grant No. 5 R01 DC00427-09 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

## Differences that make a difference: Do locus equations result from physical principles characterizing all mammalian vocal tracts?

W. Tecumseh Fitch[a] and Marc D. Hauser[b]

*[a]Program in Speech and Hearing Sciences, Harvard/MIT; [b]Department of Anthropology, Harvard University, Cambridge, MA 02138.*
**tec@wjh.harvard.edu hauser@wjh.harvard.edu**

**Abstract:** Sussman and colleagues provide no evidence supporting their claim that the human vocal production system is specialized to produce locus equations with high correlations and linearity. We propose the alternative null hypothesis that these features result from physical and physiological factors common to all mammalian vocal tracts and we recommend caution in assuming that human speech production mechanisms are unique.

We are sympathetic to many of Sussman et al.'s arguments, especially the claim that the auditory systems of humans and other



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

animals are similar and that human-specific perceptual mechanisms are likely to be evolutionarily derived. Thus, we agree that careful use of the comparative method, even when applied to animals as different from us as birds and bats, can and will fuel significant advances in our understanding of speech perception. Therefore, it is unfortunate that neither the empirical observations nor the theoretical arguments that the authors present provide significant support for these claims.

There are many problems with the locus equation story. For example, a given utterance contains information for just one point on a locus equation plot, not a line, and so provides little information by itself. To construct a locus plot for a given consonant, the listener must have already classified a number of syllables correctly, which requires the identification problem to be solved already. Furthermore, because the method to create a locus plot requires already-classified data, 100% correct classification of locus data is unimpressive. Similarly, any smooth, continuous function will yield a strong correlation between closely neighboring sample points; and as the distance between them decreases toward zero, the correlation will become perfect and perfectly linear, with a slope of unity. Because this is true for *any* smooth function, it is not surprising that locus plots of formant functions yield high correlations and linearity.

More interesting is the suggestion that, because humans rely on information in second formant (F2) transitions to categorize certain speech sounds (Liberman et al. 1954), the human articulatory system has evolved to produce such patterns. The elegant work of Ryan and colleagues (1990) on sensory exploitation in frogs provides a good indication of how valuable the comparative method can be in understanding production/perception coevolution in a communicative context. In humans, the best example of such coevolution is the hypothesis of Lieberman et al. (1969) explaining the unique position of the human larynx. Having the larynx further down in the throat than other mammals gives us a unique "two-tube" vocal tract, which allows us to produce a wider range of the formant patterns to which our auditory system is so sensitive. The discovery of a new adaptation of the human vocal system, co-evolved to a putative speech perception mechanism, would indeed be exciting. We will accordingly focus our critique on Sussman et al.'s new proposal.

Sussman et al. propose that "the articulatory system, across diverse articulators (tongue, lips, jaw, velum), adjusts consonant-vowel coarticulation . . . in order to fine-tune a feature of that output, the F2 onset/F2 vowel ratio" (sect. 5.3), and that this is a "coevolutionary adaptation of the human speech production system" (Introduction). This is somewhat puzzling, since Carré's speech modeling system presumably does not include these special co-evolved adaptations (being based on the constraints of human vocal anatomy and straightforward linear acoustics of tubes), but nonetheless reproduces the plots so exactly. Second, the poor data from the children diagnosed with developmental apraxia of speech (DAS) seem odd, because these children surely have human vocal tracts. Finally, the babbling data are more puzzling, since infants under 4 months *do not* have the adult human vocal tract configuration, but instead one more like that of other mammals (Lieberman 1984). We are left wondering precisely what this special adaptation of the human speech production system is: the computer data militate against any specially developed motor control circuitry, while the DAS and baby data argue against anything specific about human vocal anatomy.

A plausible null hypothesis is that the F2 patterns observed in both the computer speech simulations and in real data result from basic acoustic and physiological principles that hold for *any* mammalian vocal tract. If a single articulator (e.g., the tongue) tries to accomplish two goals in rapid succession (e.g., produce a vowel at one location just after producing an occlusion at another) the stiffness and inertia of this articulator will ensure an influence of the two goals on one another. Strong interference should drive locus slopes to be less than one (as in /g/ or /d/). In contrast, if another independent articulator (e.g., the lips in /b/) is brought into play, the tongue can achieve its goal more directly and dominate the F2 contour (giving the expected unity locus slope and perfect correlation of F2 onset and F2 vowel). Sussman et al. argue that the bite block data provide evidence that "the articulatory system adjusts its output in order to preserve the relationship" between F2 onset and F2 vowel (sect. 6). However, if no active control is necessary to achieve this relationship under normal conditions, no "adjustments" are necessary with the bite block in place.

Sussman et al. give no indication that the human tongue, velum, lips, or jaw differ from those of other mammals in any manner germane to these issues, and recent data (Fitch 1997; Hauser & Schön-Ybarra 1994; Hauser et al. 1993) reveal important similarities in the vocal production systems of humans and, at least, other primates. Thus, we see no reason to accept their conclusion that the locus data indicate a uniquely human co-evolved feature of the speech production system. None of the data or arguments they put forth demonstrate or even persuade that anything specifically human is required. Future work would profit from more direct comparisons with primate vocalizations and communication systems, which have much more in common with human speech, both functionally and physically, than the neural systems underlying barn owl prey detection or bat echolocation (Hauser 1996).

Because past stages in evolutionary history are not typically preserved, the comparative method provides us with one of the most valuable tools in understanding evolution. Its responsible use requires a detailed knowledge of the similarities and differences between the species under study. When it comes to humans, we are often too easily lulled into thinking of ourselves as special and unique, despite the fact that much of modern biology is a testament to the basic biochemical and evolutionary unity of life on earth. In order to understand (and appreciate) the human differences that really make a difference, we need to explore and understand the similarities as well.

## The orderly output constraint is not wearing any clothes

Carol A. Fowler

*Haskins Laboratories, New Haven, CT 06511; Department of Psychology, University of Connecticut, Storrs, CT 06269; Yale University, New Haven, CT 06520.* **fowler@haskins.yale.edu**

**Abstract:** The orderly output constraint (OOC) is extraneous. Talkers "speak in lines" in its absence. Further, there is no perceptual motivation for an OOC; perceivers ignore the linearity between F2 at consonant-vowel onset and F2 in the vowel. In any case, the analogy with bat and barn owl localization systems underlying the theory *is* extreme, Sussman et al.'s comments to the contrary notwithstanding.

I have proposed (Fowler 1994) that the linear relation between second formant (F2) onset and F2 vowel and the different line slopes for different consonants reflect characteristic resistances of consonants to coarticulatory overlap by vowels. Researchers (e.g., Farnetani 1990; Recasens 1984; 1989) have shown that consonants resist coarticulation by vowels to the extent that the vowels interfere with achieving consonantal gestural goals. For example, labial consonants generally have lower coarticulation resistances than lingual consonants, and their locus equations generally have higher slopes. I have suggested that the relation between F2 onset and F2 vowel is linear for a given consonant produced in the context of different vowels because coarticulation resistance is largely invariant for a consonant in the context of different vowels. (Vowels all use the tongue body, so their interference with a given consonant should be approximately the same.) Thus, there is a purely gestural reason why F2 onset and F2 vowel are linearly related, and the linear relation need not have any perceptual relevance.

In their target article, Sussman et al. offer two disconfirmations of these ideas. The ostensible empirical disconfirmation is evi-



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

dence that vowels coarticulate to different extents with consonants (Amerman 1970). However, Brancazio and I (1998) have found this "disconfirming" evidence to support, not challenge, our claim. My "deductive failure" was in incorrectly assuming a linear relation of articulation to acoustics. If coarticulation resistance is nearly invariant for a given consonant produced with a variety of vowels, the acoustic consequences of their coproductions should yield a nonlinear relation between F2 onset and F2 vowel. Of course, the validity of this objection depends on the magnitudes of the relevant nonlinearities. These can be estimated from Figure 13 of the target article, which depicts relations between F2 onset and F2 vowel generated by the distinctive regions model (e.g., Chennoukh et al. 1997). When Chennoukh et al. generated locus equation data from the model, for a given simulation, they held constant the extent of coarticulation between consonant and vowel. I assume that Sussman et al. did, too; their Figure 14 implies that they have no principled way to vary it. If so, the departures from linearity in the figure are those due to the nonlinear relation of articulation to acoustics. They are small and of a magnitude characteristic of human data.

My proposal has two advantages over that of an orderly output constraint. It invokes a constraint on production for which there is independent evidence and motivation, and it explains why the slope magnitudes are as they are.

As for the perceptual import of the linear relation of F2 onset and F2 vowel: if linearity reflects requirements to meet gestural goals, it should be perceptually irrelevant. Available evidence confirms this expectation. To my knowledge, Sussman's laboratory has produced just one perceptual study (Fruchter 1994) ostensibly related to the theory, which is described in the target article. This study strongly supports the viability of language as a communication system in showing that listeners tend to perceive what talkers say, but it does not test a distinctive prediction of locus-equation theory. Recently, Brancazio and I (1998) have tested and disconfirmed a claim that the linear relation between F2 onset and F2 vowel has perceptual relevance.

As for the analogy between human speech perception and bat and barn owl localization, Sussman et al. claim that it is not "extreme" considered at the proper level of abstraction. I disagree. What are the relevant similarities? Bats receive signals that have frequency-modulated parts and steady-state parts, as do humans when they hear (carefully articulated, slow rate) consonant-vowels (CVs). However, that cannot be relevant, because barn owls do not receive or require such signals, and humans do not require them. In addition, certain variables in stimulation to bats, others to barn owls, and still others to humans are linearly related. Is this significant? Not likely. First, humans do not receive lines in immediate stimulus input in the way that bats and barn owls do. A CV provides a point, not a line (and, if, in a constrained setting, a point is sufficient to specify a line, the line is redundant). Just as important, I am not aware of any evidence that linear relations between stimulus variables are distinctively perceptually informative. Physical law renders certain linear (and certain nonlinear) relations between stimulus variables informative, and bats and barn owls use some of the linear information. However, they use the information not because it is linear, but *because it is available and informative about relevant properties of environmental events.* That is the proper level and (functional) kind of abstraction relevant to comparisons among perceptual systems.

If we set aside the failed analogy and acknowledge that the linearity in speech acoustics is perceptually irrelevant, what is left? Left is the more than 40-year-old finding by investigators at Haskins Laboratories (e.g., Liberman 1996) that F2 transitions provide important information for consonant identification.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by NICHD grant HD-01994 and NIH grant DC-02717 to Haskins Laboratories.

## Listeners' perceptual mapping of locus equations and variability

Krishna Govindarajan

*Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.*
**krishna@speech.mit.edu**

**Abstract:** Although an individual speaker's productions obey locus equations, whether listeners' perceptions are based on them needs further exploration. Comparing the results from the perceptual experiments to predicted identifications, one sees qualitative similarities and some discrepancies. However, the variability of locus equations and individual consonant-vowel (CV) tokens across speakers seems problematic if listeners are using locus equations for perception.

Sussman et al. have shown that an individual speaker's productions tend to adhere to the locus equations, but the question of whether listeners use locus equations for perception needs further exploration. If one compares the results from the perceptual experiment (Fig. 15) to the expected mapping that would arise if listeners used locus equations, one sees qualitative similarities and some discrepancies. The main problem with the idea of listeners' use of locus equations, however, is the variability across speakers of both individual tokens and locus equations.

***Perceptual mapping of locus equations.*** Sussman et al. show that the "category-level variables," slope and intercept of the locus equations, can differentiate stop consonant place of articulation. However, as Fowler (1994) and the target article correctly point out, it is impossible to determine the slope or intercept from a single point – a given CV corresponds to only a single point P in the second formant (F2) vowel-F2 onset plane. Thus, in order for locus equations to be used by listeners, they would have to categorize the consonant based on which locus line was closest to point P. For this to occur, listeners must have internal, averaged locus lines for /b/, /d/, and /g/, and speakers must produce their locus line (and the CV tokens that fall along that line) closest to the mean locus equation for the intended consonant.

Assuming that listeners are categorizing each point P based on the closest locus line, one can predict the resulting identification surface. Figure 1 shows the mapping one obtains using the averaged locus lines defined in Sussman et al. (1991)[1] and a Euclidean distance metric. The mapping is qualitatively similar to

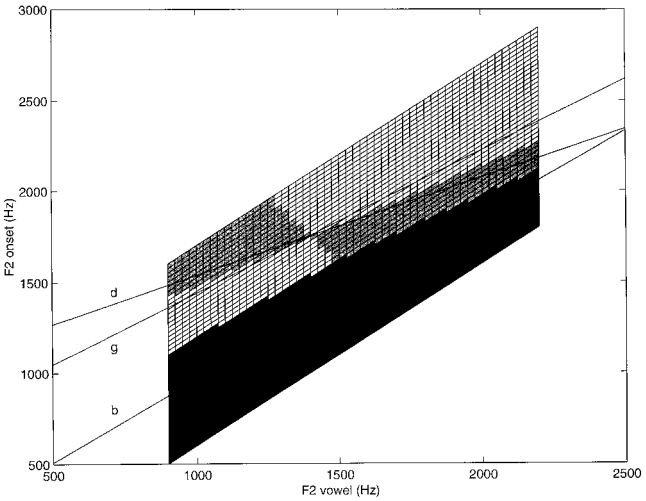



Figure 1 (Govindarajan). Predicted identification regions based on minimum Euclidean distance to the locus lines, and the locus lines based on all speakers of Sussman et al. (1991). /b/ corresponds to the black region, /d/ corresponds to the dark gray region, and /g/ corresponds to the light gray region.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

the results of the perceptual experiment (Fig. 15). Moreover, the resultant mapping shows that there is no need for the "dominance hierarchy hypothesis" (sect. 6.1) – points with a high F2 vowel and high F2 onset are closest to the /g/ line, hence /g/ dominates in front vowel contexts; similarly, /d/ dominates in back vowel contexts. Although the identification surfaces are similar, this does not necessarily validate the theory that listeners use locus equations for perception. There are some points that are incongruous with the theory. For example, one should expect the boundaries between /b/, /d/, and /g/ to correspond to the bisectors of the locus lines. However, in the back vowel context, the boundary between /b/ and /g/ occurs next to the /b/ locus line instead of the midpoint between the locus lines. Moreover, in the front vowel context, part of the /d/ region lies on top of the locus line for /b/.

In addition, the results from the neural modeling, shown in Figure 18, do not provide explicit evidence for locus equations. Instead, they show the primacy of the F2 transition. The tighter clustering one sees in Figure 18a versus 18b or 18c emphasizes the fact that the F3 transition and the information in the steady-state vowel are not as crucial as the F2 transition in determining the identity of the consonant.

***Variability of CV tokens and locus equations across speakers.*** For a listener to categorize consonants consistently, speakers should try to reduce the overlap of CV tokens in the F2 vowel/F2 onset plane by matching their locus lines to the mean locus lines across speakers. However, as Fowler (1994) has shown, the overlap of CV tokens produced by different speakers is large (Fowler's Fig. 2). Moreover, the locus line for a given consonant can vary dramatically across different speakers (e.g., Figs. 1 and 2 of Sussman et al. 1995).

This overlap is brought out further in the slope and intercept plots of Figure 6 in the target article and Figure 3 of Sussman and Shore (1996). Although the slope and intercept lead to perfect classification of /b/, /d/, and /g/, these figures also show that different speakers use different slopes and intercepts for the same place of articulation. Translating these slope and intercept points to the F2 vowel-F2 onset plane, the plane where perception occurs, one does not see segregation of the locus equations, but overlap. Figure 2 shows the locus lines for /b/, /d/, and /g/ for the 20 speakers shown in Figure 6 of Sussman et al. target article. Note that there is a large overlap across speakers' locus equations, especially for /d/ and /g/. Another example of the overlap derives from the large range of slope and intercept values for alveolars in Sussman and Shore (1996), where the slope and intercept values range from 0.1 and 1800 Hz to 0.7 and 450 Hz, respectively. The result in the F2 vowel-F2 onset plane is to produce a series of alveolar locus lines that look like spokes on a wheel, covering the majority of the F2 vowel-F2 onset plane.

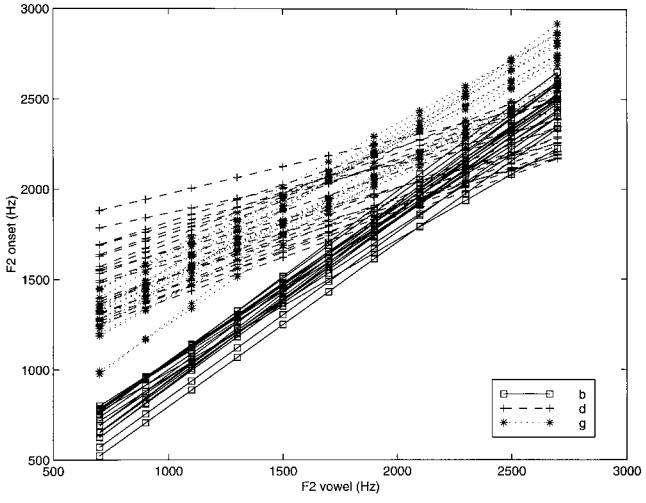



Figure 2 (Govindarajan). Locus lines for /b/, /d/, and /g/ for the 20 speakers in Sussman et al. (1991).

If speakers truly want their utterances to be perceived correctly, then one would expect little overlap of the locus equations across speakers. Thus, while speakers produce CV utterances that fall along the locus equations, the theory that listeners are using locus equations for perception seems undemonstrated.

NOTE

**1.** The locus equations in Figure 15 differ from the locus equation used in Figure 1. The locus equations in Figure 15 were derived from five of the ten male speakers in Sussman et al. (1991), whereas the locus equation in Figure 1 uses all the male and female speakers.

## In search of the unicorn: Where is the invariance in speech?

Steven Greenberg

*International Computer Science Institute, Berkeley, CA 94704.*
**steveng@icsi.berkeley.edu**

**Abstract:** Understanding spoken language involves far more than decoding a linear sequence of phonetic elements. In view of the inherent variability of the acoustic signal in spontaneous speech, it is not entirely clear that the sort of representation derived from locus equations is sufficient to account for the robustness of spoken language understanding under real-world conditions. An alternative representation, based on the low-frequency modulation spectrum, provides a more plausible neural foundation for spoken language processing.

Classical models of speech perception presume that the essence of meaning can be distilled from a linear (or quasilinear) sequence of linguistic elements. At the acoustic level these elements are most commonly associated with phonetic segments (or "phones"), through whose sequential association larger, more abstract units such as the syllable, word, and phrase are derived. In this traditional view, the phone functions as the minimal linguistic unit capable of distinguishing among lexical entities. In turn, each phone is composed of distinctive (articulatory or acoustic) features that, when bound together, yield a specific phonetic element. Within this framework each phone is commissioned to play a specific and important role in the systematic conversion of sound into meaning. Any misstep along the way potentially jeopardizes the speech decoding process, and hence it is crucial for each phonetic segment to be accurately and faithfully represented.

The locus equations so elegantly derived by Sussman and colleagues in their target article provide a neat, compact means of deriving the requisite invariant representations from the underlying acoustic signal within this traditional theoretical framework. Unfortunately, it is not entirely clear that speech understanding necessarily entails such a linear decoding process or that there are neuronal mechanisms capable of extracting the feature patterns required to functionally simulate the representational equivalence effected by locus equations.

Detailed phonetic transcription of spontaneous spoken English (4 hours of informal, conversational dialogues systematically sampled from the switchboard corpus; Godfrey et al. 1992) indicate that it is often difficult to associate much of the acoustic signal with specific phonetic symbols (Greenberg et al. 1996). Phone elements are frequently deleted or significantly transformed during the process of spoken discourse, so that words are rarely characterizable as a linear sequence of phonetic elements. Even trained phoneticians frequently have difficulty identifying a significant proportions of speech sounds contained in the switchboard corpus. However, with few exceptions, these conversations are perfectly understandable. Furthermore, the phonetic variability occasioned by dialectal, idiolectal, and entropic factors is enor-



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

mous. Many of the most common words are phonetically realized in dozens of different ways (Greenberg 1997). Often, the most reliable cues to phonetic identity are temporal, rather than spectral, in nature (Greenberg 1997; Greenberg et al. 1996).

In addition to these speaker and linguistic sources of phonetic variability, environmental factors such as reverberation and background acoustic interference cause a significant alteration of the spectrotemporal properties of the speech signal reaching the listener's ears (Greenberg & Shire 1997; Kingsbury et al. 1997). Thus, it is not entirely clear what sort of "invariance" should be sought in the signal given the nature of acoustic–phonetic variability commonly found in informal, spontaneous speech.

Yet it is tempting to search for some form of invariant representation given the robustness of speech under such a wide range of environmental and speaker conditions. Some property (or combination of properties) of the speech signal must be responsible for the hardiness of spoken communication. Locus equations, to the extent that they are associated with specific formant trajectories in the signal, are unlikely to yield the sort of invariant representation required to account for the intelligibility of speech in the real world, because they require a relatively faithful transduction of the acoustic signal in the auditory pathway. Unfortunately, auditory neurons are unlikely to provide sufficient precision of coding (at least at the level of the auditory cortex; see Schreiner's commentary, this issue) to accommodate the sort of neuronal processing implied by locus equations (at least in mammalian species other than bats).

A more likely means of providing a quasi-invariant representation of the speech signal is through neural computation of the low-frequency (<25 Hz) modulation spectrum. The magnitude of the modulation spectrum at any given frequency is derived from the modulation pattern of the speech waveform over a predefined bandwidth (typically $\frac{1}{4}$- to 1-octave wide). Preservation of this modulation information, distributed across frequency channels, is sufficient to encode natural sounding, intelligible speech (Dudley 1939). The modulation transfer function of neurons in primary auditory cortex (Schreiner & Urbas 1986) matches precisely the modulation spectrum of spontaneous speech (English: Greenberg et al. 1996; Japanese: Arai & Greenberg 1997), as well as the temporal transfer function of the vocal apparatus during speech production (Bouabana & Maeda, in press; Smith et al. 1993). An extension of the modulation spectrum, the "modulation spectrogram" (which embeds the modulation spectral information into a spectrographic format) has been used successfully in automatic speech recognition systems to preserve linguistic features otherwise degraded by acoustic interference (Greenberg & Kingsbury 1997; Kingsbury et al. 1997).

## An account of the locus equation phenomenon based on speech movement planning

Frank H. Guenther

*Department of Cognitive and Neural Systems, Boston University and Research Laboratory of Electronics, Massachusetts Institute of Technology, Boston, MA 02215.* **guenther@cns.bu.edu**
**cns-web.bu.edu/profiles/guenther.html**

**Abstract:** An alternative account of the locus equation phenomenon based on recent theories of speech movement planning is provided. It is similar to Sussman et al.'s account in positing that our productions are tuned to satisfy auditory constraints. It differs by suggesting that the locus equation effect may be an epiphenomenon of a planning process that satisfies simpler auditory constraints.

In the target article, Sussman and colleagues provide a very interesting and thought-provoking theory in which the speech production system develops to produce sounds that satisfy an "orderly output constraint," that is, a consonant-specific linear relationship between second formant (F2) onset and F2 vowel. This output constraint is presumed to reflect an attempt by the motor system to produce sounds that our auditory systems have evolved to prefer. I find the speech production aspects of the theory to be quite plausible, and I am very pleased that Sussman et al. have taken into consideration neurophysiological data in formulating their account because such data are too often overlooked in speech research.

However, equally plausible accounts for the locus equation phenomenon may well exist, and in this commentary I will outline an account based on recent theoretical work investigating speech movement planning. This theoretical work has been implemented as a computational model, called the diva model, that provides a unified explanation for a wide range of speech production phenomena in addition to the locus equation effect (Guenther 1995; Guenther et al. 1997). The account provided here is similar to the target article's account in that it hypothesizes that the speech production mechanism becomes tuned to produce sounds that satisfy important auditory constraints. Unlike the target article's account, however, this account suggests that the locus equation effect may be an epiphenomenon of a movement planning process that utilizes simpler auditory constraints, namely phonemic target regions in auditory perceptual space (Guenther et al. 1997; Perkell et al., in press; see also Savariaux et al. 1995).

Figure 1 provides a schematic view of the speech movement planning process in the diva model. The target for each phoneme is a region in auditory perceptual space (shaded boxes), and movements are planned as trajectories through these target regions. (Only one dimension of this auditory perceptual space, corresponding to F2, will be treated here.) The planned auditory trajectories are transformed into articulator movements through a learned mapping in the diva model, but this process is not important for the current purposes. The model plans auditory trajectories simply by linearly interpolating between the target regions. For a stop consonant, the portion of the movement trajectory during closure takes the form of a "virtual trajectory" passing through the consonant target region because no acoustic signal is produced during this period. It is also assumed that, for each stop category, the release of closure occurs at a roughly constant fraction of the total time for the formant transition; that is, x/T in Figure 1 is approximately constant for each stop class but can vary for different stops. It can be shown that x/T corresponds directly to the slope of the locus equation.

