


Brownian dynamics for the vowel sounds of human language

J. Burridge ^{*}*School of Mathematics and Physics, University of Portsmouth, Portsmouth PO1 3HF, United Kingdom*B. Vaux *Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, Cambridge CB3 9DA, United Kingdom*

(Received 5 July 2019; accepted 4 February 2020; published 6 March 2020)

We present a model for the evolution of vowel sounds in human languages, in which words behave as Brownian particles diffusing in acoustic space, interacting via the vowel sounds they contain. Interaction forces, derived from a simple model of the language-learning process, are attractive at short range and repulsive at long range. This generates sets of acoustic clusters, each representing a distinct sound, which form patterns with similar statistical properties to real vowel systems. Our formulation may be generalized to account for spontaneous self-actuating shifts in system structure which are observed in real languages, and to combine in one model two previously distinct theories of vowel system structure: dispersion theory, which assumes that vowel systems maximize contrasts between sounds, and quantal theory, according to which nonlinear relationships between articulatory and acoustic parameters are the source of patterns in sound inventories. By formulating the dynamics of vowel sounds using interparticle forces, we also provide a simple unified description of the linguistic notion of *push* and *pull* dynamics in vowel systems.

DOI: [10.1103/PhysRevResearch.2.013274](https://doi.org/10.1103/PhysRevResearch.2.013274)

I. INTRODUCTION

Each human language has its own inventory of sounds. Within these inventories a distinction is made between consonants, which require some restriction of airflow for their production, and vowels, resonant sounds, for which airflow is relatively unrestricted [1]. Vowels are the primary carriers of linguistic information in connected human speech [2,3]. The sound of a vowel is determined largely by the position and configuration of the tongue, although other parts of the vocal apparatus can be involved [4]. Linguists have traditionally represented vowels as points in a two-dimensional *articulatory* domain (the *vowel quadrilateral*) with coordinates given by tongue *height* and *backness* [5]. To produce the vowel sound in ⟨cat⟩, represented phonetically as [æ], the tongue is in a low, forward position. By contrast, the vowel in ⟨foot⟩, pronounced [ʊ], is articulated with the tongue dorsum in a relatively high and backed position. Vowels may also be reliably identified from the first two peaks, or *formants* (F_1, F_2), of their frequency spectrum [6] (Fig. 1). Models of vowel production [7] and experiments [8,9] suggest that F_1 strongly correlates to tongue height, and F_2 to backness, although recent work suggests that F_2 may depend on both [9]. Mathematically, there appears to be a bijective map from

articulatory to acoustic (F_1, F_2) space which is approximately affine [10], with some exceptional regions [11]. For this reason, we will view vowel sounds as existing in a closed two-dimensional domain: *vowel space*.

Vowel systems exhibit recurring patterns and regularity: the majority of languages (64.6%) have between five and seven different qualities of vowel with /i/ /a/, and /u/ occurring in over 80% of languages [12]. Moreover, certain arrangements within vowel space are particularly common [12–14]. Vowel systems, like most elements of languages, evolve over time and may therefore be viewed as dynamical systems coupled to human social dynamics, and also to geography and social networks [15–17]. Cross-linguistic similarities suggest that their internal dynamics may play a particularly powerful role, and numerous models have been proposed [13,14,18–25]. In the early work of Liljencrants and Lindblom [13], vowels were modeled as electrical charges, based on the principle of *maximal contrast* [26], yielding a single idealized vowel system for each cardinality (number of vowels). *Focalization* theory, which adds an attractive interaction to vowel dynamics, hypothesizes the convergence of formants [18,27], increasing acoustic salience and “perceptual value.” Other models involve iterative construction driven by contrast maximization [20], or agent-based imitation games between speech synthesizers [14,21]. One inspiration for our work is *exemplar dynamics* [22–25,28,29], where agents store in memory a large number of exemplars for each sound, organized into sound categories. When uttering a sound, an exemplar is reproduced (with noise) from memory and accepted as valid by listeners based on how easily it can be identified and, in some cases, how typical it is of its category [24]. When two categories get close, they may overlap, making identification