A computer simulation of this simple model was run on consonant-vowel (/CV/) utterances. Ten movements to each of ten vowels were simulated for each consonant. F2 target regions for vowels ranged from 90 to 110% of the "ideal" F2 for the vowel. The value of x/T was allowed to vary by ±10% across productions

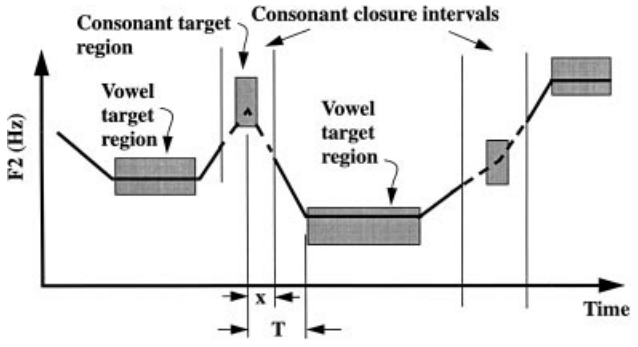



Figure 1 (Guenther). Schematic illustration of a planned F2 trajectory through phoneme target regions for the production of a /VCVCV/ sequence. This simple model of speech production, which utilizes an auditory perceptual reference frame for the planning of speech movements, appears to capture the main aspects of the locus equation effect.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

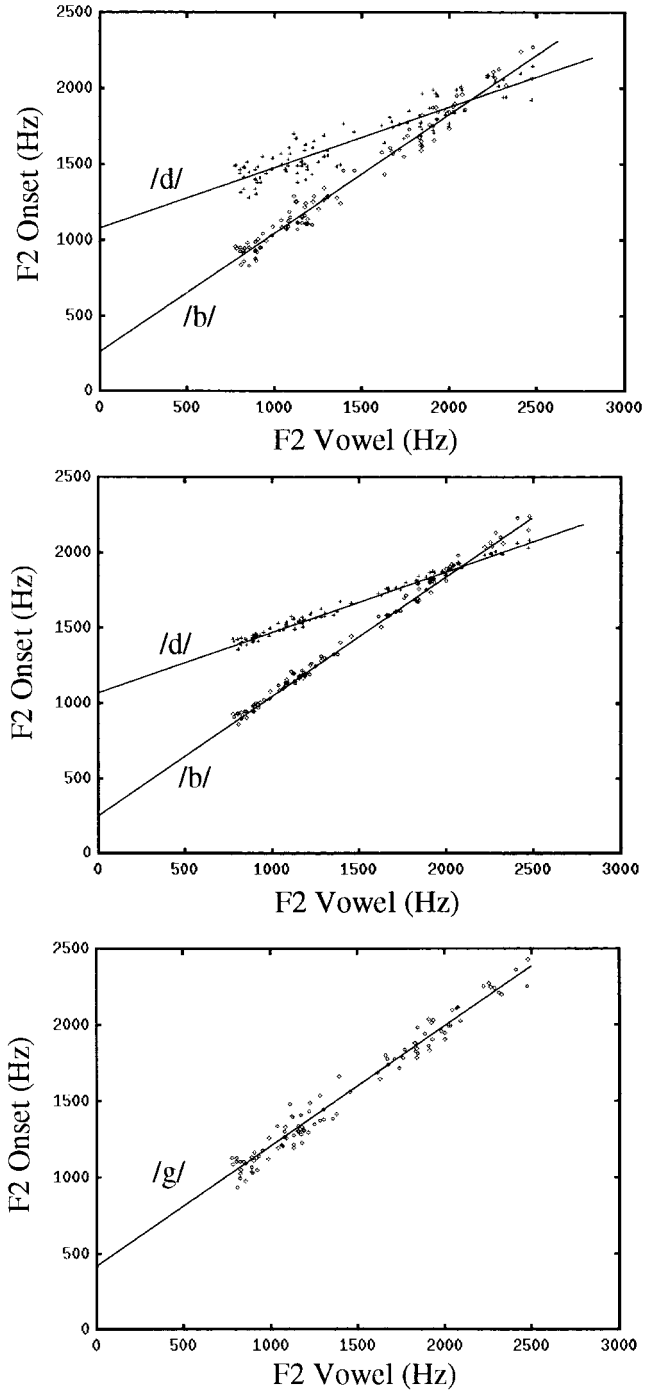



Figure 2 (Guenther). Top. F2 onset versus F2 vowel values generated by the model in Figure 1 when target regions for /b/ and /d/ are large, as may be the case early in development. Middle. Results when target regions for /b/ and /d/ are shrunk down to approximately the sizes estimated by Kewley-Port (1983) for consonant loci. Bottom. Corresponding results for /g/.

in a particular consonant class. F2 target regions for consonants were based on estimated F2 loci reported by Kewley-Port (1983). The point on each target region through which the F2 trajectory passed was chosen at random from a uniform distribution covering the F2 target region.

In the first simulation, F2 target regions for /b/ and /d/ were chosen to be significantly larger than the Kewley-Port (1983) estimates. This is meant to correspond to a young speaker who has not yet fully refined his target regions for consonant productions. The top panel of Figure 2 shows the results of this simulation. As seen in Sussman et al.'s subjects, F2 onset is linearly related to F2 vowel, and the slope and intercept values are comparable to those reported in the target article. Because of the large target regions, however, a relatively large amount of scatter is seen in the data points for each consonant. The bottom two panels of Figure 2 show the results of simulations in which the consonant target regions were shrunk down to the sizes estimated by Kewley-Port (1983). This results in tighter correspondences to the locus equations. It thus appears from these simulations that a speech production model that plans linearly interpolated trajectories through auditory target regions that shrink in size during development can account for both the linear F2 onset versus F2 vowel relationships and the increasingly tight correspondence to the locus equations as development progresses.

This explanation does not depend on the importance of the locus equation phenomenon for perception, although it clearly does not rule out this possibility. Instead, the linearity between F2 onset and F2 vowel is simply a side effect of moving in relatively straight lines between auditory targets. If it turns out that the linear F2 onset versus F2 vowel relationship is indeed central to auditory perception due to inherent properties of auditory brain regions, as hypothesized by Sussman et al., then the model described here may provide an account for how the production system can be relatively easily tuned to obey this relationship. The model as stated here does not take into account coarticulation, although this may become necessary to account for the departure from a single locus equation for /g/ in front versus back vowel contexts. Finally, it should be noted that an account similar to the one provided here, except that auditory perceptual target regions are replaced by constriction target regions, will likely also be able to account for the main aspects of the locus equation phenomenon due to the close relationship between F2 and constriction location. Thus, although I personally agree with Sussman et al.'s assertion that the phenomenon reflects an attempt by the production system to satisfy auditory constraints, more evidence is needed before ruling out the possibility that more "articulatory" sources are responsible for the effect.

ACKNOWLEDGMENTS

The work is supported by NIH grant 1R29-DC02852 and by the Alfred P. Sloan Foundation.

## Linearity or separability?

Bärbel Herrnberger and Günter Ehret

*Department of Comparative Neurobiology, University of Ulm, 89069 Ulm, Germany.* **baerbel.herrnberger@biologie.uni-ulm.de; guenter.ehret@biologie.uni-ulm.de cat.biologie.uni-ulm.de/**

**Abstract:** Sussman et al. state that auditory systems exploit linear correlations in the sound signal in order to identify perceptual categories. Can the auditory system recognize linearity? In bats and owls, separability of emergent features is an additional constraint that goes beyond linearity and for which linearity is not a necessary prerequisite.

There is great fascination in the idea that consonant classification in humans could be done by neuronal mechanisms that existed long before human speech was developed (Ehret 1992). The bat and owl studies indeed provide clear examples of two-dimensional maps of sound parameters that, through their linear correlation, imply an emergent perceptual quality such as relative velocity, object distance, or azimuthal position. Sussman et al. state, as their central point, that auditory systems make use of these linear correlations. This holds in human consonant recognition based on second formant (F2) onset and vowel.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

***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:

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

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

$$F = k * P$$

and

$$k = 1/(2PI * ITD) * P,$$

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



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

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, [gʲ] (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 |



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

successful. Spectral peak location (Heinz & Stevens 1961) or relative amplitude (Hedrick & Ohde 1993) may serve to distinguish non-sibilant /f, v, ɵ, ð/ from sibilant /s, z, ʃ, ʒ/. Within each of these groups, locus equations, spectral peak, or spectral moments (Forrest et al. 1988) can further distinguish place of articulation.

Sussman and colleagues' goal to develop a biologically plausible model of human stop perception based on known neural models of mammalian and avian sound processing is exciting. The perceptual evidence presented in section 6 suggests that listeners may use locus equation information in stop identification. The time course of this process, however, makes this unlikely. To plot a consonant in acoustic space, a locus equation approach requires F2 at onset and at vowel midpoint – an average interval of approximately 60 to 110 msec (Sussman et al. 1991). Thus, the listener would extract F2 at vowel onset and then wait nearly 100 msec for F2 at vowel midpoint to determine the place of articulation of the stop consonant under consideration. Perceptual studies, however, have shown that listeners can successfully identify stops at substantially shorter temporal intervals. For example, listeners classify /b, d, g/ with high accuracy when presented with only the first 10 to 20 msec of stop-vowel syllables (Blumstein & Stevens 1980). Thus, locus equations may be sufficient but not necessary for stop consonant identification. In order to make the temporal scale of locus equations perceptually realistic, it is important to ascertain the minimal temporal interval between F2 onset and F2 vowel that would distinguish stops in terms of place of articulation.

In summary, I believe that the locus equations approach and the neural model for consonant perception outlined by Sussman et al. hold promise. However, more research is needed to determine how well locus equations cue place of articulation across different classes of consonants and to make this locus information match the time scale of human consonant perception.

ACKNOWLEDGMENT

The research reported in this commentary was supported in part by research grant 1 R29 DC 02537-01A1 from the National Institute on Deafness and Other Communication Disorders, NIH.

## Charting speech with bats without requiring maps

Jagmeet S. Kanwal

*Georgetown Institute for Cognitive and Computational Sciences, (GICCS), Georgetown University Medical Center, Washington, DC 20007.*
**kanwalJ@giccs.georgetown.edu**

**Abstract:** The effort to understand speech perception on the basis of relationships between acoustic parameters of speech sounds is to be recommended. Neural specializations (combination-sensitivity) for echolocation, communication, and sound localization probably constitute the common mechanisms of vertebrate auditory processing and may be essential for speech production as well as perception. There is, however, no need for meaningful maps.

A clear, biologically plausible explanation of perception of speech sounds is desperately needed to advance the field of speech processing and perception from its current "muddy" status. At present, no generally acceptable hypothesis exists as to what parameters must be studied to explain categorical perception of speech sounds. Auditorally relevant parametrization of speech sounds is a major contribution of the target article. Sussman et al. present a comprehensive and well-written argument for the role of two parameters – frequency of the second formant (F2) of a vowel and its onset frequency in a consonant-vowel transition – for perception of phonemes /b/, /d/, and /g/ in different allophonic variants. The authors formulate an "orderly output constraint" to define a functional role of the highly correlated and linear relationship between these two parameters. The data on bite-block experiments strongly argue on the importance of such a constraint. These ideas extend to several less successful attempts in the past to establish such relationships.

Putative speech processing mechanisms are equated with the "specializations" for echolocation and sound localization. If similar specializations/mechanisms exist in bats, owls, and humans, then these probably constitute the common substrate of vertebrate auditory processing and may be the most basic factors driving speech perception and production. Such relationships are auditorily driven by evolutionarily conserved mechanisms and may be important for processing contrasting sound categories.

Sussman et al. are to be commended for stepping outside the realm of psychophysics for conceptualizing and integrating available data in a generally readable fashion. It is not clear why F2 onset and offset are so elaborately discussed, however, when a simpler variable, the "frequency range of modulation" or "depth of formant (consonant to vowel) transition" (frequency modulation [FM] depth) could be calculated based on these measurements of F2. This parameter can be robustly represented, because it involves multiple channels of frequency inputs instead of just two (i.e., F2 onset and offset) frequencies. Once FM depth is considered, the role of related parameters such as the slope and/or rate of frequency modulation can and should be investigated. These are biologically plausible parameters because FM selective neurons are documented in the auditory system of several mammalian species (Mendelson et al. 1993; Suga 1964; 1973). This approach would further eliminate concerns that the linear relationship described may be an epiphenomenon because the two frequencies are part of a single frequency modulation pattern. Moreover, there is no clear justification of the reasoning behind measuring the loci of F2 offsets at the F2 vowel midpoint. Would it not be more consistent to measure the extremes of the monotonic part of the formant transition itself? Perhaps FM range and consonant to vowel duration are the useful category level variables and also represent biologically important parameters because duration-selective neurons are shown to be present in the auditory system (Casseday et al. 1994).

It is premature to invoke the presence of two-dimensional maps as algorithms to solve the problem of cognition of different speech sound categories. The argument for the two-dimensional mappability of the measured parameters is weak and sounds teleological (1) because there is no well-established biological constraint suggesting that actual surface maps of these parameters are essential to carry out the necessary discriminations/identifications in the auditory system, and (2) because Sussman et al. suggest that other parameters may also be important, in which case multidimensional representations, perhaps in the form of neural clustering, are more likely to be present than surface maps for each combination of parameters. An example of this is the presence of "blobs" in the primate primary visual cortex for color perception. Similarly, for sound localization in the barn owl, gaze fields in the archistriatal forebrain contain clusters and not maps for spatial perception (Cohen & Knudsen 1995).

Multiple maps have been described in mustached bats for estimating parameters of continuously varying stereotypic stimuli (e.g., target distance encoded by pulse-echo combinations in the time domain). In barn owls, maps are present in the inferior colliculus for localizing sound within a space continuum (Knudsen & Konishi 1978). For meaningful characterization of discrete, complex stimuli, surface maps may be poor and less effective classifiers compared with parameter-related cell clusters. Moreover, if mappability were an important requirement, one would expect to find maps in several species that communicate acoustically. This does not appear to occur. Self-organizing maps represent just one computational strategy for solving many feature extraction problems and may be inadequate or simplistic for representing the complexities of speech.

Although the validity of the theory presented is inadequately tested for human speech perception, it is clearly a bold first step toward relating neurophysiological studies on animal auditory



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

systems to speech perception. In this respect, it challenges the scientific community, especially those working with modern positron emission tomography and functional magnetic resonance imaging techniques, to design experimental tests for mechanisms in the auditory system of humans similar to those found in various animal species using single unit electrophysiology.

ACKNOWLEDGMENTS

This work was supported by a NIH grant DC02054 to JK and a DOD grant DAMD17-93-V-3018 to GICCS.

## Locus equations reveal learnability

Keith R. Kluender

*Department of Psychology, University of Wisconsin, Madison, WI 53706.*
**kluender@macc.wisc.edu**

**Abstract:** Although neural encoding by bats and owls presents seductive analogies, the major contribution of locus equations and orderly output constraints discussed by Sussman et al. is the demonstration that important acoustic information for speech perception can be captured by elegant and neurally-plausible learning processes.

Analogies between communication systems of human and nonhuman animals have been made with some frequency even before the classic comparisons offered by Hockett (1960). Sussman et al. extend these lines of argument beyond communication systems – to echolocation in particular – and beyond analogy to implied homology. From the specialized systems of bats and nocturnal birds, they draw parallels in the hope of revealing mechanisms of human speech perception.

Despite restricting their theorization to the confines of contrasts in place of articulation, there is a good bit to admire in these efforts. The approach is programmatic – extending beyond English to other languages, beyond adults to infants and toddlers, and beyond intact production to acutely and chronically impaired talkers. The major point of this commentary is that the greater worth of Sussman et al.'s efforts can be found in studies with humans and with computer simulations. The downside is that bats and owls lend little more to the story.

The two-dimensional neural representations of bats and owls can be taken as model specialized systems genetically programmed for challenges of nocturnal environments, in which case the question is whether human perceptual processes are similarly specialized for the demands of communication. However, there are multiple findings that argue against recommending such specialization for human speech perception. Alternatively, bats and owls can be taken as models of what neural systems, most generally, do naturally with facility.

What biological systems do well is use multiple sources of sometimes inconsistent or noisy data toward some perceptual end. Most contemporary models of learning and of neural organization are designed to capture just this fact. Furthermore, since the early days of perceptron models, it has been known that linear combinations of attributes make for easiest learning. What Sussman and his colleagues show is that place of articulation, as characterized by covariation between second formant (F2) onset and F2 vowel, can be captured reasonably well by simple linear operations. As such, this is not an argument for specialized processes.

What the orderly output constraint (OOC) may capture best is learnability. Acoustic products, F2 onset and F2 vowel, of articulation may be ideal grist for the simplest sorts of learning. One finding in support of learnability over specialization is the fact that Japanese quail (*Coturnix coturnix japonica*) also can respond differentially on the basis of whether a stop consonant is labial, alveolar, or velar (Kluender et al. 1987; Kluender & Diehl 1987; Lotto et al. 1997). Quail are the unlikely genetic recipients of specialized processes for perception of speech. With brains the size of almonds, quail and their starling cousins (*Sturnis vulgaris*) represent biological approximations to perceptrons, as their performance is consistent with linear operations (Kluender & Diehl 1987; Kluender et al., submitted).

A second argument for learnability lies in the diverse composition of phonetic inventories across the languages of the world. Well over 800 distinct speech sounds, with over 550 consonants, are used across languages (Maddieson 1984). Sussman et al. concentrate on three places of articulation that are common, but not universal, across languages. Hindi, for example, uses four places of articulation – labial, dental, retroflex, and velar. It is likely that F2-onset $\times$ F2-vowel plots for dental and retroflex stops overlap, and it is certain that each (dental and retroflex) would overlap considerably with like plots for alveolar stops.

Should the OOC be the rule across many phonetic contrasts, it certainly would benefit children, Hindi or English, coming to acquire that subset of possible contrasts used in their language environment. This benefit would come from learnability, however, and not from human brains being predisposed specifically for all or most of the many sound contrasts used by languages.

Finally, the elegance of using multiple stimulus attributes does not lie in only using two attributes, in this case F2 onset and F2 vowel. As Sussman et al. seem to appreciate, the simplicity lies in more or less linear combinations. Even for the paradigm case of place of articulation, addition of acoustic information such as characteristics of the third formant (F3) and release burst may improve performance considerably. As impressive as success with only F2 onset and F2 vowel may be, the biological system is not so constrained and the real beauty must be found in the simplicity of the learning operation, not the poverty of the input.

Sussman et al. provide reason for optimism as they present speech perception as a tractable problem. Some caution must be exhibited, however, because there remain some thorny problems when this model is scaled up to fluent connected speech. The problem of coarticulation thus far has been tackled mostly in the forward direction – the relationship between F2 onset and F2 vowel. (Sussman et al. (1997a) have found less success with syllable-final stops.) Coarticulation is pervasive, however. Preceding phonemes can have considerable acoustic and perceptual influence. For example, syllables /da/ and /ga/ are acoustically quite different depending on whether they follow /al/ or /ar/ (Dianora et al. 1996), and locus equations certainly pay some price. Although, in this case, there appear to be general auditory processes unrelated to locus equations that ease the perceptual task (Lotto et al. 1997; Lotto & Kluender, in press), it is not yet clear that the locus-equation approach will scale up well to fluent connected speech. This presents yet another reason to embrace learnability – likely with many more than two sources of information – in the face of noise and variability inherent in natural speech.

Bats and owls might be a distraction. Use of two dimensions underestimates biological capabilities. That being said, Sussman et al. have done a fine service by revealing order where others have overlooked it. They have shown how such order meshes well with neural potential. They have shown how, at least for the cases studied thus far, the linguistic products of vocal tracts can be exquisitely learnable. By extension, they lend encouragement that the problem of speech perception is more tractable, and more general, than typically believed.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

## An articulatory perspective on the locus equation

Björn Lindblom

*Department of Linguistics, Stockholm University, S-10691 Stockholm, Sweden.* **lindblom@ling.su.se**

**Abstract:** Using an articulatory model we show that locus equations make special use of the phonetic space of possible locus patterns. There is nothing articulatorily inevitable about their linearity or slope-intercept characteristics. Nonetheless, articulatory factors do play an important role in the origin of simulated locus equations, but they cannot, by themselves, provide complete explanations for the observed facts. As in other domains, there is interaction between perceptual and motor factors.

It seems possible to look at locus equations as restating the well-known fact that physically adjacent phonemes interact. The properties of a consonant (e.g., its locus pattern) are modified by the following vowel (Öhman 1966) and conversely (Moon & Lindblom 1994). Thus, the locus equation provides a way of quantifying assimilation. Phoneticians usually think of assimilations as articulatory processes that make segments more similar to each other. The engineering approach is to represent articulators as virtual overdamped systems (Lindblom 1983). Assimilations can then be explained as consequences of "articulatory ease," defined as minimization of energy expenditure. Smaller distances and force levels reduce articulatory costs. However, motor optimization is always balanced by the listener's consent, assimilations that occur only when perception permits (Hura et al. 1992).

For a set of /dV/ syllables simulated on apex, an articulatory model (Lindblom et al. 1997; Stark et al. 1996), we investigated the effect of varying tongue parameters on achieving dental stop closures. apex takes input specifications for lips, tongue tip, tongue body, jaw opening, and larynx height and derives an articulatory profile, an area function, and a set of formant frequencies. The distinctive region model (Carré & Mrayati 1992) uses parameters derived from acoustics, whereas those of apex have physiological motivation and vary over empirically determined ranges. For a specific /dV/ syllable, the model offers numerous ways of coarticulating /d/ with /V/ and, hence, of producing many locus patterns. The stop of /du/ can be produced with the tongue body already in position for /u/. In apex this is possible if the tongue tip is raised sufficiently. Alternatively, the occlusion can be made with minimal tongue tip elevation, which calls for a more neutral tongue. The situation is similar for all other /dV/ tasks.

In all probability, this behavior is not an apex idiosyncracy. In many languages a dental/alveolar closure before a back vowel is made with a fronted/palatalized tongue, or, before a front vowel with a posterior/velarized tongue. Such variants occur in English (cf. "*clear*" /l/ in *led* and "*dark*" in *bell*). The existence of such secondary modifications and their acoustic effects suggests that, for any vowel context, second formant loci could theoretically range anywhere between 1000 and 2000 Hz.

Figure 1 shows simulated locus equations for /dV/ syllables. The solid dots pertain to the case where the tongue shape for *V* has been attained during the closure. Here the stop is the result of tip movement only. The unfilled points were obtained by minimizing tongue tip movement, leaving tongue position unchanged but making its shape more neutral. Here the coproduction with the vowel is minimal. Both situations give rise to linear patterns. Maximum coarticulation produces a slope near 1.0 and a small intercept. The neutral-tongue condition forms a more horizontal locus equation (Table 1).

Compared with published data (e.g., Figs. 4 and 8 of Sussman et al.'s target article), these observations lie somewhere between the two extremes, a finding suggesting that locus equations arise from an optimization that (1) minimizes the displacement of the tongue from neutral and (2) minimizes tip elevation.

This account is similar to the explanation of how the jaw and the tongue interact synergistically in vowel production. Acoustically successful compensatory bite-block productions of /i/ involve a superpalatalized tongue shape (Gay et al. 1981). With the jaw locked in an abnormally open position, the tongue body alone produces the palatal constriction, whereas both jaw and tongue normally contribute so as to avoid extreme movements in both articulators (Lindblom 1983).

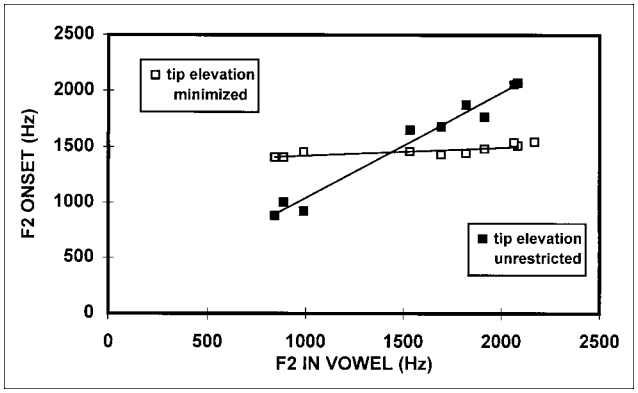



Figure 1 (Lindblom). Simulated locus equations for apical-stop-vowel sequences. The conditions of tongue tip elevation are explained in Table 1.

Should we dismiss perceptual accounts on the basis of these results, and infer that locus equations are simply articulatory in origin? Not at all. There is nothing in the mapping from articulation to acoustics that makes locus equation linearity inevitable. Rather, both the phenomenon of linearity and the specific slope-intercept values reflect implicit "choices" made by speakers and languages. Although apex simulations show that articulatory fac-

Table 1 (Lindblom). *Simulating locus equations for /dV/ syllables*

| Vowel | | Tongue tip elevation: Unrestricted maximum tongue-body coarticulation | Tongue tip elevation: Minimized tongue shape neutralized |
|---|---|---|---|
| | F2(Hz) | F2onset (Hz) | F2onset (Hz) |
| [i:] | 2084 | 2070 | 1513 |
| [e:] | 2064 | 2056 | 1542 |
| [ɛ:] | 1914 | 1770 | 1483 |
| [y:] | 1820 | 1876 | 1446 |
| [ø:] | 1693 | 1682 | 1431 |
| [ʉ:] | 1533 | 1654 | 1459 |
| [u:] | 887 | 1005 | 1406 |
| [o:] | 843 | 882 | 1406 |
| [ɑ:] | 990 | 921 | 1454 |
| Slope | | 0.94 | 0.07 |
| Intercept | | 100 | 1348 |
| $r^2$ | | 0.97 | 0.64 |

Note: The vowels were obtained by searching the APEX vowel space for articulations matching Swedish formant data (Lindblom et al. 1997). The dental stops were modeled by imposing two conditions: (1) unrestricted tongue tip elevation so as to allow complete anticipation of the tongue body shape of the vowel (*maximum coarticulation* with V)and (2) minimizing tip elevation, which results in making APEX tongue shapes more neutral (*minimum coarticulation* with V).



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

tors play an important role in determining locus equation characteristics, the space of possible locus patterns offers a great number of ways in which the locus plot could be either linear or nonlinear. We therefore conclude that, like other phenomena in speech, the patterns underlying locus equations are likely to be products of both articulatory and perceptual selections.

## Integrating cues in speech perception

Dominic W. Massaro

*Department of Psychology, University of California, Santa Cruz, Santa Cruz, CA 95064.* **massaro@fuzzy.ucsc.edu**

**Abstract:** Sussman et al. describe an ecological property of the speech signal that is putatively functional in perception. An important issue, however, is whether their putative cue is an emerging feature or whether the second formant (F2) onset and the F2 vowel actually provide independent cues to perceptual categorization. Regardless of the outcome of this issue, an important goal of speech research is to understand how multiple cues are evaluated and integrated to achieve categorization.

Speech perception represents a prototypical domain of pattern recognition (Massaro 1998). When considered in this light, one charge of the speech scientist is to determine the ecological and functional properties of the speech signal. The ecological properties refer to the information in the speech signal that is potentially informative with respect to the categories of the language. The functional cues are those properties that are actually used in perception. Not all ecological properties are functional; the speech scientist must devise ingenious ways to determine which ecological properties are actually functional in perception. Sussman and his colleagues propose an ecological property of the speech signal that is functional in perception. For consonant-vowel (CV) syllables, this property is the correlation between the second formant at onset (F2 onset) and the second formant in the steady-state vowel (F2 vowel).

Sussman et al., correctly emphasize that this property is not an invariant cue to perception; that is, it is not perfectly reliable in distinguishing among the categories of the language. They admit that other cues, such as the burst of the stop consonant, the spectrum properties at the onset of the consonant, and voice onset time can contribute to the perception of place of articulation. Apparently, however, they do not have a good feeling for how multiple sources of information might work together to influence the perceptual process and to achieve categorization of the input. In addition to not providing a clear description of how their putative cue might be combined with multiple other cues, they believe that somehow component cues in the speech signal must be correlated to achieve categorization. They state, "any learning system (even purely statistical) must rely upon correlations between the inputs to identify and organize them into categories" (sect. 7, para. 1). This statement is incorrect; all that is needed is a correlation between end of the inputs and the resulting categories. In fact, if there are two properties of the speech signal, best performance can be achieved when those properties are completely independent of one another. When there is a perfect correlation between the two properties, it is obvious that the second property cannot provide more information beyond that given by the first property. A partial correlation would provide less information than if the two properties were completely uncorrelated. A simple Bayesian-like integration provides the most information when the properties are completely uncorrelated (Massaro 1998).

Given the importance of correlation among stimulus properties, a major issue in the Sussman et al. approach is whether the perceptual system is indeed using the correlation between the F2 onset and the F2 vowel as the functional cue in perception or whether these two sources of information are being used independently of one another to achieve perceptual recognition. To test this, the framework of the fuzzy logical model of perception (FLMP) can be used to implement the independence of view. In the FLMP, certain properties are evaluated independently of one another and integrated in an optimal fashion, and a categorization decision is made on the basis of the relative goodness of match of the outcome of integration with all of the prototypes or categories in memory. In the independence model it is assumed that the F2 onset and the F2 vowel provide independent sources of information, whereas in the nonindependence model or dependence model, it is assumed that the higher-order property – what Sussman et al. call an emerging feature – is used for perceptual recognition.

It seemed possible to analyze the data provided by Sussman et al. to determine which of these models gives the best description of performance. In their design, there were 11 levels of F2 onset for each vowel, 10 levels of the vowel frequency, and 3 levels of the F3 onset to give 330 test stimuli. It became apparent, however, that this design was not factorial because the F2 onset values differ for the different vowels. Similarly, the 3 levels of F3 onset differed for the 10 different vowels. Basically, the design can be considered to be an 11 by 3 factorial within a vowel category. Thus, one needs 14 parameters for each response alternative. Unfortunately, this design can only ask the question whether F2 onset and F3 provide independent cues, not how F2 onset and F2 vowel are processed. To address this issue, F2 onset and F2 vowel must be manipulated independently of one another. Thus, the experimental design of Sussman et al. falls short of having the potential to test whether F2 onset and F2 vowel are actually evaluated independently of one another and then integrated as described by the FLMP. To address this issue, a true factorial manipulation of F2 onset and F2 vowel must be carried out. Until then, the claim that the correlation between F2 onset and F2 vowel is an emerging higher-order feature that is used in the recognition of CV syllables remains unproven.

ACKNOWLEDGMENTS

This research was supported, in part, by grants from the Public Health Service (PHS R01 DC00236), the National Science Foundation (BNS 8812728), and the University of California, Santa Cruz.

## Why did coarticulation evolve?

Ignatius G. Mattingly

*Haskins Laboratories, New Haven, CT 06511.*
**ignatius@uconnvm.uconn.edu www.haskins.yale.edu**

**Abstract:** The locus equation proposal ignores a fundamental difference between human speech perception and nonhuman echolocation and sound localization, offers a questionable account of the function of consonant-vowel coarticulation, and is further undermined if the effects of other forms of coarticulation are considered. The function of coarticulation is to convey phonetic information rapidly and reliably.

To most people who have thought about speech production and perception, the problem has seemed to be: Given the great variation in the production of particular speech sounds, how can one account for the reliability with which they are perceived? For Sussman et al., however, speech perception presents no mysteries. It requires only the neuroauditory resources known to be available to moustached bats and barn owls. What has to be explained, rather, is the *absence* of variation that is observed, if only the right perspective is adopted, in speech production.

Sussman et al. are not the first investigators to seek inspiration in the ways of bats and owls; Liberman and I have suggested that bat echolocation and owl sound localization were precedents for regarding the speech system as a neurological specialization (Mattingly & Liberman 1988). It is gratifying to see that Sussman et al. have arrived at much the same conclusion, though by a rather different path. It did not occur to us, however, as it has to these



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

authors, to look for the biological origins of speech perception in the specific neurological structures found for echolocation and sound localization, because there is a fundamental difference in function between the human and the nonhuman systems. The speech perception system is very definitely categorical, as phonology requires (Liberman et al. 1957). Within-category acoustic differences among speech sounds are ignored or discarded. The two nonhuman systems, on the other hand, are not categorizing but simply measuring: the velocity and range of the target in the case of the bat and the azimuth of the target in the case of the owl. It is quite misleading to speak of "isovelocity categories" (sect. 1.3.1) and "ITD [interaural time difference] categories" (sect. 1.3.2).

Sussman et al. claim that their locus equations are not merely invariant but linear, and that the speech production system has evolved so as to "enforce" this linearity by adjusting consonant-vowel (CV) coarticulation (sect. 5.3). Note that the requirement for variable CV coarticulation is crucial to their proposal. If the linearity simply followed from the fact that the vocal tract is a system of tubes, there would be no need to look for an auditory constraint that the speech production system must have evolved in order to satisfy. It is therefore rather surprising that, although the authors cite some articulatory evidence for variable CV coarticulation (sect. 5.2, para. 3) and show many linear locus equation plots, they never present both kinds of data for the same utterances.

Even if direct evidence existed to support locus equations in the form of variable CV coarticulation, it would be puzzling that in utterances more complex than CV syllables, second formant (F2) onset and offset are subject to numerous other forms of coarticulation that work against locus equations. For example, F2 onset may be affected by the vowel of the preceding syllable (Öhman 1966) and F2 offset by the degree of stress on the syllable (Lindblom 1963b). Although Sussman et al. and other investigators have looked at other manner classes (sect. 3.2.3) and at stops in other languages (sect. 3.1, para. 1), and have considered sources of variability such as sex, speaking style, speaking rate (sect. 3.2.2, paras. 1 and 2), and bite blocks (sect. 3.2.4, paras. 1 and 2), they do not seem to have tested the stability of locus equations in the presence of these other coarticulatory influences. If they did, they might find that different patterns of coarticulatory influence would yield different sets of locus equations. If, as in Öhman's (1966) vowel-consonant-vowel data, F2 onsets of vowels after /yb/ are consistently higher than those after /ob/, while F2 offsets are hardly affected, two different linear regression functions will result. In general, if F2 onset/offset pairs for various different coarticulatory contexts were plotted together, the result, while still nonrandom, would be quite noisy, and would reveal large areas in which clusters of points for two stops overlapped. In that situation, a combination-sensitive neuron expecting F2 onset/offset pairs falling on one of four straight lines would be in serious trouble.

But if the stability of locus equations is not the adaptive goal of coarticulation, what is? A more plausible account, appealing to perceptual requirements in a different way, might be that the overlapping of articulatory gestures in speech makes possible parallel, hence rapid, transmission of information. Moreover, the timing of the gestures is not random; they are organized into highly restricted syllabic patterns so that acoustic information sufficient to identify each gesture is made available to perception as reliably and quickly as possible. Thus, to borrow Sussman et al.'s own example (sect. 5.2, para. 3), jaw elevation adjustments during the consonant constriction in a CV syllable make information about vowel height available as soon as the constriction is released.

ACKNOWLEDGMENT

Support from NIH grant DC-02717 to Haskins Laboratories is gratefully acknowledged.

## What can auditory neuroethology tell us about speech processing?

David R. Moore and Andrew J. King

*University Laboratory of Physiology, Oxford OX1 3PT, United Kingdom.*
**david.moore@physiol.ox.ac.uk www.physiol.ox.ac.uk**

**Abstract:** A systematic relationship between the acoustic structure and phonemic content of speech raises the possibility that processing strategies similar to those described in animals with highly specialized hearing may also operate in the human brain. This idea could be tested by analyzing animal communication calls into locus equations and using those as stimulus tools in neurophysiological studies of auditory neurons.

The target article attempts the ambitious task of integrating a model of human speech perception with neurophysiological data from two animal species (barn owls and mustached bats) possessing other highly evolved auditory processing mechanisms. For us, the main issue in the target article is whether these processing mechanisms bear any clear relationship to the locus equations favored by the authors as at least a partial resolution of the "noninvariance dilemma." Sussman et al. argue that the processing strategies that have evolved in these animals are likely to have been conserved and that speech processing in humans may also be based on neural processing (combination sensitive neurons and auditory maps) underlying primitive functions such as prey detection and obstacle avoidance. It is also possible, of course, that these functions and speech evolved in a parallel rather than a serial fashion. For example, both birds and mammals evolved from reptiles and many features of the avian auditory system are mechanistically different, although functionally similar, to those of mammals. In some respects the barn owl represents an extreme form of this parallelism. It has sound localization acuity that is equal to or better than that of humans, yet it uses specializations, such as vertically misaligned ears, that are very different from those used by mammals.

Whatever their evolutionary history, the existence of combination-sensitive neurons in amphibians, song birds, and primates may suggest a general mechanism for processing complex sounds. For example, the existence of delay-sensitive frequency-modulated (FM)-constant frequency (CF) neurons in the bat cortex does, in our view, imply that "similar types of auditory neurons could easily have evolved in human auditory substrates to encode the FM and CF components of consonant-vowel utterances" (sect. 1.1, para. 2). However, caution is required in deciding whether these neurons might also represent the relatively more complex locus equations of speech. Although the evidence is strong that the neural maps in barn owls and mustached bats play a role in the processing of signals used in sound localization and echolocation, respectively, these representations are, in both cases, based on a limited range of clearly defined stimulus parameters that are relatively invariant between individuals. In contrast, locus equations do not seem to do a particularly good job in describing differences in place of articulation between subjects (see Fig. 6). In addition, Sussman et al.'s examples of combination-sensitive neurons represent a very broad definition of the term. They are "tuned to coincidence . . . of impulses . . . in the time, frequency and/or amplitude domains" (Suga 1994, p. 143). In fact, coincidence detection of this type is found in all neurons of the central nervous system receiving convergent input. More positively, we believe that the ideas developed by Sussman and colleagues offer the potential for testing whether information-bearing parameters in communicative calls are processed in the way they suggest. A crucial element of this would involve an analysis of animal calls into locus equations. If this were possible, equations defining behaviorally significant features of these calls might be a useful tool for further neurophysiological studies.

It is worth noting that at least some aspects of speech perception are unlikely to be represented as simple linear maps in the brain. Recent studies in the bat have examined the responses of cortical



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

neurons to social communication calls. In contrast to the biosonar signals used for echolocation, there does not appear to be a discrete syllable map. Rather, activity patterns across different cortical areas seem to provide the basis for discriminating different calls (Kanwal 1997). Imaging studies of the human brain suggest that neural activity is distributed across several cortical areas during language processing. Given the variety of sound combinations involved, it seems certain that speech signals are encoded by the spatiotemporal pattern of activity across different areas, although certain cortical fields may be more concerned with processing the semantic or phonetic structure of speech (e.g., Price et al. 1992; Zatorre et al. 1992).

Nevertheless, as Sussman et al. have shown, the relationship between second formant (F2) onset and F2 vowel may provide a particularly reliable and robust cue for identifying the stop place of articulation. Although other features of speech sounds must also be considered, their findings do suggest that certain perceptually relevant aspects of the acoustic structure of speech may be encoded by specific cortical areas in ways that, in other animal species, may be studied using electrophysiological techniques.

## Locus equations and pattern recognition

Terrance M. Nearey

*Linguistics 4-32, University of Alberta, Edmonton, Alberta T6G2E7 Canada.*
**nearey@nova.ling.ualberta.ca**

**Abstract:** Although the relations between second formant (F2) onset and F2 vowel are extremely regular and contain important information about place of articulation of the voiced stops, they are not sufficient for its identification. Using quadratic discriminant analysis of a new data set, it is shown that F3 onset and F3 vowel can also contribute substantial additional information to help identify the consonants.

I am sympathetic to many of the ideas expressed by Sussman and his colleagues (see Nearey 1997). I also look on with interest at their efforts to help build the intellectual bridges to neurophysiology. However, I believe that commitment to the rather strong constraints implied by the authors in their dismissal of third formant (F3) locus equations (sect. 6.1) may be premature.

In 1987, Shammass and I investigated locus equations from a pattern recognition perspective. The following passage from the abstract summarizes our position:

> A regression line fitted to each plot [of F2 onset by F2 vowel] represents an invariant relational property of the corresponding consonant. F2 trajectories are not sufficient to specify the stops uniquely since the lines for the three consonants intersect (indicating category overlap). However, the slopes and the intercepts for the three consonants are distinct and thus represent partly distinctive invariant properties or partial invariants. (Nearey & Shammass 1987)

We went on to show that grossly similar, though somewhat weaker relations obtained in F3 and information from F2 and F3 trajectories could be exploited for pattern recognition.

A new study reported below confirms this finding. The data involved 12 speakers (7 female and 5 male) who each produced stop + vowel + /k/ syllables, with the stop ranging over /b, d/ and /g/ and the vowel over the 10 nonrhotacized vowels traditionally treated as monophthongs in Canadian English. Stimuli were sampled at 16 kHz and low passed filtered at 7.5 kHz and analyzed by linear prediction. Fifteen millisecond Hamming windows with 2 msec frame advance were used. Signals were preemphasized with transfer function ($1.0 - 0.98\ z^{-1}$). A number of coefficients appropriate for the formant ranges of each speaker was determined by examining a few syllables from that speaker. Candidate formant peaks were examined using the graphic display of CSRE 3.0 software. For F1, F2, and F3, piecewise linear formant tracks were drawn by graduate student assistants who were instructed to try to fit the general trajectory with a small number of line segments. For the present analysis, a simple, conservative second-stage formant tracking procedure was used to align the formant candidates to the manually specified guidelines. A candidate peak was assigned to a formant slot if and only if that candidate and a corresponding guideline formant estimate were mutually nearest neighbors. Three point median filtering was applied to the resulting tracks.

The onset of voicing and temporal midpoint of the vocalic section were chosen to extract F2 onset and F2 vowel. Locus plots for F2 pooled across subjects showed similar patterns to those reported elsewhere. Although the variance about the regression lines was greater and differences among the three constants was less salient, linear patterns were also found for F3 and for F1.

Rather than discuss the locus equations themselves, I will focus on results of pattern recognition experiments with the individual stop tokens. A "leave-out-one-subject" cross validation procedure was used throughout. Each of the 15 speakers' data sets was classified using statistics trained on the other 14. (This method is probably more appropriate than the traditional "leave-out-one" cross validation scheme available in most statistical packages, because it is better matched to the repeated measures nature of the original data.)

As noted by Nearey and Shammass (1987), quadratic discrimination is able to exploit the differences in covariance relations represented by varying slopes of some of the locus equations. It can also exploit (more thoroughly than locus equations) other configurational information, including the means of the categories in the pattern space. Quadratic discrimination results based on F2 onset and F2 vowel alone yielded a 62.2% identification rate. (Standard errors of all the identification scores across speakers were less than 3.5 percentage points.) Using F3 onset and F3 vowel alone yielded 54.2% correct. Although less effective than F2, this still represents a substantial gain over chance rate of 33.3%. More importantly, perhaps, combining the F2 and F3 measures lead to a substantial increase in identification scores, 75.6%. This constitutes a reduction of the error rate by more than one third compared with F2 information alone. (Adding F1 to the mix resulted in a slight increase in the score to 78.7%. F1 onset and F1 vowel alone produce 55.0% correct.)

Nossair and Zahorian (1991) have presented results of quadratic discrimination using more elaborate characterization of F1, F2, and F3 trajectories and also using formant amplitudes. They achieved fully cross-validated place identification rates for voiced stops of approximately 85%. This result based on 60 msec sections is quite similar to the performance (approximately 86% correct) Nossair and Zahorian obtained from a panel of five listeners labeling the first 50 msec the same tokens in the most similar listening condition they studied. Quadratic discrimination methods are somewhat more powerful than locus-equation based approaches. However, compared with many current artificial neural network or exemplar-based schemes in the psychological literature, they are capable of carving out only modestly complex decision regions in the pattern space.

The strong linear relations evident in the F2 locus plots are important and deserve additional attention. (They seem particularly promising in providing summary statistics for cross-population studies.) However, if we limit our attention to phenomena that exhibit such striking bivariate regularity to the exclusion of other factors, we may be ignoring much of what makes speech intelligible and relatively noise resistant.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

## Feature extraction and feature interaction

Frank W. Ohl[a] and Henning Scheich[b]

[a]*Department of Molecular and Cell Biology, Division of Neurobiology, University of California, Berkeley, Berkeley, CA 94720;* [b]*Department of Auditory Plasticity and Speech, Federal Institute of Neurobiology (IfN), D-39118 Magdeburg, Germany.* **frankohl@socrates.berkeley.edu; staak@ifn-magdeburg.de**

**Abstract:** The idea of the orderly output constraint is compared with recent findings about the representation of vowels in the auditory cortex of an animal model for human speech sound processing (Ohl & Scheich 1997). The comparison allows a critical consideration of the idea of neuronal "feature extractors," which is of relevance to the noninvariance problem in speech perception.

Sussman et al. hypothesize that neuronal mechanisms of speech processing in humans could exploit evolutionarily conserved auditory processing strategies found in nonhuman species. Part of their argument, which focuses on the second formant (F2) transients in consonant-vowel transition, is based on the robustness of F2 locus equations as linear regressions of the onset frequency and offset frequency of the transient.

A similarly robust feature of auditory perceptual categories when plotted in an acoustically motivated coordinate system is given by the so-called Peterson-Barney map, which reveals the clustering of human vowels by plotting their first two formants (F1 and F2) against each other (Peterson & Barney 1952). Since that discovery, attempts to translate such a map into neuronal space have failed because orthogonal (or at least nonparallel) frequency axes spanning a sufficient frequency range for vowel representation could not be demonstrated in any of the known mammalian auditory maps. Experiments using complex tones with sinusoidally shaped spectral envelops ("ripple spectra") revealed ubiquitous interactions between spectral components of complex sounds in auditory cortical units (Schreiner & Calhoun 1994; Shamma et al. 1995). By a reformulation of the Peterson-Barney-type map emphasizing spectral interactions (a demonstrated neuronal property) rather than spectral filtering (the classically envisaged role of neuronal operation), it was recently possible to show that spectral interaction characteristics are organized in the auditory cortex in such a way that the reformulated mapping is indeed neuronally represented (Ohl & Scheich 1997). The reformulation makes use of a spectral interaction of the form F2-F1 (or similar relations).

In this commentary, we want to point out that these results might have general implications for strategies to solve the noninvariance problem in speech perception. Sussman et al. consider locus equations a partial solution as they focus on acoustic cues for stop consonant place of articulation across vowel contexts (sect. 2). In the subsequent section they allow for other "cues," such as the stop release burst preceding the F2 transition. Generally, these are examples of the question of how to determine the number and set of relevant "cues" or "features" in the acoustic signal. However, even when (a) systematic variations of response selectivity for such features is demonstrated in single neurons, (b) orderly representations of features are shown across neuronal maps, and (c) perceptual relevance is suggested psychophysically by manipulating features as elements of auditory stimuli, it should not be overlooked that such isolated features are in the first place arbitrarily chosen coordinate dimensions thought to provide a suitable acoustic description of the perceptual categories. Consequently, the quantitative nature of the relation between features is likewise a result of that coordinate choice.

On the basis of these arguments it is possible that some "correlation of features" is important in establishing perceptual categories, and facilitates discrimination between different categories. It is probably not a relevant characteristic, however, that correlated aspects fall on a linear regression line. Instead, nonlinearities might be the rule, because the noise resistance of categorization and discrimination depends on various parameters such as the structure of the embedding coordinate space and the internal structure of the categories on the one hand, and the biophysical characteristics of the receptor structures and the interaction characteristics of the processing neuronal network on the other. In the case of vowel representation in cortex, spectral interaction functions of second-formant vowels have been found to be highly nonlinear. By virtue of their topographic organization in the auditory cortex, however, they still give rise to an orderly, that is, monotonic, map. The scaling of the map again reflects the selection of the coordinate space, as discussed for F2-F1 versus F2/F1 mappings of formant interactions (Ohl & Scheich 1997).

Extending these lines of reasoning, the usefulness of the "feature" concept might be questioned even quite generally as a descriptor of neuronal processing mechanisms. Clearly, "features" can always be determined as prototypical attributes of perceptual categories and "features" can also be defined along physical dimensions of receptive fields, and so forth. In the case of vowels, formant coding has been studied classically under the (implicit) assumption that units contribute to the coding of only those parts of the vowel spectrum that correspond to their characteristic frequencies ("feature extraction"). These would in turn require specific convergence circuits for combining relevant formants ("feature binding"). It has been proposed that this idea can be replaced by the assumption of truly parallel processing from the receptive structures to higher brain centers, circumventing the need for separate feature filtering at lower brain stations. Such a neuronal coding strategy resembles psychophysical vowel coding models eliminating the need for spectral peak extraction (e.g., Bladon & Lindblom 1981).

## Locus equation: Assumption and dependencies

Richard E. Pastore and Edward J. Crawley

*Center for Cognitive and Psycholinguistic Sciences, Binghamton University (SUNY-Binghamton), Binghamton, NY 13902-6000.* **pastore@binghamton.edu; br00437@binghamton.edu**

**Abstract:** Evaluating the current locus equation under ideal conditions identifies important and unexpected parameter dependencies. Locus equation (LE) utility, either as a valid laboratory tool or possible invariant cue, depends on stringent specification of critical parameters and rigorous empirical testing.

For decades, researchers have attempted to identify invariant perceptual cues for place of articulation for syllable-initial consonants. Although many individual properties were found to specify categories under very limited conditions, none qualified as an invariant cue. Thus researchers began to evaluate possible complex or relational invariance, including the locus equation (LE). Sussman and colleagues assert that a refocused conceptualization of the second formant (F2)-LE may define an acoustic relational invariant for place of articulation. Does the modern LE add to our knowledge of consonant category perception? The modern LE essentially indicates the rising, flat, or falling nature of the average F2 transition as a function of vowel F2. The LE or its parameters (LE-slope, LE-intercept, $R^2$), suffers from two major limitations; the degree to which (1) F2 attributes specify consonant categories and (2) LE reflects important consonant-relevant, as opposed to consonant-irrelevant, variables.

***Specifying consonant categories.*** Scatterplots of place category exemplars in coordinates of potential cues consistently result in large regions of category overlap, demonstrating clear limitations for the potential cues. Linear regression reduces the graphic representation of category overlap but cannot eliminate actual cue overlap. Unless the regression equation somehow identifies a highly salient emergent perceptual property ("locus" of initial



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

resonance?), Sussman et al. must ultimately fall back on the traditional assumption that place category perception involves multiple cues as a function of vowel context.

Many studies, including recent empirical results from our lab, identify attributes of F2 and F3 transitions and the release burst as contributors to the specification of place categories. In contrast to most previous work, which evaluated only restricted classification for stimuli varying in one stimulus attribute, our work used multiple behavioral measures (open classification, category goodness ratings, and pair-wise similarity) to evaluate perception for matrices of stimuli varying factorially across a number of dimensions (e.g., Pastore et al. 1996). Our results confirm that some aspect of the F2 transition is important in differentiating /b/ from the other voiced stops (/d/ and /g/), but *only* when the vowel F2 is low. This very limited role of F2 is apparent in the target article where, allowing for response bias, classification accuracy can be predicted from the scatterplots (with or without the LE). Our results were consistent with most research (e.g., Stevens & Blumstein 1981), indicating that perceptual contributions from the F2 and F3 transitions, the release burst, and combinations of these stimulus properties are all functions of vowel context.

***Consistent measurement of consonant-relevant dimensions.*** We evaluated locus equation using ideal stimuli: a pure tone of fixed frequency ($F_{initial}$) connected by a transition to a second tone whose frequency ($F_{vowel}$) is varied. LE was computed from the scatterplot of $F_{onset}$ and $F_{vowel}$ for linear and nonlinear (exponential) transitions as a function of (1) temporal location ($T_{onset}$) for measuring *nominal* transition onset frequency ($F_{onset}$), (2) transition duration ($T_{transition}$), and (3) various distributions of $F_{vowel}$ defined by range (1 to 2 octaves) and skewness relative to $F_{initial}$. $F_{vowel}$ is measured at $T_{vowel}$ or transition termination, with $T_{onset}$ and $T_{vowel}$ defined relative to transition onset (thus, $T_{vowel} = T_{transition}$). This evaluation indicates that for $T_{onset}$ specified within a *linear* transition, LE slope equals $T_{onset}/T_{transition}$ (for constant $T_{onset}$, LE-slope is inversely proportional to $T_{transition}$) and, as originally proposed, $F_{initial}$ is indicated by a flat transition ($@F_{onset} = F_{vowel}$). Therefore, under ideal linear conditions, the LE does accurately reflect properties of the initial stimulus. When the temporal stimulus and measurement parameters are unstable or indeterminate, LE reflects this variability, reducing both $R^2$ and accuracy in indicating underlying stimulus properties. When the formant transition is *nonlinear,* the picture becomes even more complicated (including LE reflecting the sampling distribution of $F_{vowel}$), with LE now even more dependent upon experimenter decisions and even less an accurate reflection of any consistent stimulus properties.

Thus, although LE may reflect aspects of resonance properties of initial consonant place categories, it does so in a manner dependent on other variable properties of consonants (e.g., $T_{transition}$) as well as measurement decisions (e.g., $T_{onset}$) and implicit assumptions (e.g., linear transitions) made by the researcher. While we agree that the precise location within the vowel for measuring $F_{vowel}$ is of little consequence (if a consistent criterion is employed), our concern is with other more critical parameters that have not been adequately addressed in the modern LE. In fact, Sussman and Shore (1996) describe the inherent difficulty and lack of stability in specifying $T_{onset}$. If LE is to be considered as a potential laboratory tool for studying place categories, or for the machine recognition of such categories, a careful analysis of the impact of these variables on the LE is required. An empirical validation of this analysis then would be needed before turning to the question of whether LE may be a useful tool for the laboratory classification of place consonant categories.

Despite these limitations as a laboratory research tool, it is possible, at least in theory, that listeners use something such as the LE to perceptually estimate the initial formant resonance. If listeners employ some consistent, but inaccurate, indicator of transition onset and offset (thus defining $T_{transition}$), as well as a consistent criterion for $T_{onset}$, the listener would simply exhibit a consistent (but not unusual) perceptual error in estimating the initial resonance. However, solid perceptual tests are required to make this description of perceptual LE relevance anything more than loose conjecture.

ACKNOWLEDGMENT

This work was supported by Air Force Office of Scientific Research.

## Merits of a Gibsonian approach to speech perception

Jörgen Pind

*Department of Psychology, University of Iceland, Oddi, IS-101 Reykjavík, Iceland.* **jorgen@rhi.hi.is www.hi.is/~jorgen**

**Abstract:** Neurobiologically inspired theories of speech perception such as that proposed by Sussman et al. are useful to the extent that they are able to constrain such theories. If they are simply intended as suggestive analogies, their usefulness is questionable. In such cases it is better to stick with the Gibsonian approach of attempting to isolate invariants in speech and to demonstrate their role for the perceiver in perceptual experiments.

A longstanding problem in studies of speech perception is the invariance question, the fact that speech is highly variable and yet the listener shows impressive constancy in perception. This variability is caused, among other things, by coarticulation, individual differences in vocal tract sizes and shapes, and an ever-changing speaking rate. Confronted with such a variable stimulus it is only natural that speech researchers have spent a good part of their efforts attempting to specify putative invariants, expressed as these may have been in acoustic (Stevens 1989) or motoric terms (Liberman et al. 1967). Now Sussman et al. take another stab at this problem, inspired by neuroethological studies of hearing in barn owls and bats. This is a highly suggestive approach to phonemic perception. The major question it raises is whether it adds substantively to our knowledge of speech perception at this stage. The authors themselves seem ambivalent about the status of their specific hypothesis, the role of second formant (F2) locus equations as "information bearing elements" processed by combination sensitive neurons. They base their own modeling of speech perception on "species-specific" auditory specializations found in the bat and barn owl, though interpreted in "sufficiently abstract terms." I am not quite sure I follow their meaning here. What, indeed, is sufficiently abstract, and what does such abstraction entail for the species-specific status of, for example, bat echolocation?

Sussman et al.'s arguments for the role of locus equations does little to solve the longstanding problem of the perception of stop-consonants. Locus equations are of limited help in explicating the course of stop consonant perception because they do not yield an invariant; there is simply too much overlap between different stop categories. The authors do succeed in putting some order into this by suggesting a "dominance hierarchy hypothesis" whereby, for example, "b" identification "will tend to prevail when tokens fall in the region of overlap between [d] and [b]" (sect. 6.1, para. 3). But this hypothesis lacks independent motivation except that it serves to fit the locus equations to the perceptual facts. Thus it appears to me that the very specific neural model that the authors start out with turns into a rather vague analogy as complicating factors in the perception of stop consonants are added to the picture.

James Gibson, in his quest for a stimulus based theory of perception, eschewed speculations as to underlying neural mechanisms, arguing that perception be understood on its own terms. ("The question is not how the receptors work, or how the nerve cells work, or where the impulses go, but how the system works as a whole"; Gibson 1966, p. 6). This point of view has been considered by many to be too constraining. Good arguments can indeed be made for this in those cases where physiological mea-



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

surements can be conducted in concert with perceptual experiments (Nakayama 1994). Where this is not possible, physiological hypothesis are only of value to the extent that they can constrain or substantially add to perceptual theories; for example, by showing that some putative invariant is specifically tuned to facts of auditory neural processing. The "orderly output constraint" might have served this purpose if in fact the locus equations could serve a similar role in the perception of stop consonants as interaural time difference arrays do in the barn owl. However, they cannot carry this burden on their own since "other information, such as the release burst, shape of the onset spectra, and voice onset time will also contribute to stop place identification during normal speech perception" (sect. 6.1, para. 4). In the absence of a detailed model of the interaction of these various cues, speculations as to a perceptual role for locus equations is difficult to evaluate.

Let me illustrate with an example from my own work of what I take to be the advantage of Gibsonian approach to speech perception. I have for some time been looking at the question of invariance as it relates to the perception of quantity in Icelandic, a language that distinguishes long and short vowels and consonants in stressed syllables (Pind 1986; 1995). Of particular interest are those kinds of syllables where a long vowel is followed by a short consonant or vice versa. Consider typical production data as shown in Figure 1. It can readily be seen that speaking rate affects the overall durations of vowels and consonants. Indeed, a close examination of the figure would reveal that a phonemically short vowel, spoken slowly, can easily become longer than a phonemically long vowel spoken at a fast rate. Because listeners are usually not troubled by changing speaking rates, it may be surmised that some invariant can be found for the speech cue of duration. Indeed, looking at the figure, it can readily be seen that there is no overlap in the data as plotted here on a two-dimensional scatterplot, showing simultaneously vowel and consonant duration. This suggests that a ratio of vowel to consonant duration could serve as the higher-order invariant. This is borne out by perceptual studies that show (Pind 1995) that the listener more or less bisects the vowel-consonant (VC)-plane as shown in Figure 1, hearing syllables of type V-C if the vowel is longer than the consonant and vice versa.

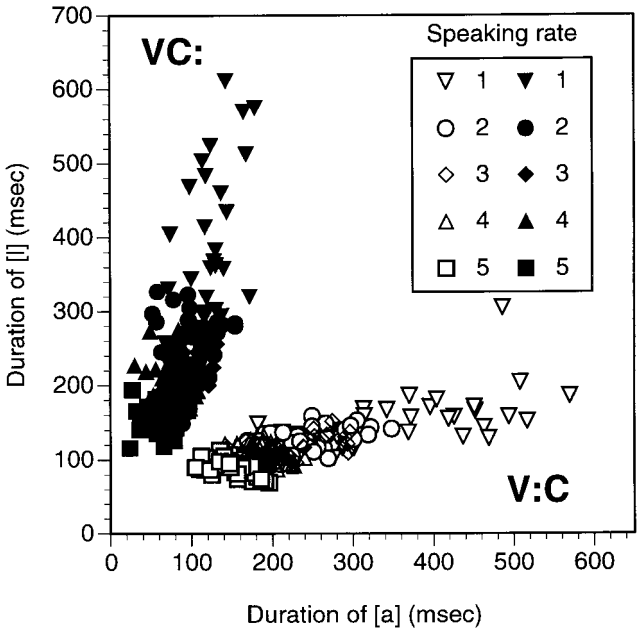



Figure 1 (Pind). Measurements of the durations of the vowel [a] followed by [l] in two-syllabic Icelandic words, spoken by four speakers at five different speaking rates from very slow (1) to very fast (5). The words either have a long vowel followed by a short consonant (type V:C -- open symbols) or vice versa. The distributions of these durations suggest an invariant for quantity expressed in terms of the ratio of vowel to consonant durations (from Pind 1995).

The interesting thing about this relational cue is that it is self-normalizing with respect to speaking rate. Changes in speaking rate will affect the durations of vowels and consonants, and the overall durations of the syllables. The relational speech cue needs no rate adjustments; it will stay invariant in the face of quite large transformations of rate.

Although it has been claimed that the case for invariants in speech is often overstated (Lindblom 1986), I would argue that the notion of invariants provides a convenient reference from which to pursue the study of speech perception. As an exhortation to experimental studies it is still without equal.

## On the ontogeny of combination-sensitive neurons in speech perception

Athanassios Protopapas[a] and Paula Tallal[b]

*[a]Scientific Learning Corporation, Berkeley, CA 94704; [b]Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, NJ 07102.*

**protopap@scilearn.com www.scientificlearning.com; tallal@axon.rutgers.edu**

**Abstract:** The arguments for the orderly output constraint concern phylogenetic matters and do not address the ontogeny of combination-specific neurons and the corresponding processing mechanisms. Locus equations are too variable to be strongly predetermined and too inconsistent to be easily learned. Findings on the development of speech perception and underlying auditory processing must be taken into account in the formulation of neural encoding theories.

The issue of acoustic invariance in phonetic perception has long baffled speech scientists. Reliable derivation of place of articulation from acoustic information remains essentially an unsolved problem, for both automatic speech recognition and human perceptual modeling. Sussman et al. propose that locus equations constitute a consistent cue and speculate on the possibilities for the emergence of the observed regularity and its perceptual significance. Despite several remaining questions, the idea that combination-responsive neurons constitute a cross-species mechanism for solving species-specific problems touches on many important issues. We would like to comment on the interplay between genetic and environmental constraints in the ontogeny of speech perception as it might apply to locus-equation specific, combination-sensitive neurons.

Several lines of evidence support the notion that humans are born with the capacity to discriminate between phonetic contrasts despite cross-linguistic differences that influence subsequent phonetic development (see Jusczyk, 1997, for discussion and review of findings). Neural mechanisms are likely to exist for the detection of formant frequencies, perhaps as an evolution of species-specific call detectors (Rauschecker et al. 1995) or for the estimation of body size (Fitch 1997). Neurons sensitive to spectral energy transitions of specific slopes such as those found in the ferret cortex (Shamma et al. 1993) may in turn constitute formant transition detectors. Whatever the specifics turn out to be, there is certainly a strongly innate component to basic auditory processing that underlies the infant's earliest phonetic perception.

On the other hand, support for a learning-based notion of relatively low-level perceptual functions comes from findings on the phonetic development of language-learning impaired (LLI) children showing that (1) there exist individuals with severe impairments in phonetic perception and in nonspeech auditory processing (Tallal & Piercy 1973; 1974), and (2) the observed deficits in these individuals can be substantially ameliorated through specialized training in auditory processing of speech and nonspeech stimuli (Merzenich et al. 1996; Tallal et al. 1996). There is now mounting evidence to suggest that the perceptual



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

deficits in LLI children are not speech-specific but stem from a generalized impairment in auditory processing (Wright et al. 1997; see Bishop, 1992, and Farmer & Klein, 1995, for review). This impairment has been found to be present within the first 6 months of life in children genetically at risk for LLI and to predict subsequent language delay (Benasich & Tallal 1996). The relatively rapid improvement that can be brought about by specialized auditory training indicates that basic auditory perception underlying speech perception is subject to powerful learning effects, as language-specific phonetic perception must also be.

Analogies from nonhuman species can be powerful when operating on similarly predetermined processing mechanisms, either genetically "hardwired" or strongly biased in terms of physiological and environmental constraints. The speech perception literature, in particular, has gained substantially from cross-species research. The analogies from nonhuman species offered by Sussman et al., however, differ from locus equations and speech perception in some important respects. Specifically, the overlap between locus-equation combination cues for different places of articulation stands in contrast to the unambiguous mapping from combination cues for both the isovelocity categories in the mustached bat and the iso-interaural time difference (ITD) categories in the barn owl. Consequently, what is relatively straightforward for the bat to learn may be very difficult if at all possible in the case of speech perception.

Furthermore, velocity and ITDs are well-defined physical properties that do not vary between individuals, groups, or time frames. In the cases of the nonhuman species used to illustrate the orderly output constraint principle, the corresponding combination-specific neural responses to a great extent may be genetically encoded, as a result of adaptation on an evolutionary time scale. Human listeners, however, must learn (or at least fine-tune) during development the specific places of articulation and their combinations with manner of articulation of their language. In contrast to the nonhuman analogies of Sussman et al., a hardwired processing mechanism for locus equation cues in human speech perception seems unwarranted.

In summary, it is doubtful that locus equations for speech perception are on par with isovelocity or iso-ITD cues, regardless of the relative degree of environmental (signal-bound) and genetic (physiology-bound) constraints. It remains possible, however, that a neural mechanism of cue combination exists that forms higher-order features from perceptual inputs. Advances in neural network simulations have shown many ways in which such learning is possible and, indeed, functional (if still speculative with respect to human perceptual learning). It remains to be specified, however, where in the speech/auditory processing system such combination-sensitive neurons are to be found, to what extent their connectivity (and function) is dependent on the acoustic environment, and how language-specific properties are fine-tuned throughout development.

## Listening to speech in the dark

Robert E. Remez

*Department of Psychology, Barnard College, New York, NY 10027-6598.*
**remez@paradise.barnard.columbia.edu**
**www.columbia.edu/ barnard/psych/fac-rer.html**

**Abstract:** This commentary questions the proposed resemblance between the auditory mechanisms of localization and those of the sensory registration of speech sounds. Comparative evidence, which would show that the neurophysiology of localization is adequate to the task of categorizing consonants, does not exist. In addition, Sussman et al. do not offer sensory or perceptual evidence to confirm the presence in humans of processes promoting phoneme categorization that are analogous to the neurophysiology of localization. Furthermore, the computational simulation of the linear model of second formant variation is not a plausible sensory mechanism for perceiving speech sounds.

Osteoarthritis is universal in humans by age 70. It is also observed in elderly fish, amphibia, reptiles (including dinosaurs), birds, bears, whales, and dolphins. The universality of this form of articular disorder has been taken to reflect the action of a paleozoic mechanism of joint repair rather than a specific disease afflicting humans. A satisfactory account of the biology of osteoarthritis would describe the cellular functions by which the tissues are established, and the mechanical, biochemical, and enzymatic forces that promote hypertrophy. To accomplish this descriptive and explanatory goal, animal models are exploited, and only the species that exhibit the ailment are suitable to model it. Despite wide distribution of degenerative joint disease among vertebrates, it is nonetheless possible to make an unlucky choice of animal model. Bats do not manifest it at all, nor do sloths, though both are bony and are similar in evolutionary history and physiology to animals that, like the rest of us, exhibit structural changes in aged joints.

When contemplating the biology of language, far rarer among species than joint disease, there can be little hope of exploiting an animal model. There is simply no veterinary instance of language. Without an animal model of language, Sussman et al. propose instead to use the mustached bat as a partial model. In doing so, they went out on a limb already well populated by those of us who have asserted analogies between aspects of language and all sorts of ways that animals think or act. The present case is distinguished by a reliance on assertions of rough similarity, on claims that are cautious albeit hopeful, and on indirect empirical tests. Despite its ambition and its well-informed rendition of the neurophysiology of localization, the target article is not convincing about language, leaving even this modest and partial correspondence of human and animal nature merely arguable and conjectured.

The target article does succeed in a goal it set for itself: to propose an analogy between the auditory functions that promote phonetic perception and the neurophysiological vignettes of bats and owls. Indeed, the exposition is a profusion of analogies: (1) Localization by bats is analogous to localization by owls, both using combination-sensitive neurons (sect. 1, para. 2). (2) Auditory localization is analogous to phonetic categorization (sect. 1.2), both requiring the recognition of acoustic elements in combination and permutation. (3) An owl or bat recognizing an auditory pattern is analogous to a human listener recognizing an auditory pattern (sect. 1.3.1). (4) The auditory systems that support these functions are analogous, perhaps necessarily so, if not homologous (sect. 1.3.2). (5) The auditory maps representing interaural phase differences as iso-velocity contours are analogous to maps that represent frequency transitions in formant-centers as iso-stop-place territories, regions within the space unique to phonetic features of place of articulation (sect. 7; Figs. 2 and 16). (6) Localization in bat and owl exploits low-variance linearities in an impinging signal correlated with direction; by analogy, so would an auditory mechanism responsible for pattern recognition in speech (sect. 6.2). (7) The coevolution of auditory and motor components of speech is analogous to the coevolution of the visual sensitivity of bees and the production of pigment by flowers (sect. 6.2). Throughout the exposition, analogies pile up with no defense of the aptness of any of them, a circumstance in which an allegation of unelaborated similarity between localization and categorization of phonetic segments fits. This format allows Sussman et al. to endorse an answer that appeals to them – linearity and low-variance sensory maps – before defining the compliant question. We should find nothing unusual about this. It is a customary pretheoretical way to appraise the psychological applicability of findings in sensory physiology, and is the only way available to us for devising a physiologically justified account of the causes of phonetic perceptual impressions (cf. Rock 1970). When we discover a specific mechanism, we consider the likelihood that its operating characteristic is global, rather than local. Does the strategy work here?

The enterprise fares poorly in implementing a computational analog of this neural mapping mechanism that proves adequate to



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

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



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

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



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

though the correlation between acoustic measurements of F2 and human labeling data is strong.

ACKNOWLEDGMENT
Preparation of this commentary was supported by NIH grant R01 DC00219 to SUNY at Buffalo.

## Input limitations for cortical combination-sensitive neurons coding stop-consonants?

Christoph E. Schreiner

*Coleman Laboratory, W. M. Keck Center for Integrative Neuroscience, Sloan Center for Theoretical Neurobiology, University of California at San Francisco, San Francisco, CA 94143-0732.* **chris@phy.ucsf.edu**
**www.keck.ucsf.edu/**

**Abstract:** A tendency of auditory cortical neurons to respond at the beginning of major transitions in sounds rather than providing a continuously updated spectral-temporal profile may impede the generation of combination-sensitivity for certain classes of stimuli. Potential consequences of the cortical encoding of voiced stop-consonants on representational principles derived from orderly output constraints are discussed.

The basic premise of the target article by Sussman and colleagues – a cortical realization of speech representation as orderly maps of combination-sensitive neurons – is a reasonable working hypothesis. It is supported by some preliminary evidence that combination-sensitive cortical neurons also exist for certain aspects of species-specific vocalizations, particularly on a syllabic level (e.g., Ohlemiller et al. 1995; Rauschecker et al. 1995). However, the neuronal implementation of the proposed representational principles is not entirely straightforward on the level of formant transitions in view of the experimental evidence for representation of voiced stop-consonants in the primary auditory cortex of cats and monkeys. The general electrophysiological finding is that voiced stop-consonant consonant-vowels (CVs), such as /ba/, /da/, and /ga/, result in a single "phasic" or "onset" response at the beginning of the stimulus marking the initial segment of the formant transitions (Eggermont 1995; Schreiner et al. 1996; 1997; Steinschneider 1982; 1994). However, these onset responses show little evidence of the coding of the end of the formant transition that marks the beginning of the steady-state frequency information needed to satisfy the locus equations. By contrast, voiceless transitions, as in /pa/, /ta/, and /ka/, do show a second phasic response corresponding to the onset of voicing and the moment that the formant transition has reached its steady-state value. This implies that at the level of the primary auditory cortex the information for the onset-frequency of the formant transition and the steady-state frequency at the end of the transition are coded robustly and explicitly only for sufficiently long voice-onset times. Hence, the "raw material" for the creation of combination-sensitivity – e.g., for second formant (F2) onset and F2 steady state – in higher cortical regions is potentially available for voiceless but not for voiced stop consonants.

Alternative solutions to this problem may postulate different auditory pathways, special neuronal subpopulations, perhaps located subcortically, or more complex coding schemes that may provide the necessary information to higher cortical stations. For example, it may be sufficient to assume that the rate-of-change in the formant frequencies at the beginning of the transition, in combination with their onset frequency, can be substituted for the parameters currently used in the locus equation. The rate of change in the formant transition also represents a linear correlate and is likely to suffice as an orderly output constraint. Studies of frequency-sweep (frequency-modulated) coding in the auditory cortex in mammals other than bats show neurons with selectivity for different rates of change and sweep directions (e.g., Gaese & Ostwald 1995; Heil et al. 1992; Mendelson et al. 1993; Tian & Rauschecker 1994) suitable for coding formant transitions. In addition, it has been electrophysiologically demonstrated that cortical neurons can be tuned to specific formant ratios (Schreiner & Calhoun 1994; Shamma et al. 1995), making systematic encoding of spectral envelop properties another potential representational basis of static and dynamic speech-sound structures.

Co-existing systematic and overlapping tonotopic, frequency-modulation, and spectral envelop organization of cortical fields may provide a representation of the stop-consonant place information that is based on a distributed population code utilizing spatially dispersed and temporally synchronized cortical cell assemblies (Creutzfeldt et al. 1980; Schreiner & Wong 1996; Wang et al. 1995). Such a code would suffice without the explicit need for combination-sensitive neurons. Which of these scenarios or which combination of them is actually utilized in the human brain requires detailed investigation in several different auditory cortical fields at the cellular level, allowing distinctions between neuronally based combination sensitivity and population based distributed coding.

ACKNOWLEDGMENT
The work is supported by NIH Grants DC 02260 and NS 34835.

## Locus equations in models of human classification behavior

Roel Smits

*Department of Phonetics and Linguistics, University College London, London NW1 2HE, United Kingdom.* **roel@phon.ucl.ac.uk**
**www.phon.ucl.ac.uk/home/roel/home.htm**

**Abstract:** The potential role of locus equations in three existing models of human classification behavior is examined. Locus equations can play a useful role in single-prototype and boundary-based models for human consonant recognition by reducing model complexity.

Sussman et al. make a convincing case that speakers producing consonant-vowel (CV) syllables actively control the movement of their articulators so that the frequency of second formant (F2) sampled at voicing onset and in the vowel nucleus show a high degree of regularity. They argue that speakers do so for a communicative purpose. However, if we reason strictly from the perspective of the listener, whose task it is to classify a *single* stimulus at a time, it is not trivial exactly how the observed regularity actually aids the classification process. Indeed, although Sussman et al. observe a fair match between acoustic data and perceptual data (sect. 6.1), it is not made clear what explicit role is played here by the locus equations.

There are two basic ways in which a regularity in the incoming data might be beneficial to a classifier: (1) by increasing classification accuracy; and (2) by reducing classifier complexity. In this commentary I will first examine how existing quantitative models of human classification behavior would deal with the problem of classifying stimuli on the basis of (F2 vowel, F2 onset). Then I will consider whether the regularity captured by locus equations would actually aid classification in either of the two ways mentioned.

Currently, three models of human classification behavior are successful and popular: the single-prototype similarity-choice model (SPSCM), the multi-exemplar similarity-choice model (MESCM), and the boundary-based recognition model (BBRM). In the SPSCM each response class is represented by a single prototype, which is a point in a multidimensional perceptual space (the F2 vowel/F2 onset plane, in this case). A stimulus is mapped to a point in this space and the similarity of the stimulus to each of the classes is inversely related to the distance between the stimulus and each of the prototypes. The probability of choosing a particular response is proportional to the similarity of the stimulus to the prototype associated with that response (Shepard 1958). Locus equations can be considered prototypes for consonantal



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

place of articulation. However, locus equations are one-dimensional prototypes (lines), rather than the conventional zero-dimensional ones (points). Thus, additional assumptions are needed to quantify the similarity calculation, for example, using the distance of a stimulus to its projection on the locus equation.

Sussman et al. kindly provided a set of 450 (F2 vowel, F2 onset) data for three speakers. Figure 1a displays these data as well as the associated locus equations. I calculated response probabilities for /b, d, g/ on the basis of these locus equations and the classification strategy proposed above, using a Euclidean distance measure and a Gaussian distance-to-similarity mapping. The resulting territorial plot is presented in Figure 1b. The different regions in this plot indicate regions of the perceptual space in which a particular response (indicated by the phonetic symbols) is the most likely. The dotted parallelogram indicates the region containing the stimuli in the reported perception experiment. In all examples the response biases for /b/, /d/, and /g/ were set to 1.0, 1.5, and 1.0, respectively.

In the MESCM (Nosofsky 1986) each class is represented by a large number of "prototypes" or exemplars. The similarity of a stimulus to a class is defined as the sum of the similarities of the stimulus to all exemplars in the class. Figure 1c represents the territorial plot for the MESCM based on the same data.

In the BBRM (Ashby & Perrin 1988) optimal class boundaries are computed on the basis of observed distributions of data. Assuming the data are normally distributed, optimal quadratic class boundaries were calculated for the locus equation data. These are shown in figure 1d.

Inspection of Figures 1b, 1c, and 1d reveals that, although the three classification models are based on very distinct assumptions, their predicted classification behavior is not vastly different, at least not within the parallelogram. It would be very interesting to fit these models on the classification data from Sussman et al.'s perception experiment, and hypothesize on the underlying mechanisms in the listeners' classification behavior on the basis of the goodness of fit for each of the models. Prior to such evaluation, however, it should be considered what role is actually played by the locus-equation regularity in each of the models, and whether this regularity actually aids the classification process, using the criteria of model complexity and classification accuracy. In the SPSCM the locus equation plays a very explicit role, resulting in low model complexity. The linearity of the data allows each class to be represented by a single, albeit one-dimensional, prototype (the locus equation), thus using only six parameters plus two biases. The MESCM classification is essentially based on comparisons to exemplars, and locus equations do not play any role whatsoever; nor does the extreme linearity of the data necessarily enhance class separability. Finally, in the BBRM, locus equations as such do not play an explicit role. However, the regularity of the data does allow each class to be accurately represented by a single two-dimensional Gaussian, which keeps the model complexity relatively low at 15 parameters plus two biases.

In conclusion, locus equations can play a useful role in single-prototype and boundary-based models for human consonant recognition by reducing the model complexity. Locus equations and multi-exemplar-based models, on the other hand, are incompatible.

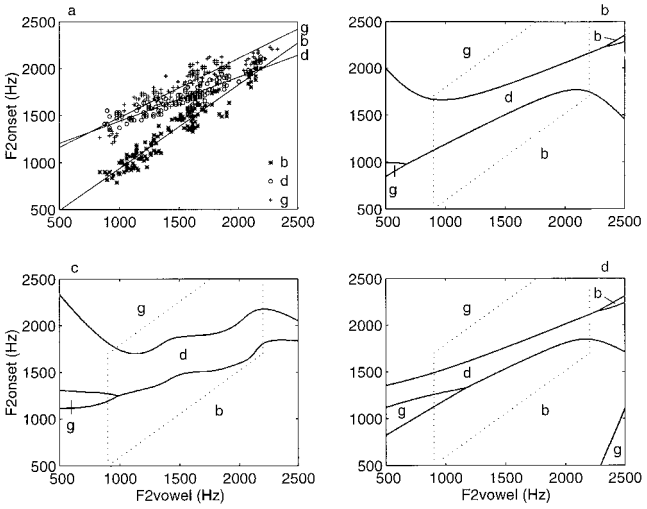



Figure 1 (Smits). *a.* Acoustic data and fitted locus equations; *b.* territorial plot for SPSCM; *c.* territorial plot for MESCM; *d.* territorial plot for BBRM.

## Evolutionary conservation and ontogenetic emergence of neural algorithms

Hermann Wagner and Dirk Kautz

*Institut für Biologie II, RWTH Aachen, D-52074 Aachen, Germany.*
**wagner@tyto.bio2.rwth-aachen.de; kautz@nke.de**
**birdland.bio2.rwth-aachen.de**

**Abstract:** Neural algorithms are conserved during evolution. Neurons with different shapes and using different molecular mechanisms can perform the same computation. However, evolutionary conservation of neural algorithms is not sufficient for claiming the realization of an algorithm for a specific computational problem. A plausible scheme for ontogenetic emergence of the structure of the algorithm must also be provided.

In their target article, Sussman et al. use examples from neuroethology to suggest a partial solution to the noninvariance dilemma in speech perception. Speech perception depends on neural computations just as do the determination of sound locus and the extraction of biosonar information. The authors claim that "[s]peech sounds . . . are not, in principle, that different from biologically important sounds" (introduction). The processing of spatial sound attributes such as Doppler shifts or interaural time difference is assumed to pose computational problems equivalent to some of the problems underlying speech perception. Therefore, algorithms found in other species might serve as models for analyzing processes involved in speech perception. Can such an analogy be drawn? In our opinion, the key here is the question about the evolutionary conservation of neural algorithms, because, obviously, the neural substrates are different.

Studies of neural computations in various species as well as within different nuclei in one species suggest that neural algorithms are conserved. We shall discuss the example of coincidence detection that is crucial for the combination-sensitive neurons postulated in the target article (sect. 1.1).

The term coincidence detection means that a neuron's response depends on the temporal difference in the time of arrival of inputs to this neuron. If two spikes arrive simultaneously, they may fire a neuron, but if they arrive at different times, the neuron will be silent, because the firing threshold is only reached if two simultaneous spikes produce a high enough excitatory postsynaptic potential. In this way, a neuron can bind together two inputs. Coincidence detection plays a role in many neuronal computations in different neural substrates and on a large range of time scales: associative learning, motion detection, measurement of interaural time difference for sound localization, long-term potentiation, synchronization of neural activity, range detection in bats, depth vision by spatial or temporal disparity, and coordination of cerebellar activity. The computation is performed by neurons having quite different morphology: from pyramidal cells in the mammalian cortex to dendrite-lacking cells in the nucleus laminaris of the owl. Likewise, several different molecular mechanisms are involved; for example, NMDA-receptors in the hippocampus, outward-rectifying potassium channels in the auditory system, 5-hydroxytryptamine receptors in conjunction with G-protein–



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

dependent intracellular cascades in sensitization in the snail. The different computational speeds of the molecular processes, together with neural gross morphology and conduction times, account for the wide range in relevant time scales: from microseconds in measuring interaural time differences to seconds in associative learning. Nevertheless, the formal description, the algorithm, is always the same. These observations make it plausible that coincidence-detecting or combination-sensitive neurons also play a role in speech perception. The response properties of such neurons may be shaped in ontogeny by a Hebbian-type of mechanism.

Coincidence detection does not suffice, however, for the extraction of phonemes. The integration of the information contained in the responses of many combination-sensitive neurons is necessary. The locus equations (sect. 3) suggest a possible way of combining the information. A mechanism that can achieve this integration has been described in the owl (Wagner et al. 1987). Sussman et al. refer to this work in the target article and emphasize the linear dependence between two variant "input" parameters in the emergence of a new, invariant parameter (sect. 1.3.2). Can the mechanisms detected in the owl be transferred to speech perception? The linearity requirement might be more mathematical hocus-pocus than real biological necessity. We know that in the owl, inputs are linked together that originate from the same location in space, not those that have the input combinations that fit to a straight line (Brainard et al. 1992). Because interaural time differences (ITDs) originating from one location may vary, this is biologically sensible, because what the owl needs to localize is one locus in space and not one ITD. This leads us to a second comment: since ITDs originating from one point in space change during development and depend on the individual shape of the sound-receiving systems (ruff, ear flap, middle ear), their relation cannot be genetically preprogrammed. It must be shaped during ontogeny. In the case of the arrays in the owls, this is possible because the signals that have to be bound together are always present simultaneously in the acoustic signal. Hebbian types of synapses might do this job. We see problems here for the proposed speech-perception analogy. As with the sound localization cues, the variation in vocal tract morphology is too large for genetic preprogramming of the locus equations. Thus, there should be a plausible explanation of how the information-bearing units can emerge in ontogeny. This topic is not dealt with in the target article. The problem we see is that the signals that should form the spectrotopic maps (Fig. 17 in target article) are not present in the signal simultaneously. How, then, can they be combined?

## Combination-sensitive neurons: A flexible neural strategy for analyzing correlated elements in sounds

Jeffrey J. Wenstrup

*Department of Neurobiology, Northeastern Ohio Universities College of Medicine, Rootstown, OH 44272-0095.* **jjw@neoucom.edu**
**web.neoucom.edu/depts/neur/web/graduate/wenstrup.html**

**Abstract:** Combination-sensitive neurons serve as the fundamental processing unit in Sussman and colleagues' proposal for the neural representation of stop consonants. This commentary describes recent studies in the mustached bat that show how ubiquitous and flexible this neural strategy can be. Sussman et al.'s proposal is an important contribution to a neuroethological consideration of speech perception.

It is central to Sussman et al.'s view that the higher order mechanisms and representations used to analyze complex sounds in other vertebrates are exploited and specialized for the analyses of human speech sounds. For those who share this view (as I do), the resulting questions include: What mechanisms and representations, and what kinds of specializations for what kinds of speech sounds? The target article describes how a particular class of neuron (combination-sensitive) implements an analytic approach (locus equations) in categorizing phoneme-level speech sounds. This commentary will focus on the neural substrate, combination-sensitive neurons, describing their capabilities and whether their encoding features are useful in the way described in the target article. Emphasis is on the mustached bat, because these neurons are perhaps best described in that species. The main point is that combination-sensitive neurons provide a flexible neural strategy for the analysis of correlated elements within acoustic signals.

The two best known classes of combination sensitivity in the mustached bat are CF/CF and FM-FM neurons (Suga et al. 1983). These compare different harmonic elements in the emitted pulse and returning echo for constant frequency (CF) or frequency modulated (FM) sonar components. Each neuronal class is located in functionally specialized regions devoted to the analysis and systematic representation of target velocity or target distance. These highly specialized representations are generally viewed as utilizing a processing strategy that is distinct from what occurs in the tonotopically organized parts of the ascending auditory pathway.

Recent studies in the auditory cortex and inferior colliculus have changed this view substantially. Fitzpatrick et al. (1993) reported large numbers of neurons combining sensitivity to an FM component in the fundamental of the emitted pulse and a higher harmonic CF component in the returning echo. Because these neurons occur in primary auditory cortex, the finding showed that combination-sensitive responses are used in processing by tonotopically organized parts of the auditory pathway. Ohlemiller et al. (1996), recording from echo delay-sensitive FM-FM neurons in auditory cortex, found that many also responded to communication signals with a frequency structure similar to sonar calls, but a different temporal structure.

Studies of the inferior colliculus (IC) demonstrate that combination-sensitive responses are not unique to the auditory forebrain. They are abundant in the IC of the mustached bat, where roughly two-thirds of the neurons display two separately tuned frequency sensitivities to sounds (Leroy & Wenstrup 1996; Mittmann & Wenstrup 1995; Portfors & Wenstrup, in preparation; Yan & Suga 1996). What is particularly surprising is the broad range of combinatorial properties – both in terms of the frequency bands that contribute to these combinations and the interactions between the inputs. The frequency combinations include responsiveness to harmonic elements in social communication calls; for example, a common form of combination-sensitive tuning is to frequency bands just below and just above the first sonar harmonic, frequency bands containing two major energy peaks for a class of social communication signals (Kanwal et al. 1994). For combination-sensitive interactions, previous descriptions have generally emphasized the facilitative interactions between the spectral components. However, in the IC, the frequency combinations display a broader range of interactions – inhibitory, neutral, and facilitative. These can be useful in a correspondingly broader range of computational solutions.

Two other response features add to the flexibility of these neuronal populations. One property is their temporal specificity. FM-FM neurons, tuned to pulse-echo delay, are the best example in the mustached bat, but most other combination-sensitive neurons are also sensitive to the timing of the two inputs. Thus, the focus on the comparison between spectral elements should not obscure the temporal specificity of the interactions. The second feature is the ability to create additional response selectivities. An example, from the mustached bat's IC, is a subpopulation of neurons tuned to nonsonar frequency bands (Leroy & Wenstrup 1996). Like other combination-sensitive neurons, they display responsiveness to two frequency bands, but these require the activating signals in each frequency band to be acoustically complex (e.g., narrow-band noise). This additional selectivity allows these neurons to respond well to one class of social vocalizations but not to another having the same spectral peaks. Thus, studies in



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

the mustached bat's IC demonstrate the potential to initiate a wide range of specialized, higher order analyses at relatively early stages of the ascending auditory pathway. These frequency combinations, interactions, temporal selectivity, and additional specializations all seem well-suited for representations of elements of speech sounds.

But do they compute locus equations? I focused on the potential of combination-sensitive neurons to perform these analyses because the details of an implementation are difficult to predict. As an example, it is clearly possible for combination-sensitive neurons to form a two way frequency matrix for the frequency ranges of the second formant (F2) onset and F2 vowel. However, the tuning of neurons in the F2 frequency range is broad, at least as observed at the level of the auditory nerve. At sound levels characteristic of speech, the response areas of single neurons probably include both F2 onset and F2 vowel frequencies, providing little of the discrimination of frequencies important to this model. Temporal features of neuronal responses (e.g., selectivity for the extent, direction, or rate of frequency transition, response to the burst) may be crucial in coding the F2 transition and would be essential for any implementation. Regardless of the details of this implementation, the target article makes an important contribution by showing how a flexible and widely used vertebrate processing scheme may function in a particularly complex perceptual system.

ACKNOWLEDGMENTS

I thank C. V. Portfors for helpful comments. Some work was supported by the National Institute for Deafness and Other Communication Disorders.

# Authors' Response

## Human speech: A tinkerer's delight

Harvey M. Sussman,[a] David Fruchter,[b] Jon Hilbert,[c] and Joseph Sirosh[d]

*[a]Department of Linguistics and Communication Sciences and Disorders; [b]Department of Linguistics; [c]Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712; [d]HNC Software, Inc., San Diego, CA 92121.* **sussman@mail.utexas.edu; fruchter@mail.utexas.edu; sirosh@hnc.com**

**Abstract:** The most frequent criticism of the target article is the lack of clear separability of human speech data relative to neuroethological data. A rationalization for this difference was sought in the tinkered nature of such new adaptations as human speech. Basic theoretical premises were defended, and new data were presented to support a claim that speakers maintain a low-noise relationship between F2 transition onset and offset frequencies for stops in pre-vocalic positions through articulatory choices. It remains a viable and testable hypothesis that the phenomenon described by the locus equation is a functional adaptation of production mechanisms to processing preferences of the auditory system.

There are commonalities between animal and human communication systems. In the target article we focused on a subset of commonalities deemed pertinent to the neural processing of human speech sounds, especially sounds characterized by noninvariance – stop consonants produced in varying vowel contexts. A strategy of comparing speech to neuroethological models was adopted because we observed empirical phenomena in both realms that shared several intriguing features. François Jacob once said: "To produce a valuable observation, one has first to have an idea of what to observe, a preconception of what is possible" (1977, p. 1161). Upon looking at the linear regularities in Doppler-shifted harmonic relationships coding target velocities in the mustached bat and phase/frequency relations coding interaural time differences (ITDs) in the barn owl, it became apparent that our locus equation data bore a reasonable resemblance to the form of these input signals. The processing mechanisms common to the animal models – combination-sensitive neurons and two-dimensional (2-D) mapping of correlated variables to yield an emergent property – were viewed as possible examples of evolutionarily conserved auditory processing strategies that humans could use to encode speech sound categories. A programmatic plan of study ensued to extend the locus equation phenomenon, explore constraints, and speculate on possible functional origins.

### R1. The overlap problem

In response to the criticism most frequently encountered in the commentaries – the overlap of locus equations in selected regions of acoustic and auditory/perceptual space (**Blumstein, Brancazio, Carré, Diehl, Fowler, Govindarajan, Herrnberger & Ehret, Jongman, Moore & King, Nearey, Pastore & Crawley, Pind, Protopapas & Tallal, Sawusch**) – we offer the following. The lack of complete separability among consonant-vowel (CV) categories stands in stark contrast to the perfect separation of ITDs and velocity formulations in the barn owl and mustached bat. To paraphrase **Herrnberger & Ehret,** the identification of a given CV does not uniquely fall out of the coordinate's position in the 2-D decision space. The neuroethology examples take advantage of laws of physics that uniquely specify input signals in 2-D space. The F2 transition in CV utterances does not operate like a reliable Doppler-shift or a pulse/echo-delay distance function. Why does the single most important cue in speech perception, the F2 transition, need so much additional help? A likely place to start looking for answers is the nature of the speech production-perception process itself. The overlaid speech production-perception system involves several factors that preclude coding by simple physical laws: (1) speakers can modulate their style and rate of speaking and thus the acoustic integrity of the signal; (2) control of the speech motor system is characterized by comparatively many degrees of freedom; (3) motor equivalence is the norm precluding any simple acoustic-to-articulation mapping (the inverse problem); (4) phonemes assimilate with neighboring sounds creating coarticulated entities lacking any transparent isomorphism to linguistic units; (5) the cues for segments are redundantly coded; (6) acoustic correlates of a segment often exhibit trading-relations in specific contexts. These characteristics of spoken languages are some of the reasons the average two-year-old can outperform the most advanced speech recognition system.

If the lack of single-cue-dependent separability of human CV processing is to be compared to the non-overlapped scenarios in bats and barn owls, then an evolution-based account is necessary to justify its encoding complexity and obvious imperfections. Engineers design machine-based recognition systems, and evolution designed the human brain: "natural selection does not work as an engineer works. It works like a tinkerer" (Jacob 1977, p. 1163). Human speech perception is the late-comer with



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

respect to sound processing. It was not designed de novo to handle overlapped speech sounds. What worked so perfectly in ancestral forms was not completely adequate for the task at hand. The computational mechanisms that were evolutionarily conserved had to be tinkered with as these new signal forms necessitated altered combinatorial algorithms using already functioning processors. There were lots of "spare parts" to work with. It is important to note that these "spare parts" worked very well for the tasks that they had already evolved to handle, for example, CF and FM analysis, noise analysis, and combinatorial spectral and durational analyses. Some reshuffling was needed, together with a division of labor, to handle the complex nuances of this new signal. Some acoustic parameters worked well in some contexts and failed in others. Where F2 transitions were confusable, a greater reliance on burst cues, voice onset times (VOT), or F3 onsets were built in. In terms of elegance and simplicity it was far from perfect, but it worked nevertheless.

The encouraging news coming out of neuroethology is that every potential acoustic cue for speech CV processing can be related to documented neural mechanisms. Whether it be dynamic frequency changes over time, FM-CF relationships, the coding of noise burst features both spectral and temporal, or transforms between aspects of the above, neurons have been found that can, in principle, detect and signal such properties. The exact combinatorial arrangements still need to be specified to make sense of the tinkered human system. Passing the buck to a direct-realist position (**Fowler, Brancazio**) or to a speech-is-special module (**Mattingly**) will not solve the problem. Brancazio and Fowler's locus equation+ model correctly classified input tokens into "bdg" categories with 77.1% accuracy (chance = 33%). This "poor" showing, according to those authors, was taken as cause for dismissing the perceptual relevance of locus equations. We argue instead that the missing 22.9% will be found when the added elements of the tinkered system are included in the modeling – burst, VOT, and F3 information.

**Damper's** commentary is relevant to this point in the sense that he recommends greater utilization of data-driven or, as he puts it, "ignorance-based" research strategies, in the quest to uncover other statistical regularities hidden within the variability of the complex speech signal. Neural networks can serve as expedient research tools to rule out and/or uncover additional self-learning and self organizing relationships in the input signal. For example, as yet undiscovered correlated relationships must exist between spectral properties of the noise burst and F2 and F3 formant information. We do not necessarily agree, however, that such automatic discovery procedures should totally supplant traditional knowledge-based scientific inquiry, but they certainly can eliminate traveling down many ill-fated garden paths.

## R2. Inappropriateness of the neuroethology analogy

**R2.1. Lack of signal correspondence.** The analogy is "too extreme," **Fowler** suggests, even if considered at an abstract level. She points to dissimilarities in the make-up of input signals to the bat, barn owl, and human. Exactness in matching all details of the analogy between human and nonhuman systems is unrealistic and, more importantly, irrelevant to the thrust of the argument. Our primary concern is the computational commonalities that can be identified across species. Ehret (1992), in comparing species as diverse as the mouse, chinchilla, monkey, cat, and bat, has outlined four examples of general preadaptations for speech-specific perceptual features: categorical perception, perceptual constancy, perception of formant structure, and phoneme-like perception. As concluded by Ehret: "Mammalian auditory pathways are adequate systems for testing hypotheses about mechanisms of human speech perception, provided that species-specific calls are used as stimuli, not human speech" (p. 108). **Wagner & Kautz** concur: "speech perception depends on neural computations just as do the determination of sound locus and the extraction of biosonar information."

**Ryan et al.** suggest that sound localization processing is a suitable source of preadaptations because there were by necessity (viz., acoustic laws) "low-noise relationships of acoustic parameters that can be used in localization." This led to the general use of combination-sensitive neurons and auditory maps to process and represent these highly correlated acoustic variables. As stated by Ryan and colleagues: "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." Evidently, evolutionary biologists and neuroethologists have no objection to generalizing across species, despite a lack of precise signal congruence in species-specific sound processing.

**R2.2. Lack of hard evidence.** It is suggested by **Remez** that "there can be little hope of exploiting an animal model [as] there is simply no veterinary instance of language." Our analogy deals with a very early stage of language processing – "phonetic pre-processing," which represents simply another case of complex sound processing for communication purposes, for which there are many, many animal models. Remez adopts a very parochial approach and will not accept the logic/data of any analogy until bats and barn owls are found that encode stop consonant + vowel syllables, or until humans perform biosonar navigation. We agree that it is important to choose animal models judiciously, and that an "unlucky choice of models" is quite possible. However, despite the fact that there are so few well understood animal models of complex-sound processing, we were willing to risk generalizing from only two examples because of the potential importance of any insight into speech perception. We have never claimed that the OOC hypothesis is at this point anything other than "arguable and conjectured" (Remez), but the possibility is intriguing enough that it deserves to be aired. It is an early hypothesis lacking in conclusive proof. However, we should not fear to generalize because we might be overgeneralizing; if no generalizations are made, the appropriate degree of generalization will never be determined.

**R2.3. Continuous versus categorical processing.** The adequacy of the analogy was questioned by **Mattingly** because "there is a fundamental difference in function between the human and the nonhuman systems." Humans process speech categorically, while the bat and barn owl process sounds continuously. This point was also brought up by **Herrnberger & Ehret.** Note 3 in the target article readily admitted this fact and apologized for our liberal use



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

of the term "category." The continuity-discontinuity feature of the acoustic signals is, however, irrelevant to our argument. This dichotomy is simply a reflection of species-specific ecological requirements. For example, the house mouse exhibits classic categorical perception of mouse pup ultrasounds in both frequency and temporal domains (Ehret & Haack 1981; 1982). Mother mice need only detect alarm calls from their pups, a category-type classification. Does this penchant for categorizing make the mouse neural system a more pertinent model for human speech? Another example of species-specific categorical perception can be found in Japanese macaques that process two categories of "coo" sounds carrying very different information about the sender (May et al. 1989). For the bat, prey speeds and ranges do not call for discrete representations with "allophonic" insensitivities within and heightened "phonemic" discriminability across categories.

At another level of argument the categorical-continuous difference is not relevant to locus equation data. Our basic premise is that an initial representation of stop place categories is likely formed by a neural representation of F2 onset vs. F2 vowel frequencies. At this initial processing stage there is a continuous representation of frequency coordinates, as they range across speakers. Categorization of input signals is not critical (or 100% possible) at this stage, but rather representation of all useful combinations of frequencies. Parameters from the noise burst, F3, and VOT must be integrated with F2 transition information to eventually signal, at higher levels, categorical identity. Seen in this perspective, F2 onset versus F2 vowel representations are quite similar to ITD arrays in the inferior colliculus of the barn owl. At this early stage of processing, in both systems, ambiguity exists in the continuous representations of partial cues. In the barn owl the ITD arrays do not yet reflect auditory space and, similarly, locus equation representations do not unequivocally represent stop categories.

Both types of processing – continuous and categorical – are well-represented across a variety of mammalian auditory systems. What is most relevant is not whether processing is continuous or categorical but the auditory mechanisms used in the computations. In this sense, there are more commonalities than differences. **Wagner & Kautz** list several types of neural processing that are all dependent on coincidence detection of various inputs. Despite differing time scales, neuron morphologies, and molecular mechanisms "the algorithm is always the same." It is these basic formal correspondences that we call attention to and that **Remez, Fowler,** and **Mattingly** choose to denigrate.

**R2.4. One CV does not generate a locus equation.** A basic premise of the locus equation perspective is that orderliness is found at the level of the category, not the single, online token. While this helps one aspect of the coding problem – finding order where others often found disorder – it creates a processing dilemma, namely, how does the online input find its way to the orderly and presumably stored representations? This conundrum was mentioned by several commentators.

**Wagner & Kautz** point out a major difference between the barn owl's resolution of ambiguity in ITDs and humans resolving the ambiguity of vowel-context induced variability in the F2 transition. We have long been aware of the lack of co-temporality in the locus equation story (see Sussman 1989). This issue was also expressed, in one form or another, by **Fowler, Fitch & Hauser,** and **Smits.** Sussman et al. (1991) offered a "neural flow-chart" to conceptualize one possible solution. A multi-tier network of coincidence detectors was schematized. One tier responded to correspondences between burst information in relation to F2 onset frequency. A second tier processed online F2 onset in relation to a "predicted F2 onset." The extreme linearity of locus equations allows accurate prediction of the dependent variable, F2 onset, from the independent variable, F2 vowel. If the calculated and predicted F2 onsets "matched," the output signalled stop place information to higher centers.

Another scenario would entail neural population clusters for F2 onset-F2 vowel coordinates as typified by an exemplar-based model. A single CV stimulus would maximally activate a subset of neurons, and by virtue of their position in this neural space a (partial) signalling of stop place affiliation is effected. This scenario does not lead to perfect categorization (see **Brancazio, Smits, Massaro, Govindarajan**) as models using either locus equation lines as prototypes or exemplar labelled coordinates yield less than 100% accuracy in categorization. However, the success rate in all models is significantly well above chance (33%). The limitation to accurate categorization lies not in the locus equation algorithm per se, but in the fact that the models are not playing with a full deck – the relevant cues are multiple and need to be integrated (**Massaro, Jongman, Blumstein, Diehl, Nearey**).

A successful use of higher-order locus equation parameters, slope/intercept, to improve speech recognition performance was described by **Deng.** The key to this improvement (15% reduction of error rate) was constraining the hidden Markov model (HMM) to reduce the number of needed parameters. In current HMM-based speech recognition systems there are typically 50 million model parameters needed to handle the context-dependencies of speech (Deng, personal communication). Slope/intercept parameters are vowel-independent and consonant-specific. Deng's succinct parametrization of some context-dependencies of speech using locus equation regularities provides a modelling/statistical example of how higher-order category parameters can be used to process on-line CVs.

## R3. Alternative neural mechanisms for CV auditory processing

The commentaries by **Greenberg, Wenstrup,** and **Schreiner** bring up realistic concerns for a locus equation-type analysis. Greenberg points out the well-known difficulties of extracting phonemic elements from informal speech and suggests temporal, rather than spectral, cues for deriving phonetic identity. The amplitude of the low frequency ($<$25 Hz) modulation spectrum derived across frequency bands has had success in encoding natural speech. Although we do not doubt the important contributions of amplitude X time information carried in the speech signal, especially for hypoarticulated speech, this does not rule out using spectral information. The intelligibility of speech carried solely by temporal envelope information improves greatly as the number of frequency bands increases (Shannon et al. 1995). From personal experience (HMS) listening to the House Ear Institute demo tape containing primarily temporal cues, intelligibility was



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

"zero" with one- and two-band processors, and first became intelligible when the three-band processor was used. The four-band processor provided the most intelligible signal. This demonstrates, to us at least, the value of even "watered-down" spectral information to the overall identification process.

What should not be lost sight of in the debate over the primacy of temporal versus frequency information is that children learn language not by listening to input signals resembling the switchboard corpus (Godfrey et al. 1992) but rather classic "parentese." Fernald (1984) has shown infant-directed speech to be produced with higher pitch levels, extended intonational contours, and slower rates. Kuhl et al. (1997) have extended this to the point vowels /i/, /a/, and /u/. In a cross-language study comparing the acoustic make-up of infant-directed versus adult-directed speech, Kuhl and her coworkers found an expansion of acoustic vowel space in infant-directed speech. These hyperarticulated vowels were more distinctive, provided better exemplars for establishing phonetic categorization, and, by creating more variation within each vowel category, "it highlights the parameters on which speech categories are distinguished and by which speech can be imitated by the child" (p. 686). What is crucial in phonological neurogenesis is the state of the input signal when representations are initially being formed. When we attain a better understanding of how idealized speech signals are neurally encoded we will be in a better position to understand how underspecified transforms are processed.

**Wenstrup** mentions the relatively broad tuning properties of auditory neurons at intensity levels typical of human speech. This wide response area would preclude separate analyses of F2 onset and F2 vowel frequencies. **Schreiner** cites yet another problem as findings from studies presenting voiced stop consonants to cats and monkeys while recording from single neurons in primary auditory cortex show a single "phasic" or "onset" response at the beginning of the initial portion of the F2 transition but no second response that could be coding the end of the F2 transition or the vowel nucleus. These facts of neuronal activation patterns argue against orthogonal processing and representations for two separate frequency axes (see **Ohl & Scheich**). Alternative processing solutions are available, however, that are well within documented neuronal capabilities. One possibility would entail specialized speech-specific neurons operating in a fashion similar to FM-FM "delay-tuned" neurons (Olsen & Suga 1991b) encoding echo delays to signal target distances. In a personal communication **Wenstrup** wrote:

> For the type of analysis proposed . . . the neurons should exhibit a time-delayed response to the F2 onset. This permits the excitation evoked by F2 onset to coincide with F2 vowel-evoked excitation, and serves to prohibit responses to other frequencies in the F2 transition that would presumably code for other consonant-vowel combinations. The delay of F2 onset-excitation should be in the range of tens of milliseconds, since F2 onset typically precedes the F2 vowel by that interval. This is clearly within the capabilities of combination-sensitive neurons described in many species.

In fact, Ohlemiller et al. (1996) describe FM-FM neurons responding to social communication calls in the mustached bat with delays on the order of 50–75 msec, an interval very much similar to that between F2 onset and F2 vowel.

In addition to a specialized delay-tuned neuronal processor, other possibilities exist. Wenstrup mentions specialized combination-sensitive neurons in the inferior colliculus of the mustached bat that were responsive to two (non-sonar) frequency bands and for which each activating signal needed to be acoustically complex, for example, a narrow-band noise. Such selectivity could easily be adapted for human speech to relate noise burst information to F2 onsets. Other examples include subpopulations of specialized neurons that compute FM depth (**Kanwal**), thus encoding the same information as F2 onset and F2 vowel but along a single axis.

A more challenging coding dilemma concerns whether 2-D maps of orthogonally coded parameters are justified (**Moore & King, Kanwal, Schreiner**). Bat social calls comprise concatenated sound elements very similar to human speech, with harmonic structure, constant and frequency-modulated segments, as well as noise bursts (Kanwal et al. 1994). Unlike the cortical 2-D maps found for biosonar signal processing, neuronal analysis of bat communication calls does not suggest 2-D representations (Kanwal 1997). Kanwal's preliminary finding of "parameter-related cell clusters" (rather than a series of separate 2-D maps) provides suggestive evidence for a multi-dimensional coding for discrete, complex stimuli similar to human speech.

The problem of finding neural maps corresponding to acoustically motivated coordinates was the main focus of the commentary by **Ohl & Scheich.** Explorations of mammalian auditory cortices failed to reveal F1 × F2 representations for vowel categories similar to the familiar Peterson and Barney (1952) data. Ohl and Scheich (1997) had success in going from acoustic-to-neural space when a "spectral interactions" approach was adopted that transformed spectral peak frequencies to an auditory distance metric captured by the simple transform F2 minus F1. This suggests that neural encoding strategies for complex speech sounds may use transformed data rather than direct representations of sets of independent cues.

## R4. Is linearity necessary?

Several commentators (**Brancazio, Fowler, Guenther, Herrnberger & Ehret, Kanwal, Ohl & Scheich, Wagner & Kautz**) brought up the issue of whether or not linearity is a necessary prerequisite for establishing perceptually distinguishable categories. In terms of neurobiological necessity it is probably safe to conclude that linearity is, strictly speaking, not a prerequisite for establishing neural representations of input signals belonging to different equivalence classes. However, for learnability reasons (**Kluender**), we maintain that statistical regularities among cues are extremely beneficial and go well beyond "mathematical hocus-pocus" (**Wagner & Kautz**).

A linear relationship between stimulus variables creates a unique and tight coupling that stands out as a very clear and prominent signal amidst all the other nonlinear and noisy relationships in the signal. It is a stamp of uniqueness unlike any other. While one could argue that a sinusoidal relationship or any other curvilinear relationship could also be a stamp of uniqueness provided that it is consistently adhered to, linear relationships are information-theoretically minimal. That is, it takes the least amount of information to specify a linear relationship and to identify it. When a computational system is forced to establish some kind of a tight relationship between two perceptual variables to help



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

discrimination (and has freedom to, as in speech), it is only natural that by the principle of least effort the linear relationship is preferred. Linear relationships are simply the most efficient "marker" in a mathematical sense.

**R4.1. Articulatory origins of linearity.** An alternative account of locus equation linearity based on a theory of speech motor planning computationally implemented by his diva model is proposed by Guenther. This type of simulation promises to be very valuable in understanding the articulatory origins of locus equation linearity, but only if the underlying assumptions are realistic. These assumptions are that (1) there are invariant auditory target loci for the stop consonants, and (2) his parameter X/T determines the onset point of the F2 transition. The first assumption has its origin in motor control principles derived from arm movement studies – achievement of target position is conceptualized in terms of resting spring lengths. The movement planning process in diva is based on these notions and is described as a "virtual trajectory." The endpoint of a planned consonant-target auditory region is analogous to achieving a resting length for a spring, and, moreover, this endpoint target region can exist beyond a hard boundary of the vocal tract. When the tongue tip, for example, hits the alveolar ridge for a /d/ closure, movement stops but applied force does not. The amount of force is proportional to the distance yet to be attained to reach the consonant-target. The consonantal release movement, to the vowel-target region, initially involves the dissipation of this force, also realized as moving back along the "virtual trajectory," until it arrives at the alveolar ridge, where force now is zero. Movement to the vowel-target region then commences, and at some point along this trajectory, corresponding to a fixed percentage "X" of it, F2 onset is derived. We see several problems with this scenario. First, unlike the arm, the tongue has no joint and also, unlike arm muscles, which have parallel muscle fibers, intrinsic muscles of the tongue run in three different directions, transverse, vertical, and longitudinal (Sussman 1972). For the varied contour shapes achieved at closure there are a myriad of muscle resting lengths that would have to be simultaneously programmed, without even considering extrinsic tongue musculature. In the case of labial stops, a sphincter muscle system (orbicularis oris) + mandibular elevators + lower lip elevators + upper lip depressors must be coactivated, in a non-stereotyped, motor equivalent fashion, to achieve closure (Sussman et al. 1973). Setting targets equal to "resting lengths" for such a multidimensional system seems unrealistic, impractical, and extremely inelegant.

Even if "resting muscle length" is the parameter of choice in motor planning, it certainly is not the case that it is invariant across all vowel contexts for a given stop, as the tongue configuration at closure varies with the upcoming coarticulated vowel (Öhman 1966). Despite a different theoretical rationale and new terminology, the motor planning process in diva strongly resembles the classic, but disproven, "virtual locus" concept (Delattre et al. 1955). In **Guenther**'s model this invariant auditory target is tuned during development. Leaving aside the ontogenetic question of how the articulatory system gets tuned by a silent target during the closure interval, the main problem would be that a consonant like /b/, with one articulator for the consonant and another for the vowel, has no demonstrable locus, and the vowel may begin even during stop closure.

Another important assumption concerns the parameter X/T, which is held constant within a consonant (across vowels). "X" is a proportion of the "virtual trajectory" between a consonant and vowel's auditory targets, while T is the duration of the transition. Our first concern is that this parameter be interpreted properly for the purposes of a locus equation simulation. It is not clear to us how to interpret it if the consonant and vowel have different major articulators, but leaving aside this problem, X/T should be, for our purposes, the interval between stop closure and the onset of voicing, thus yielding the sample point for F2 onset. Instead, X/T was defined as the interval between maximum stop closure and stop *release,* so that voice onset time (VOT) was left out of the simulation. Whether X/T is constant within a stop and across vowels is also unclear, and we would like to see some empirical support for this assumption. **Guenther**'s "X" values (per stop) were selected to match locus equation slopes from our data. The unrealistic assumptions and circularity in specifying "X" makes the locus equations simulated by diva somewhat less than compelling.

**R4.2. Utilization of locus equations in perception.** An interesting query is put forth by **Massaro** concerning locus equations and perception, but he falls well short of providing an answer. In distinguishing between information that is potentially informative (ecological properties) and information that is actually used in perception (functional cues), he asks whether the emerging property, based on the correlation between F2 onset and F2 vowel, is being used by the perceptual system, or is it a simpler utilization of the two sources of information independently of each other. Massaro raises several points which are quite correct within the context of his concern but ignores the context upon which the target article is based. This is best illustrated in his discussion of input correlation. He states that "they believe that somehow component cues in the speech signal must be correlated to achieve categorization. . . . In fact, if there are two properties of the speech signal, best performance can be achieved when those properties are completely independent of one another." The confusion here is between the self-organization phase, that is, inference of categories from properties of the set of exemplars (the context within which we were speaking) and subsequent classification of exemplars given some model. Massaro views the issue of perception from the perspective of machine-based pattern recognition. Such a program of research has many degrees of freedom to achieve categorization optimality which are not available to humans. The OOC hypothesis is informed (and constrained) only by known models of animal neural processing and representation.

There is also some confusion about our stance regarding the perceptual role of the locus equation lines. Some commentators (**Brancazio, Govindarajan, Smits**) have assumed that we favor a model in which the regression lines are computed mentally and serve as prototypes of the stop place categories. We do not support such a model. In our view, the information-bearing locus equation parameters in each token serve as inputs that organize a map, such as the SOM Kohonen-style maps shown in Figure 18 of the target article. It is not entirely clear how organized SOMs, in a neurally realistic way, yield category identifications of unlabeled stimuli; however, until this problem is solved, we can envision a rough algorithm such as, for example, the



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

profile of the closest tokens associated with a set of the most active neural units in the map following a stimulus (Hilbert et al. 1994).

**Govindarajan** points out that the form of the self-organizing maps does not provide explicit evidence for locus equations. By this we presume he means that the maps do not manifest lines per se, as in the linear prototype models. This is exactly right. We do not claim that there are "lines in people's heads." Nevertheless, there are "lines" in the acoustic data, and our hypothesis to explain this is that there may be some important relationship between the ability of the maps to organize and the linearity in the set of inputs.

Both **Wagner & Kautz** and **Protopapas & Tallal** bring up the issue of ontogenetic experience (versus genetic programming). In our view, experiential exposure to spoken language during the first few years of life will, similarly to other sensory representations, lead to the formation of "phonological humunculi," formed from information-bearing dimensions of input sounds. This is one reason we favor the Kohonen map, which was designed to simulate topographic sensory maps such as the somatosensory cortex (Obermayer et al. 1990).

## R5. Coarticulation

Both **Fowler** and **Carré** support a "uniform coarticulatory resistance" hypothesis to account for the linearity of locus equations. In this view each stop consonant is thought to possess an inherent resistance to being affected by overlapping articulatory influences of the preceding or following vowel. Consonants are thought to resist coarticulatory effects of vowels "to the extent that the vowels interfere with achieving consonantal gestural goals" (Fowler). The only way uniformity in a consonant's coarticulatory resistance can be maintained is if each and every vowel exerts the same interference with the consonant's shape at closure. In examining the coarticulatory data from one of the key studies cited to support the notion of uniform coarticulatory resistance, Recasens (1984), it was noted that *vowel specific effects* were observed for all the Catalan obstruents studied (one approximant, two nasals, and a lateral). Recasens states that "carryover and anticipatory effects can be large or small depending on the quality of the transconsonantal vowel" (p. 72). In no case was the extent of either carryover or anticipatory coarticulation identical for /i/, /a/, or /u/ contexts. It is hard to imagine how the concept of uniform coarticulatory resistance was derived from the data of this study.

Regardless of the lack of explicit articulatory or acoustic support, there are two basic requirements for this concept to be meaningful. First, it must be defined in quantitative articulatory terms, and second, it needs to be shown why uniform coarticulatory resistance is good for the speaker/hearer. Neither **Fowler** nor **Carré** have provided any information to satisfy either requirement. Our view, as shown schematically in Figure 14 of the target article, holds that coarticulation is non-uniform across vowel contexts within a stop place category, but the vowel-context normalization of the F2 transition, as it is organized in a locus equation plot, achieves a characteristic level of the acoustic coding of coarticulation, per category, that is contrastive across place of articulation.

To test whether or not coarticulation is uniformly achieved across vowel contexts for a given stop place we utilized a physiologically motivated computational model of speech production, apex (Lindblom et al. 1997; Stark et al. 1996).[1] From formant data derived from productions of [d] (retroflex) in $V_1/d/V_2$ contexts (Krull et al. 1995) apex was instructed to provide the optimized articulatory parameters for the retroflex configuration that matched ±5% the acoustic formant values obtained from real speech tokens. The criterion of minimizing the extent of tongue tip elevation was used to limit the extensive range of possible articulatory configurations capable of producing the target formant values. The obtained parameters provided anterior-posterior position and constriction values in apex space for both the vowel and the coarticulated apical stop. The extent of articulatory movement (Euclidean distance) between a neutral tongue body configuration and that observed during coarticulation was calculated for each of six Swedish vowels produced with the retroflex stop /d/ and the results are shown in Figure R1. It can be clearly seen that the extent of movement varies as a function of vowel context. High front vowels /I:/ and /e:/ had the most extensive movement of the tongue body and low or mid back vowels /a:/ and /o:/ had the least extensive movement. These APEX-defined distances capture the extent of the vowel's influence on the subsequent configuration of the tongue body for the apical stop. If coarticulatory resistance were uniform within a stop consonant, as maintained by **Fowler** and **Carré,** one would expect to see uniform articulator excursions across vowels. Such was not the case.

**Carré**'s argument for a uniform coarticulation hypothesis to account for locus equation linearity is based on simulation data. Carré's DRM model captures coarticulation by adjusting the temporal phasing of the consonantal closure of the acoustic tube vis-à-vis opening for the vowel. If the consonantal closure occurs simultaneously with opening for the vowel, then coarticulation resistance is minimal and the vowel shape will maximally influence the consonant. Uniformity of coarticulation is the only outcome possible if the same temporal phasing is used across all vowel contexts. There is no principled way to vary the temporal onset of the vowel in the DRM model according

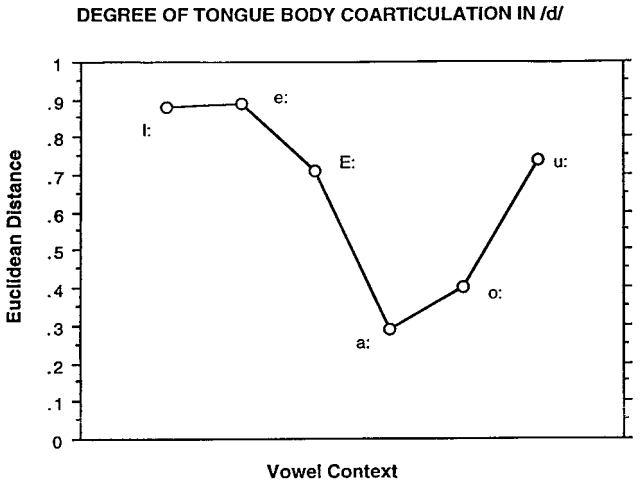



Figure R1. Degree of tongue body coarticulation measured in Euclidean distance for six Swedish vowels in the context of a retroflex [d]. Values were obtained from apex model simulations.



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

to what speakers actually do. Achievement of uniformity of coarticulation simply results from operator-based timing decisions – there is no other outcome possible.

## R6. Optimization of articulatory motor control and ontogeny of locus equations

**Fitch & Hauser** express puzzlement at the evidence offered to support our contention of a co-evolutionary adaptation of the human articulatory system to produce consonant + vowel utterances acoustically conforming to processing strategies favored by auditory processors. Fitch & Hauser are sympathetic to a preadaptation view but feel we did not prove our case. Their puzzlement is brought about by the following set of facts: (1) **Carré**'s DRM model reproduces linear plots and cannot be said to possess uniquely human motor control adaptations; (2) children with developmental apraxia of speech (DAS) have normal vocal tract shapes yet fail to produce linear locus equations; and (3) babbled CVs do not yield linear locus equations. Our babbling data are derived from infants seven months and older, and therefore their comment that infants under four months do not yet have the normal "two-tube" vocal tract (Lieberman et al. 1969) is irrelevant. They support a "non-uniqueness" view of the human vocal tract and claim that any and all mammalian "vocal tracts" would yield linear locus equations.

We believe that **Fitch & Hauser** may be confusing vocal tract shape/resonance properties with the dynamic properties of speech motor control. Our interpretation of these assorted facts is as follows. It has not been shown that nonhuman "vocal tracts" could yield linear "locus equation" functions, and it is doubtful that they would since the range of "F2 vowel" would be greatly restricted in a one-tube configuration (Lieberman et al. 1969). The DRM model, designed on the basis of acoustic resonance properties of human vocal tracts, can, assuming consistent places of constriction, successfully yield linear locus equation plots, but not with proper slope contrasts (reflecting different coarticulation levels) varying as a function of place of articulation. Speakers learn to exert the proper levels of coarticulation to tweak the distribution of CVs into unique functions that vary as a function of place. When the motor control system is not 100% functional, as in DAS, both linearity and slope distinctiveness suffer (likewise perceptual quality of the output) as scatterplots are characterized by large SEs.

Prelinguistic babbling is hypothesized to be generated by simple mandibular cyclicity – opening for the "vowel" and closing for the "consonant" (Davis & MacNeilage 1995). By 8–16 months of age the child's vocal tract is the normal two-tube configuration, but fine-tuned articulatory control and segmental independence is lacking and so is the signature linearity of locus equations. There are linear trends in the babbling data, but SEs are quite large (exceeding 300 Hz) compared to adult norms. So, there must be an articulatory maturation factor contributing to the locus equation story. We assume at this point that with normal maturation comes greater articulatory control over and precision of place of constriction and degree of coarticulation, and as these mature, so does the locus equation form.

A recent study examining syllable position effects has shown that final stops do not evidence the signature linearity of initial stops (Sussman et al. 1997). CVC words with initial and final stops /bdg/ produced with 10 medial vowel contexts were analyzed across 10 speakers to derive "offset" (VC) as well as "onset" (CV) locus equations (see **Idsardi**). Slope values for final stops were statistically less distinctive relative to initial stops. The mean SE for offset locus equations was almost double that obtained for traditional onset locus equations (CV = 144 Hz; VC = 252 Hz). Mean $R^2$ values were .84 for CV and .60 for VC locus equations. CV and VC entities appear to be phonetically fundamentally dissimilar with more articulatory (and hence acoustic) precision in the control of the F2 transition for CVs than for VCs. A simple vocal tract tube explanation for locus equation linearity (as supported by **Fitch & Hauser**), divorced from motor control factors, cannot explain these findings as closed-to-open (CV) and open-to-closed (VC) alterations of a tube should not, in principle, affect the resulting modulation of the resonance frequencies. The greater articulatory precision in the production of CVs is congruent with higher rates of initial relative to final consonant identifications (Ahmed & Agrawal 1969; Redford & Diehl 1996).

A convincing demonstration that articulatory factors play an important role in producing linear locus equations with unique slope/y-intercept characteristics (see **Lindblom**) has recently been given by Lindblom and his colleagues working with the apex articulatory model (Stark et al. 1996). The apex model differs from the DRM model (Carré & Mrayati 1992) in that it is physiologically motivated, based on analyses of X-ray images of real speakers. From input specifications for key articulatory dimensions (lip position, tongue tip elevation, tongue body shape, jaw elevation, larynx height) apex derives an articulatory profile, a computed area function from this profile, and an array of formant frequencies characterizing the acoustic output. For /dV/ syllables there were numerous articulatory configurations that could achieve the proper formant matches. The total possible locus space (for F2 onsets) across vowel contexts was found to be quite large, with nonlinear locus equations just as likely as linear locus equations. Yet speakers seem to utilize a restricted and linearly arranged portion of this possible acoustic space. As stated by Lindblom in his commentary: "There is nothing in the mapping from articulation to acoustics that makes locus equation linearity inevitable. Rather, both the phenomenon of linearity and the specific slope-intercept values reflect implicit 'choices' made by speakers and languages."

The capability of simulating different degrees of coarticulation in apex allows for a systematical examination of the role of motoric choices and the effect of such choices on the output signal. With respect to **Lindblom**'s elegant example of two levels of coarticulation for /dV/ productions – maximum tongue-body coarticulation with no constraints on tongue tip elevation versus a minimization of coarticulation combined with optimized constraints on tongue tip elevation for closure – two distinct, yet linear, locus equations were produced. The maximal coarticulation function had a slope of .94 and the minimized coarticulation condition a slope of .07. Real speakers produce /d/ locus equations with slopes near .40. They probably sacrifice some degree of coarticulation to minimize tongue tip excursion to the alveolar ridge.

One way to empirically study ontogenetic development is to track the acquisition of coarticulatory tuning and achievement of locus equation form in a child learning a language. A recent study addressed this developmental



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

issue (Sussman et al. 1997b). A longitudinal analysis of a single child, tape recorded at regular intervals from 7 to 40 months was performed.[2] Locus equations were derived for /bV/, /dV/, and /gV/ syllables, first from babbling, then early first words, and finally from conversational speech. In all, 3,153 /bV/, 3,040 /dV/, and 1,521 /gV/ syllables were acoustically analyzed. Signature locus equations had not yet been achieved by the age of 3;4 in this child when recording ceased. Mean SEs were >250 Hz throughout the entire 3-year period and had not reached adult norms at 40 months when recording ceased. In a related study, Minifie et al. (1997) tested the perceptual quality of babbling CVs with respect to their distance from locus equation functions. When the "best judged" and "worst judged" exemplars of each stop were grouped and plotted as locus equation scatterplots, the former yielded more tightly clustered scatterplots with more adult-like slopes relative to the CV coordinates based on more ambiguously perceived tokens.

The fact that locus equation scatterplots mature from a noisier form to an adult form with little noise is an argument for a perceptual constraint (versus simply an articulatory one). If the perceptual system were really flexible and could learn any kind of linear or non-linear relationships, there would be little need to train the articulatory system so finely. Instead, it could adapt itself during development to learn a broad variety of F2 onset–F2 vowel relationships. The fact that the articulatory system is the one that is being trained suggests that the perceptual system is the one that is less flexible and more constrained in this respect.

## R7. Locus equation stability

**Mattingly** states that locus equations are likely to be adversely affected by such phonetic factors as (1) vowel context from a preceding syllable, (2) degree of stress, or (3) coarticulatory influences. If locus equation stability is adversely affected by such factors, then their utility is limited. Along similar lines **Deng, Kluender,** and **Greenberg** ask what would happen if the locus equation paradigm were scaled up to fluent speech. Deng's use of informal speech contained in the timit data base showed reduced linearity in locus equations. In Sussman et al. (1997a) the issue of prior vowel contexts was empirically examined. Speakers produced VCV tokens, and locus equation slopes (for the CV portion) were compared in three different V1 environments: /i/, /ae/, and /u/. Slopes for /bdg/ were very stable across the three V1 contexts and showed only minimal alterations – .04 (/b/), .04 (/d/), and .056 (/g/). These small changes do not affect relative separability of categories when plotted in slope/y-intercept coordinates. In terms of "scaling up" to fluent connected speech, we are in the midst of a large-scale study to determine how locus equations are affected by alterations in speaking style. Previous studies have shown that locus equation slopes are somewhat steeper for spontaneous relative to citation-style speech (e.g., Krull 1989). At the moment we have analyzed five speakers, three male and two female. Citation-style locus equations were derived from clearly articulated CVC stimuli read from lists. Locus equations, from the same speakers, are also derived from spontaneous speech. The slope differences between the two speaking styles (averaged across speakers) are /b/ = .096, /d/ = .109, and /g/ = .095. Figure R2 shows the relative separability, in locus equation-defined space, of the stop place categories for the two speaking styles. While these results are only preliminary, it appears that modulation of degree of coarticulation caused by adoption of different speaking styles does not significantly alter locus equations.

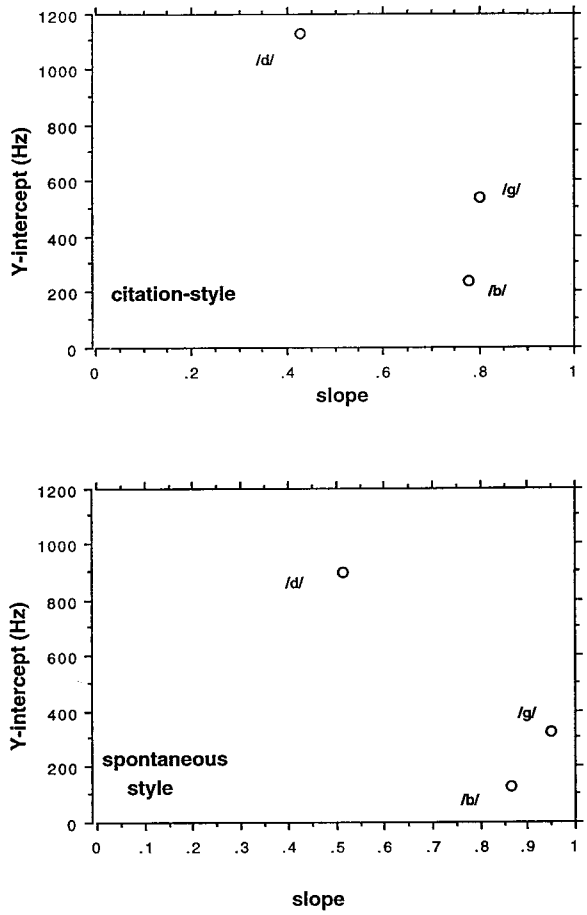



Figure R2. Comparison of slope/y-intercept coordinates for /bdg/ locus equations derived from citation-style (top) and spontaneous (bottom) speaking styles.

## R8. Phonetic concerns

**Kluender** questioned whether locus equation parameters would be successful in contrasting stop place in languages containing more than three places of articulation. Sussman et al. (1993) studied four stop place contrasts in Urdu and Cairene Arabic. Cairene Arabic contains a dental [d] contrasted with a pharyngealized [d], and Urdu contrasts a dental [d] and retroflex [d]. Where slope values were fairly similar (as in Arabic: dental [d] slope = .25; pharyngealized [d] slope = .21), the y-intercepts were quite distinctive (1307 Hz and 933 Hz, respectively). Slopes/y-intercepts for the Urdu coronals were .50/857 Hz and .44/1070 Hz for the dental and retroflex stops, respectively. When both stop contrasts from a given speaker were plotted together in locus equation space, a clear separability was maintained throughout all vowel contexts, brought about by differences in F2 onsets.

**Jongman** tested locus equations as phonetic descriptors



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

of place of articulation in fricatives. Slopes for labiodental (.768), dental (.53), alveolar (.517), and palato-alveolar (.505) fricatives failed to show a systematic change as place was varied. Our response to Jongman's data is that locus equations were not originally derived to characterize "unencoded" obstruents such as fricatives. Fricatives are continuant sounds that lack the dynamic and transient nature of stops and, hence, do not provide the paradigmatic case of noninvariance. Most fricatives can be transposed (via tape-splicing) across words without destroying the identity of the fricative (e.g., splicing the [z] from "zap" onto an "ip" and hearing "zip"). This cannot be done with stops.

**Idsardi** faults locus equations because they fail to abstractly capture a single English /g/, but rather reflect the phonetic allophones colored by vowel place features. A single regression function could be fit to /g/ tokens (described as a "Procrustean" fit by Lindblom), but it was felt that two linear fits were a more accurate way to describe this problematic phoneme. Clearly, /g/ presents a unique coding problem and we do not claim to have the answer. Both allophonic representations of /g/ must be integrated at a higher level of analysis, most likely aided by top-down processes at the lexical level.

**Pastore & Crawley** argue for a stringent analysis of the effect of taking F2 onset measurements at various positions before locus equations can be confidently utilized as a laboratory research tool to investigate consonantal place, especially as different manner classes create unique problems with respect to sampling the F2 resonance. We welcome such suggestions.

**Govindarajan** comments that there is considerable cross-speaker variability in locus equations. In our perception data (Fruchter & Sussman 1997), there is a large range of tolerable variation within categories that matches the range of variation in production. This variability, however, is well-quantified. The speech perception mechanism does not necessarily need a lack of variability, but sufficient discriminability between categories.

The integration of phonetic cues to form a percept was mentioned by many commentators (**Blumstein, Diehl, Jongman, Massaro, Pind, Sawusch**). None of the models used to test locus equations to date (**Brancazio, Fowler, Smits**) has advanced to the point where they can incorporate other cues; they are consequently unrealistic in a crucial sense. Integrating sensory data channels (the "binding problem") is a very general problem in cognitive neuroscience. We envision binding in speech perception in a way that is conceptually similar to the integration of azimuth and elevation signals in the barn owl for localization – a hierarchical integration of different feature maps coding independent aspects of the signal.

ACKNOWLEDGMENTS

We would like to thank the following individuals for their helpful comments during preparation of this response: Li Deng, Günter Ehret, Frank Guenther, Jag Kanwal, Björn Lindblom, Jim Talley, and Jeff Wenstrup.

NOTES

**1.** We thank Björn Lindblom for performing the apex simulations and providing the Euclidean distance measures for Figure R1.

**2.** Audiotapes of this child were generously provided to us by Peter MacNeilage and Barbara Davis.

## References

**Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.**

Ahmed, R. & Agrawal, S. S. (1969) Significant features in the perception of (Hindi) consonants. *Journal of the Acoustical Society of America* 45:758–63. [rHMS]

Amerman, J. D. (1970) A cinefluorographic investigation of the coarticulatory behavior of the apex and body lingual articulators. Ph.D. dissertation, University of Illinois. [CAF, aHMS]

Arai, T. & Greenberg, S. (1997) The temporal properties of spoken Japanese are similar to those of English. *Proceedings of Eurospeech* 2:1011–1114. [SG]

Ashby, F. G. & Perrin, N. A. (1988) Toward a unified theory of similarity and recognition. *Psychological Review* 95:124–50. [RS]

Benasich, A. A. & Tallal, P. (1996) Auditory temporal processing thresholds, habituation, and recognition memory over the first year. *Infant Behavior and Development* 19:339–57. [AP]

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S. E. & Mehler, J. (1987) Discrimination in neonates of very short CVs. *Journal of the Acoustical Society of America* 82:31–37. [SEB]

Bladon, R. A. W. & Lindblom, B. (1981) Modeling the judgement of vowel quality differences. *Journal of the Acoustical Society of America* 69:1414–22. [FWO]

Blumstein, S. E. & Stevens, K. N. (1979) Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America* 66:1001–17. [aHMS]

(1980) Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America* 67:648–62. [SEB, AJ]

Bishop, D. V. M. (1992) The underlying nature of specific language impairment. *Journal of Child Psychology and Psychiatry* 33:3–66. [AP]

Bouabana, S. & Maeda, S. (in press) Multipulse LPC modeling of articulatory movements. *Speech Communication*. [SG]

Brainard, M. S., Knudsen, E. I. & Esterly, S. D. (1992) Neural derivation of sound source location: Resolution of spatial ambiguities in binaural cues. *Journal of the Acoustical Society of America* 91:1015–27. [HW]

Brancazio, L. & Fowler, C. A. (1998) On the relevance of locus equations for production and perception of stop consonants. *Perception and Psychophysics* 60:24–50. [LB, CAF]

Carré, R. & Chennoukh, S. (1995) Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gesture. *Journal of Phonetics* 23:231–41. [RC]

Carré, R. & Mody, M. (1997) Predictions of vowel and consonant place of articulation. In: *Proceeding of the Third Meeting of the ACL Special Interest Group in Computational Phonology, SIGPHON 97, (Madrid)*. [RC]

Carré, R. & Mrayati, M. (1992) Distinctive regions in acoustic tubes. Speech production modeling. *Journal d'Acoustique* 5:141–59. [BL, aHMS]

Casseday, J. H., Ehrlich, R. & Covey, E. (1994) Neural tuning for sound duration: Role of inhibitory mechanisms in the inferior colliculus. *Science* 264:847–50. [JSK]

Catford, J. C. (1988) *A practical introduction to phonetics*. Clarendon. [RER]

Celdran, E.M. & Villalba, X. (1995) Locus equations as a metric for place of articulation in automatic speech recognition. *Proceedings of the XIIIth International Congress of Phonetic Sciences (Sweden)* 1:30–33. [aHMS]

Chennoukh, S., Carré, R. & Lindblom, B. (1997) Locus equations in the light of articulatory modeling. *Journal of the Acoustical Society of America* 102:2380–89. [RC, CAF]

Churchland, P. S. & Sejnowski, T. (1989) Neural representation and neural computation. In: *Neural connections, mental computations*, ed. L. Nadel, L. A. Cooper, P. Culicover & R. M. Hornish. MIT Press. [aHMS]

Cohen, Y. E. & Knudsen, E. I. (1995) Binaural tuning of auditory units in the forebrain archistriatal gaze fields of the barn owl: Local organization but no space map. *Journal of Neuroscience* 15:5152–68. [JSK]

Creutzfeldt, O., Hellweg, F. C. & Schreiner, C. E. (1980) Thalamocortical transformation of responses to complex auditory stimuli. *Experimental Brain Research* 39:87–104. [CES]

Damper, R. I., Harnad, S. & Gore, M. O. (1997) A computational model of the perception of voicing in initial stops. *Journal of the Acoustical Society of America*. (submitted). [RID]

Davis, B. L. & MacNeilage, P. F. (1995) The articulatory basis of babbling. *Journal of Speech and Hearing Research* 38:1199–1211. [rHMS]

Delattre, P. C., Liberman, A. M. & Cooper, F. S. (1955) Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America* 27:769–73. [arHMS]



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

Deng, L. (1994) A statistical model for formant-transition microsegments of speech incorporating locus equations. *Signal Processing* 37(1):121–28. [LD]

Deng, L. & Braam, D. (1994) Context-dependent Markov model structured by locus equations: Applications to phonetic classification. *Journal of the Acoustical Society of America* 96(4):2008–25. [LD]

Deng, L., Kenny, P., Lennig, M. & Mermelstein, P. (1992) Modeling acoustic transitions in speech by state-interpolation hidden Markov models. *IEEE Transactions on Signal Processing* 40(2):265–72. [LD]

Dianora, A., Hemphill, R., Hirata, Y. & Olson, K. (1996) Effects of context and speaking rate on liquid-stops sequences: A reassessment of traditional acoustic cues. *Journal of the Acoustical Society of America* 100:2601. [KRK]

Diehl, R. L. (1981) Feature detectors for speech: A critical reappraisal. *Psychological Bulletin* 89:1–18. [RLD]

Diehl, R. L. & Kluender, K. R. (1987) On the categorization of speech sounds. In: *Categorical perception*, ed. S. Harnad. Cambridge University Press. [RLD]

Doherty, J. & Hoy, R. (1985) Communication in insects III. The auditory behavior of crickets: Some views of genetic coupling, song recognition, and predator detection. *Quarterly Review of Biology* 60:453–72. [MJR]

Dorman, M. F. & Loizou, P. C. (1997) Relative spectral change and formant transitions as cues to labial and alveolar place of articulation. *Journal of the Acoustical Society of America* 100:3825–30. [JRS]

Dudley, H. (1939) Remaking speech. *Journal of the Acoustical Society of America* 11:169–77. [SG]

Eggermont, J. J. (1995) Representation of voice onset time continuum in primary auditory cortex of the cat. *Journal of the Acoustical Society of America* 98:911–20. [CES]

Ehret, G. (1992) Preadaptations in the auditory system of mammals for phoneme recognition. In: *The auditory processing of speech: From sounds to words*, ed. M. E. H. Schouten. Mouton de Gruyter. [BH, rHMS]

Ehret, G. & Haack, B. (1981) Categorical perception of mouse-pup ultrasounds by lactating females. *Naturwissenschaften* 68:208. [rHMS]

(1982) Ultrasound recognition in house mice: Key-stimulus configuration and recognition mechanism. *Journal of Comparative Physiology A* 148:245–51. [rHMS]

Farmer, M. E. & Klein, R. M. (1995) The evidence for a temporal processing deficit linked to dyslexia: A review. *Psychonomic Bulletin and Review* 2:460–93. [AP]

Farnetani, E. (1990) V-C-V lingual coarticulation and its spatiotemporal domain. In: *Speech production and speech modeling*, ed. W. J. Hardcastle & A. Marchal. Kluwer. [CAF]

Fernald, A. (1984) The perceptual and affective salience of mothers' speech to infants. In: *The origins and growth of communication*, ed. L. Feagans, C. Garvey & R. Golinkoff. Ablex. [rHMS]

Fitch, W. T. (1997) Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America* 102(2):1213–22. [WTF, AP]

Fitzpatrick, D. C., Kanwal, J. S., Butman, J. A. & Suga, N. (1993) Combination-sensitive neurons in the primary auditory cortex of the mustached bat. *Journal of Neuroscience* 13:931–40. [aHMS, JJW]

Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R. N. (1988) Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America* 84:115–24. [AJ, JRS]

Fowler, C. A. (1994) Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics* 55:597–610. [RC, CAF, aHMS]

Fruchter, D. (1994) Perceptual significance of locus equations. *Journal of the Acoustical Society of America* 95:2977. [LB, CAF, aHMS]

Fruchter, D. & Sussman, H. M. (1997) The perceptual relevance of locus equations. *Journal of the Acoustical Society of America* 102:2997–3008. [rHMS]

Fuzessery, Z. M. & Feng, A. S. (1983) Mating call selectivity in the thalamus and midbrain of the leopard frog (*Rana p. pipiens*): Single and multi-unit analyses. *Journal of Comparative Physiology* 150:333–44. [aHMS]

Gaese, B. H. & Ostwald, J. (1995) Temporal coding of amplitude and frequency modulation in the rat auditory cortex. *European Journal of Neuroscience* 7:438–50. [CES]

Gay, T., Lindblom, B. & Lubker, J. (1981) Production of bite-block vowels: Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America* 69:802–10. [BL]

Gibson, J. J. (1966) *The senses considered as perceptual systems.* Houghton-Mifflin. [JP]

Godfrey, J. J., Holliman, E. C. & McDaniel, J. (1992) switchboard: Telephone speech corpus for research and development. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-92)* 1:517–20. [SG, rHMS]

Greenberg, S. (1997) On the origins of speech intelligibility in the real world. In: *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels* 23–32. [SG]

Greenberg, S., Hollenback, J. & Ellis, D. (1996) Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In: *Proceedings of the Fourth International Conference on Spoken Language, Philadelphia (ICSLP)*, S24–27. [SG]

Greenberg, S. & Kingsbury, B. (1997) The modulation spectrogram: In pursuit of an invariant representation of speech. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, Munich (ICASSP-97)* 1647–50. [SG]

Greenberg, S. & Shire, M. (1997) Temporal factors in speech perception. In: *CSRE-based teaching modules for courses in speech and hearing sciences*. AVAAZ Innovations. [SG]

Guenther, F. H. (1995) Speech sound acquisition, coarticulation, and speaking rate effects in a neural network model of speech production. *Psychological Review* 102:594–621. [FHG]

Guenther, F. H., Hampson, M. & Johnson, D. (1997) A theoretical investigation of reference frames for the planning of speech movements. Boston University Technical Report CAS/CNS-97–002. *Psychological Review*. [FHG]

Halle, M. (1991) Phonological features. In: *Oxford international encyclopaedia of linguistics*, ed. W. Bright. Oxford University Press. [WJI]

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech* 1:1–7. [AJ]

Hauser, M. D. (1996) *The evolution of communication*. MIT Press. [WTF]

Hauser, M. D., Evans, C. S. & Marler, P. (1993) The role of articulation in the production of rhesus monkey (*Macaca mulatta*) vocalizations. *Animal Behaviour* 45:423–33. [WTF]

Hauser, M. D. & Schön-Ybarra, M. (1994) The role of lip configuration in monkey vocalizations: Experiments using xylocaine as a nerve block. *Brain and Language* 46:232–44. [WTF]

Hebb, D. O. (1949) *The organization of behavior*. Wiley. [aHMS]

Hedrick, M. S. & Ohde, R. N. (1993) Effect of relative amplitude of frication on perception of place of articulation. *Journal of the Acoustical Society of America* 94:2005–27. [AJ]

Heil, P., Rajan, R. & Irvine, D. R. F. (1992) Sensitivity of neurons in cat primary auditory cortex to tones and frequency-modulated stimuli: I. Effects of variation of stimulus parameters. *Hearing Research* 63:108–34. [CES]

Heinz, J. M. & Stevens, K. N. (1961) On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America* 33:589–96. [AJ]

Hilbert, J., Fruchter, D., McWilliams, M., Sirosh, J. & Sussman, H. M. (1994) Self-organizing maps of stop consonant place from token-level locus equation inputs. Paper presented at semi-annual meeting of the Acoustical Society of America, Austin, Texas. [rHMS]

Hinton, G. E. & Lang, K. J. (1988) The development of the time-delay neural network architecture for speech recognition. *Technical Report Carnegie-Mellon University,CMU-CS* 88–152. [aHMS]

Hockett, C. D. (1960) The origin of speech. *Scientific American* 203:88–96. [KRK]

Hoffstetter, K. M. & Ehret, G. (1992) The auditory cortex of the mouse: Connections of the ultrasonic field. *Journal of Comparative Neurology* 323:370–86. [aHMS]

Hura, S. L., Lindblom, B. & Diehl, R. (1992) On the role of perception in shaping phonological assimilation rules. *Language and Speech* 35(1,2):59–72. [BL]

Jacob, F. (1977) Evolution and tinkering. *Science* 196:1161–66. [rHMS]

Jakobson, R., Fant, G. & Halle, M. (1963) *Preliminaries to speech analysis*. MIT Press. [SEB]

Jongman, A. (1989) Duration of fricative noise required for identification of English fricatives. *Journal of the Acoustical Society of America* 85:1718–25. [AJ]

Jongman, A. & Sereno, J. A. (1995) Acoustic properties of non-sibilant fricatives. *Proceedings of the XIIIth International Congress of Phonetic Sciences* 432–35. [AJ]

Jusczyk, P. W. (1997) *The discovery of spoken language*. MIT Press. [AP]

Kanwal, J. S. (1997) A multidimensional code for processing social calls in the primary auditory cortex of the mustached bat. *Proceedings of the 33rd International Congress of Physiological Sciences*, Abstract No. L081.05. [DRM, rHMS]

Kanwal, J. S., Matsumura, S., Ohlemiller, K. & Suga, N. (1994) Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *Journal of the Acoustical Society of America* 96:1229–54. [arHMS, JJW]

Kewley-Port, D. (1982) Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America* 72:379–89. [FHG, aHMS]

(1983) Time-varying features as correlates of place of articulation in stop



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

consonants. *Journal of the Acoustical Society of America* 73:322–35. [aHMS]
Kingsbury, B., Morgan, N. & Greenberg, S. (1997) Improving ASR performance for reverberant speech. In: *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, 87–90. [SG]
Kluender, K. R. & Diehl, R. L. (1987) Use of multiple speech dimensions in concept formation by Japanese quail. *Journal of the Acoustical Society of America* 82:S84. [KRK]
Kluender, K. R., Diehl, R. D. & Killeen, P. R. (1987) Japanese quail can form phonetic categories. *Science* 237:1195–97. [KRK]
Kluender, K. R., Lotto, A. J., Holt, L. L. & Bloedel, S. L. (1997) Role of experience in language-specific functional mappings for vowel sounds. (submitted). [KRK]
Knudsen, E. I. & Konishi, M. (1978) A neural map of auditory space in the barn owl. *Science* 200:795–97. [JSK]
Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43:59–69. [aHMS]
(1990) The self-organizing map. *Proceedings of the Institute of Electrical and Electronics Engineers* 78:1464–80. [RID, aHMS]
Konishi, M. (1994) An outline of recent advances in birdsong neurobiology. *Brain Behavior and Evolution* 44:279–85. [MJR]
Konishi, M., Takahashi, T., Wagner, H., Sullivan, W. E. & Carr, C. E. (1988) Neurophysiological and anatomical substrates of sound localization in the owl. In: *Auditory function*, ed. G. M. Edelman, W. E. Gall & W. M. Cowan. Wiley. [aHMS]
Krull, D. (1988) Acoustic properties as predictors of perceptual responses: A study of Swedish voiced stops. *Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS)* VII:66–70. [aHMS]
(1989) Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *Phonetic Experimental Research at the Institute of Lingusitics, University of Stockholm (PERILUS)* X: 87–101. [arHMS]
(1990) Relating acoustic properties to perceptual responses: A study of Swedish voiced stops. *Journal of the Acoustical Society of America* 88:2557–70. [LB]
Krull, D., Lindblom, B., Shia, B.-E. & Fruchter, D. (1995) Cross-linguistic aspects of coarticulation: An acoustic and electropalatographic study of dental and retro-flex consonants. In: *Proceedings of the International Congress of Phonetic Sciences, Stockholm* 3:436–39. [rHMS]
Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U. & Lacerda, F. (1997) Cross-language analysis of phonetic units in language addressed to infants. *Science* 277:684–86. [SEB, rHMS]
Kugel, K., Leishman, L. I., Bahr, R. H. & Montgomery, A. (1995) Procedural influences on the measurement of locus equations. Paper presented at the annual meeting of the American Speech-Language-Hearing Association, Orlando, Florida, December 7–10. [aHMS]
Lahiri, A., Gewirth, L. & Blumstein, S. E. (1984) A reconsideration of acoustic invariance in stop consonants: Evidence from cross-language studies. *Journal of the Acoustical Society of America* 76:391–404. [JRS, aHMS]
Leroy, S. A. & Wenstrup, J. J. (1996) Combination-sensitive neurons in the inferior colliculus of the mustached bat: Possible analysis of social communication signals. *Society for Neuroscience Abstracts* 22:404. [JJW]
Liberman, A. (1996) *Speech: A special code*. MIT Press. [RID, CAF]
Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review* 74:431–61. [SEB, RLD, JP, aHMS]
Liberman, A. M., Delattre, P. C., Cooper, F. S. & Gerstman, L. J. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs* 68:1–13. [LB, WTF, aHMS]
Liberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. (1957) The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54:358–68. [IGM]
Liberman, A. M. & Mattingly, I. (1985) The motor theory of speech perception revised. *Cognition* 21:1–36. [aHMS]
(1989) A specialization for speech perception. *Science* 243:489–94. [RID]
Lieberman, P. (1975) *On the origins of language*. Macmillan. [SEB]
(1984) *The biology and evolution of language*. Harvard University Press. [WTF, aHMS]
Lieberman, P., Klatt, D. H. & Wilson, W. H. (1969) Vocal tract limitations on the vowel repertoires of rhesus monkeys and other nonhuman primates. *Science* 164:1185–87. [WTF, rHMS]
Lindblom, B. (1963a) On vowel reduction. Report No. 29, Speech Transmission Laboratory, The Royal Institute of Technology, Sweden. [RID, aHMS]
(1963b) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35:1773–81. [IGM]
(1983) Economy of speech gestures. In: *The production of speech*, ed. P. F. MacNeilage. Springer-Verlag. [BL, aHMS]
(1986) Explaining phonetic variation. A sketch of the H & H theory. In: *Speech production and speech modeling*, ed. W. Hardcastle & A. Marchal. Klüwer. [JP]
Lindblom, B., Stark, J. & Sundberg, J. (1997) From sound to vocal gesture: Learning to (co)-articulate with APEX. In: *Fonetik-97, Phonum 37–40*, Umeå Universitet. [BL, rHMS]
Lotto, A. J. & Kluender, K. R. (in press) General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics*. [KRK]
Lotto, A. J., Kluender, K. R. & Holt, L. L. (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix japonica*). *Journal of the Acoustical Society of America* 102:1134–40. [KRK]
Maddieson, I. (1984) *Patterns of sound*. Cambridge University Press. [KRK]
Margoliash, D. (1983) Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *Journal of Neuroscience* 3:1039–57. [aHMS]
Margoliash, D. & Fortune, E. S. (1992) Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. *Journal of Neuroscience* 12:4309–26. [aHMS]
Massaro, D. W. (1998) *Perceiving talking faces: From speech perception to a behavioral principle*. MIT Press. [DWM]
Mattingly, I. G. & Liberman, A. M. (1988) Specialized perceiving systems for speech and other biologically significant sounds. In: *Auditory function: Neurobiological bases of hearing*, ed. G. M. Edelman, W. E. Gall & W. M. Cowan. Wiley. [IGM]
May, B., Moody, D. B. & Stebbins, W. C. (1989) Categorical perception of nonspecific communication sounds by Japanese macaques, *Macaca fuscata*. *Journal of the Acoustical Society of America* 85:837–47. [rHMS]
McDermott, E. & Katagiri, S. (1988) Phoneme recognition using Kohonen's Learning Vector Quantization. *ATR Workshop on Neural Networks and Parallel Distributed Processing, Japan*. [aHMS]
Mendelson, J. R., Schreiner, C. E., Sutter, M. L. & Grasse, K. L. (1993) Functional topography of cat primary auditory cortex: Responses to frequency-modulated sweeps. *Experimental Brain Research* 94:65–87. [JSK, CES]
Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L. & Tallal, P. (1996) Temporal processing deficits of language-learning impaired children ameliorated by training. *Science* 271:77–81. [AP]
Minifie, F. D., Sussman, H. M., Hall, S. & Stoel-Gammon, C. (submitted) Assessing the perceptual relevance of locus equations in early infant speech. [rHMS]
Mittman, D. H. & Wenstrup, J. J. (1995) Combination-sensitive neurons in the inferior colliculus. *Hearing Research* 90:185–91. [aHMS, JJW]
Moon, S.-J. & Lindblom, B. (1994) Interaction between duration, context and speaking style in English stressed vowels. *Journal of the Acoustical Society of America* 96(1):40–55. [BL]
Mrayati, M., Carré, R. & Guérin, B. (1988) Distinctive region and modes: A new theory of speech production. *Speech Communication* 7:257–86. [RC]
Mudry, K. M., Constantine-Paton, M. & Capranica, R. R. (1977) Auditory sensitivity of the diencephalon of the leopard frog, *Rana p. pipiens. Journal of Comparative Physiology* 114:1–13. [aHMS]
Nakayama, K. (1994) James J. Gibson – An appreciation. *Psychological Review* 101:329–35. [JP]
Nearey, T. (1997) Speech perception as pattern recognition. *Journal of the Acoustical Society of America* 101:3241–54. [TMN]
Nearey, T. M. & Shammass, S. E. (1987) Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics* 15:17–24. [TMN, aHMS]
Neuweiler, G. (1983) Echolocation and adaptivity to ecological constraints. In: *Neurorethology and behavioral physiology: Roots and growing pains*, ed. F. Huber & H. Markl. Springer-Verlag. [aHMS]
(1984) Foraging, echolocation and audition in bats. *Naturwissenshaften* 71:446–55. [aHMS]
Nossair, Z. B. & Zahorian, S. A. (1991) Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America* 89:2978–90. [TMN]
Nosofsky, R. M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115:39–57. [RS]
Obermayer, K., Blasdel, G. G. & Schulten, K. J. (1992) Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps. *Physical Review A* 45:7568–89. [aHMS]
Obermayer, K., Ritter, H. J. & Schulten, K. J. (1990) Large-scale simulation of a self-organizing neural network: Formation of a somatotopic map. *Proceedings of the International Conference on Parallel Processing in Neural Systems and Computers (ICNC), Düsseldorf*. Elsevier. [rHMS]
(1991) Development and spatial structure of cortical feature maps: A model study. In: *Advances in neural information processing systems*, ed. R. P. Lippmann, J. E. Moody & D. S. Touretzky. Morgan-Kaufman. [aHMS]



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

Ohl, F. W. & Scheich, H. (1997) Orderly cortical representation of vowels based on formant interaction. *Proceedings of the National Academy of Sciences USA* 94:9440–44. [FWO, rHMS]

Ohlemiller, K., Kanwal, J. S., Butman, J. A. & Suga, N. (1994) Stimulus design for auditory neuroethology: Synthesis and manipulation of complex communication sounds. *Auditory Neuroscience* 1:19–37. [aHMS]

Ohlemiller, K., Kanwal, J. S. & Suga, N. (1996) Facilitative responses to species-specific calls in cortical FM-FM neurons of the mustached bat. *NeuroReport* 7:1749–55. [CES, rHMS, JJW]

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America* 39(1):151–68. [RC, BL, IGM. rHMS]

Oller, D. K. (1978) Infant vocalizations and the development of speech. *Allied Health and Behavioral Sciences* 1:523–49. [aHMS]

Olsen, J. F. (1994) Medial geniculate neurons in the squirrel monkey sensitive to inter-component delays that categorize species-typical calls. *Abstracts of the Association for Research in Otolaryngology* 17:21. [aHMS]

Olsen, J. F. & Rauschecker, J. P. (1992) Medial geniculate neurons in the squirrel monkey sensitive to combinations of components in a species-specific vocalization. *Society of Neuroscience Abstracts* 18:883. [aHMS]

Olsen, J. F. & Suga, N. (1991a) Combination-sensitive neurons in the medial geniculate body of the mustached bat: Encoding of relative velocity information. *Journal of Neurophysiology* 65:1254–73. [arHMS]

(1991b) Combination-sensitive neurons in the medial geniculate body of the mustached bat: Encoding of target range information. *Journal of Neurophysiology* 65:1275–96. [arHMS]

Pastore, R. E., Liberto, J. W. & Crawley, E. J. (1997) Mapping multidimensional perceptual consonant spaces for place contrasts. *Journal of the Acoustical Society of America* 100:2694 (Abstract). [REP]

Perkell, J. & Klatt, D. H. (1986) *Invariance and variability in speech processes*. Erlbaum. [aHMS]

Perkell, J. S., Matthies, M. L., Lane, H., Guenther, F. H., Wilhelms-Tricarico, R., Wozniak, J. & Guiod, P. (in press) Speech motor control: Acoustic segmental goals, saturation effects, and the use of auditory feedback. *Speech Communication*. [FHG]

Peterson, G. E. & Barney, H. L. (1952) Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24:175–84. [FWO, rHMS]

Pind, J. (1986) The perception of quantity in Icelandic. *Phonetica* 43:116–39. [JP]

(1995) Speaking rate, VOT and quantity: The search for higher-order invariants for two Icelandic speech cues. *Perception and Psychophysics* 57:291–304. [JP]

Pollak, G. D., Winer, J. A. & O'Neill, W. E. (1995) Perspectives on the functional organization of the mammalian auditory system: Why bats are good models. In: *Springer handbook of auditory research*, vol. II, ed. A. N. Popper & R. R. Fay. Springer-Verlag. [aHMS]

Price, C., Wise, R., Ramsay, S., Friston, K., Howard, D., Patterson, K. & Frackowiak, R. (1992) Regional response differences within the human auditory cortex when listening to words. *Neuroscience Letters* 146:179–82. [DRM]

Rauschecker, J. P., Tian, B. & Hauser, M. D. (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268:111–14. [AP, CES]

Recasens, D. (1984) V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study. *Journal of Phonetics* 12:61–73. [CAF, rHMS]

(1989) Long range coarticulatory effects for tongue dorsum contact in VCVCV sequences. *Speech Communication* 8:293–307. [CAF]

Redford, M. A. & Diehl, R. L. (1996) A study on the relative perceptibility of syllable-initial and syllable-final consonants. *Journal of the Acoustical Society of America* 100:2693(A). [rHMS]

Remez, R. E., Fellowes, J. M., Pisoni, D. B., Goh, W. D. & Rubin, P. E. (in press) Audio-visual speech perception without speech cues: A second report. In: *Proceedings of the ESCA workshop on audio visual speech processing: Cognitive and computational approaches*, ed. C. Benoit & R. Campbell. European Speech Communication Association. [RER]

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S. & Lang, J. M. (1994) On the perceptual organization of speech. *Psychological Review* 101:129–56. [RER]

Richardson, K. H. (1992) An analysis of invariance in English stop consonants. Paper presented at the 123rd meeting of the Acoustical Society of America, Salt Lake City, Utah. [JRS]

Ritter, H. J. (1990) Self-organizing maps for internal representations. *Psychological Research* 52:128–36. [aHMS]

Rock, I. (1970) Perception from the standpoint. In: *Perception and its disorders*, ed. D. A. Hamburg. Williams & Wilkins. [RER]

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) Learning representations by back-propagating errors. *Nature* 323:533–36. [RID]

Ryan, M. J. (1997) Sexual selection and mate choice. In: *Behavioural ecology, an evolutionary approach, 4th edition*. Blackwell. [MJR]

Ryan, M. J., Fox, J. H., Wilczynski, W. & Rand, A. S. (1990) Sexual selection for sensory exploitation in the frog, *Physalaemus pustulosus. Nature* 343:66–67. [WTF]

Savariaux, C., Perrier, P. & Orliaguet, J. P. (1995) Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *Journal of the Acoustical Society of America* 98:2428–42. [FHG]

Sawusch, J. R. (1986) Auditory and phonetic coding of speech. In: *Pattern recognition by humans and machines: Volume 1.*, ed. E. C. Schwab & H. C. Nusbaum. Academic Press. [JRS]

Schreiner, C. E. & Calhoun, B. M. (1994) Spectral envelope coding in cat primary auditory cortex: Properties of ripple transfer functions. *Auditory Neuroscience* 1:39–61. [FWO, CES]

Schreiner, C. E. & Urbas, J. V. (1986) Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Research* 21:227–41. [SG]

Schreiner, C. E. & Wong, S. W. (1996) Spectral-temporal representation of syllables in cat primary auditory cortex. In: *Proceedings of the ESCA tutorial and advanced research workshop on the auditory basis of speech perception*, ed. W. Ainsworth & S. Greenberg. Keele. [CES]

Schreiner, C. E., Wong, S. W. & Bonham, B. (1997) Spectral-temporal representation of syllables in cat primary auditory cortex. In: *Psychophysical and physiological advances in hearing*, ed. A. R. Palmer, A. Rees, A. Q. Summerfield & R. Meddis. Grantham. Whurr Publishers. [CES]

Shamma, S. A., Fleshman, J. W., Wiser, P. R. & Versnel, H. (1993) Organization of response areas in ferret primary auditory cortex. *Journal of Neurophysiology* 69:367–83. [AP]

Shamma, S. A., Versnel, H. & Kowalski, N. (1995) Ripple analysis in ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra. *Auditory Neuroscience* 1:233–54. [FWO, CES]

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. (1995) Speech recognition with primarily temporal cues. *Science* 270:303–04. [rHMS]

Shepard, R. N. (1958) Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology* 55:509–23. [RS]

Simpson, G. G. (1961) *Principles of animal taxonomy*. Columbia University Press. [aHMS]

Smith, C., Browman, C., McGowan, R. & Kay, B. (1993) Extracting dynamic parameters from speech movement data. *Journal of the Acoustical Society of America* 93:1580–88. [SG]

Stark, J., Lindblom, B. & Sundberg, J. (1996) APEX – an articulatory synthesis model for experimental and computational studies of speech production. In: *Fonetik-96: Papers presented at the Swedish Phonetics Conference, Stockholm*. TNH-QPSR 2:45–48. [BL, rHMS]

Stebbins, G. L. (1974) *Flowering plants: Evolution above the species level.* Belknap Press. [aHMS]

Steinschneider, M., Arezzo, J. & Vaughan, H. G., Jr. (1982) Speech evoked activity in the auditory radiations and cortex of the awake monkey. *Brain Research* 252:353–65. [CES]

Steinschneider, M., Schroeder, C. E., Arezzo, J. C. & Vaughan, H. G., Jr. (1994) Speech-evoked activity in primary auditory cortex: Effects of voice onset time. *Journal of Electroencephalography and Clinical Neurophysiology* 92:30–43. [CES]

Stevens, K. N. (1989) On the quantal nature of speech. *Journal of Phonetics* 17:3–45. [SEB, JP]

Stevens, K. N. & Blumstein, S. E. (1978) Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America* 64:1358–68. [SEB, aHMS]

(1981) The search for invariant acoustic correlates of phonetic features. In: *Perspectives on the study of speech*, ed. P. D. Eimas & J. L. Miller. Erlbaum. [SEB, REP]

Suga, N. (1964) Recovery cycles and responses to frequency modulated tone pulses in auditory neurons of echolocating bats. *Journal of Physiology* 175:50–80. [JSK]

(1973) Feature extraction in the auditory system of bats. In: *Basic mechanisms in hearing*, ed. A. R. Moller. Academic Press. [JSK]

(1988) Neuroethology, audition, and speech. In: *Auditory function*, ed. G. M. Edelman, W. E. Gall & W. M. Cowan. Wiley. [aHMS]

(1994) Multi-function theory for cortical processing of auditory information: Implications for single unit and lesion data for future research. *Journal of Comparative Physiology* (A) 175:135–44. [aHMS]

Suga, N. & Jen, P. H.-S. (1976) Disproportionate tonotopic representation for



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press

processing species-specific CF-FM sonar signals in the mustached bat auditory cortex. *Science* 194:542–44. [aHMS]

Suga, N., O'Neill, W. E., Kujirai, K. & Manabe, T. (1983) Specificity of combination-sensitive neurons for processing of complex biosonar signals in auditory cortex of the mustached bat. *Neurophysiology* 49:1573–1627. [aHMS, JJW]

Suga, N., O'Neill, W. E. & Manabe, T. (1978) Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the mustached bat. *Science* 200:778–81. [aHMS]

Sullivan, W. E. & Konishi, M. (1986) Neural map of interaural phase difference in the owl's brainstem. *Proceedings of the National Academy of Sciences USA* 83:8400–04. [aHMS]

Sussman, H. M. (1972) What the tongue tells the brain. *Psychological Bulletin* 77:262–72. [rHMS]

(1986) A neuronal model of vowel normalization and representation. *Brain and Language* 28:12–23. [aHMS]

(1988) The neurogenesis of phonology. In: *Phonological processes and brain mechanisms*, ed. H. Whitaker. Springer-Verlag. [aHMS]

(1989) Neural coding of relational invariance in speech: Human language analogs to the barn owl. *Psychological Review* 96:631–42. [arHMS]

(1994) The phonological reality of locus equations across manner class distinctions: Preliminary observations. *Phonetica* 51:119–31. [aHMS]

Sussman, H. M., Bessell, N., Dalston, E. & Majors, T. (1997a) An investigation of stop place of articulation as a function of syllable position: A locus equation perspective. *Journal of the Acoustical Society of America* 101:2826–38. [KRK, rHMS]

Sussman, H. M., Dalston, E., Duder, C. & Cacciatore, A. (1997b) An acoustic analysis of the development of CV coarticulation: A case study. *Journal of Child Language*. (submitted). [rHMS]

Sussman, H. M., Fruchter, D. & Cable, A. (1995) Locus equations derived from compensatory articulation. *Journal of the Acoustical Society of America* 97:3112–24. [aHMS]

Sussman, H. M., Hoemeke, K. & Ahmed, F. (1993) A cross-linguistic investigation of locus equations as a relationally invariant descriptor for place of articulation. *Journal of the Acoustical Society of America* 94:1256–68. [WJI, arHMS]

Sussman, H. M., Hoemeke, K. & McCaffrey, H. A. (1992) Locus equations as an index of coarticulation for place of articulation distinctions in children. *Journal of Speech and Hearing Research* 35:769–81. [aHMS]

Sussman, H. M., MacNeilage, P. F. & Hanson, R. J. (1973) Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *Journal of Speech and Hearing Research* 16:397–420. [arHMS]

Sussman, H. M., McCaffrey, H. A. & Matthews, S. A. (1991) An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America* 90:1309–25. [WJI, AJ, arHMS]

Sussman, H. M., Minifie, F. D., Buder, E. H., Stoel-Gammon & Smith, J. (1996) Consonant-vowel interdependencies in babbling and early words: Preliminary examination of a locus equation approach. *Journal of Speech and Hearing Research* 39:424–33. [aHMS]

Sussman, H. M. & Shore, J. (1996) Locus equations as phonetic descriptors of consonantal place of articulation. *Perception and Psychophysics* 58:936–46. [REP, aHMS]

Sutter, M. L. & Schreiner, C. E. (1991) Physiology and topography of neurons with multipeaked tuning curves in cat primary auditory cortex. *Journal of Neurophysiology* 65:1207–26. [aHMS]

Takahashi, T. & Konishi, M.(1986) Selectivity for interaural time difference in the owl's midbrain. *Journal of Neuroscience* 6:3413–22. [aHMS]

Tallal, P. & Piercy, M. (1973) Developmental aphasia: Impaired rate of non-verbal processing as a function of sensory modality. *Neuropsychologia* 11:389–98. [AP]

(1974) Developmental aphasia: Rate of auditory processing and selective impairment of consonant perception. *Neuropsychologia* 12:83–93. [AP]

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., Schreiner, C., Jenkins, W. M. & Merzenich, M. M. (1996) Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science* 271:81–84. [AP]

Tian, B. & Rauschecker, J. (1994) Processing of frequency-modulated sounds in the cat's anterior auditory field. *Journal of Neurophysiology* 71:1959–75. [CES]

Unnikrishnan, K. P., Hopfield, J. J. & Tank, D. W. (1988) Learning time-delayed connections in a speech recognition circuit. Neural Networks for Computing Conference, Utah. [aHMS]

Wagner, H., Takahashi, T. & Konishi, M. (1987) Representation of interaural time difference in the central nucleus of the barn owl's inferior colliculus. *Journal of Neuroscience* 7:3105–16. [aHMS, HW]

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K. (1989) Phoneme recognition using time-delay neural networks. *Proceedings of the Institutute of Electrical and Electronics Engineers (IEEE) Transactions on Acoustics and Signal Processing* 37:328–339. [RID, aHMS]

Wang, X. Q., Merzenich, M. M., Beitel, R. & Schreiner, C. E. (1995) Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: Temporal and spectral characteristics. *Journal of Neurophysiology* 74:2685–706. [CES]

Watrous, R. L. (1988) Speech recognition using connectionist networks. Ph.D. dissertation, University of Pennsylvania. [aHMS]

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, C. M. & Merzenich, M. M. (1997) Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature* 387:176–78. [AP]

Yan, J. & Suga, N. (1996) The midbrain creates and the thalamus sharpens echo-delay tuning for the cortical representation of target-distance information in the mustached bat. *Hearing Research* 93:102–10. [JJW]

Zatorre, R. J., Evans, A. C., Meyer, E. & Gjedde, A. (1992) Lateralization of phonetic and pitch discrimination in speech processing. *Science* 256:846–49. [DRM]



https://doi.org/10.1017/S0140525X9851117X Published online by Cambridge University Press