*james.burridge@port.ac.uk

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

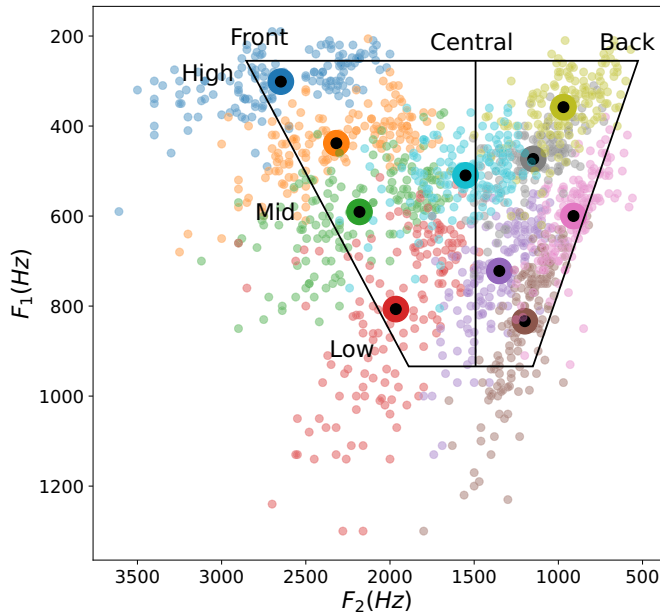


FIG. 1. Formant data for the vowel sounds of 76 speakers of American English from the original study by Peterson and Barney [6], where speakers were recorded reading the words *heed, hid, head, had, hod, hawed, hood, who'd, hud, and heard*. Axes have been reversed so that the positions of vowels correspond to those in the IPA vowel chart [5]. The approximate shape of this chart has been superimposed on the formant data, to illustrate the correspondence between mean formant positions (large points), and the arrangement of the traditional chart. Front vowels: blue *i*, orange *ɪ*, green *ɛ*, red *æ*. Back vowels: light green *u*, gray *ʊ*, pink *ɔ*, purple *ʌ*, brown *ɑ*. Central vowel: turquoise *ɜ*.

of utterances in the intersecting regions more difficult, leading to their rejection as exemplars. This shifts the mean sounds of the two categories apart, as if subject to a repulsive force. Exemplar models have recently been analyzed [23,30], with the aim of determining conditions for the merger of sound categories, and the behavior of the boundaries between them in vowel space. Repulsive mechanisms are common to all the models above and generate distributions of sounds which are dispersed in vowel space. They are known as *dispersion theories* [12,31].

While there is clearly value in defining models with greater linguistic realism (using realistic synthesizer equations [21], or metrics of *perceptual distance* [19]), this approach also makes the determination of their general behavior difficult. It is hard to evaluate whether the complexity they add in order to better match reality is scientifically justified, or whether it is a form of interpolation. Exemplar theory is a step away from this approach, and may be defined as a simple iterative computational model. However, rigorous analysis is challenging [30], and because the atomic constituents of the theory are tokens in the memories of speakers, simulations of large numbers of different sounds and words are also computationally expensive. The theory has been used to explore interactions between words and phonemes [23,24], but it has not so far been used to model the evolution of realistic vowel systems.

The model that we present builds on the ideas described in the above models, but we aim for an analytical definition

which is simpler to simulate and analyze. Sound change in language may be thought of as the diffusion of word pronunciations in acoustic space, making Brownian (Langevin) dynamics [32,33] the natural mathematical description of their motion. We think of vowel systems as a “soup” of words or, more technically, *phonological frames*. These frames interact via the vowel sounds that they contain, and the interaction forces may be seen either as a phenomenological model [34], based on established qualitative models of sound change [35], or as a simplified version of exemplar dynamics. Interactions between frames are mediated by a cloud of utterances, in our case formant clouds, but unlike exemplar theory, we do not explicitly model this collection of sounds, only the mean sound for each frame. This dramatically simplifies the model, yielding one stochastic differential equation per frame, facilitating analytical calculations, and allowing the simulation of a large number of words. Our formulation is analogous to a physical model (frames form a charged colloid [36–39]) allowing us to clarify traditional qualitative descriptions of change in terms of *pushing* and *pulling*. Its simplicity also allows a number of extensions to describe a range of different phenomena in one unified model. These include self-actuating sound change [40–42], allophonic sound variations, the effect of nonlinearities in the relationship between articulatory and acoustic parameters [11,43], and word-frequency effects which have recently been observed in a purely computational exemplar model [24].

II. A BRIEF INTRODUCTION TO PHONOLOGY

For the benefit of nonlinguists, we now review the relevant elements of *phonology*: the branch of linguistics that deals with systems of sounds.

A. Phonemes and allophones

Individual speech sounds are usefully viewed as existing within phonological frames [24]. For instance, in English the frame /m_p/, if it forms a single word, will accept one of two sound categories, creating either ⟨map⟩ or ⟨mop⟩. Similar variants appear in other words, with subtle variations depending on the frame. Because of these variations, a distinction is made between the contrastive sound categories, *phonemes*, and their context dependent versions, termed *allophones*. For example, ⟨pea⟩, ⟨spin⟩, and ⟨sip⟩ all contain what English speakers might call a p sound, but these three sounds are all slightly different. The transcriptions of these words into phonetic symbols, which represent specific sounds, are [p^hiː], [spɪn], and [sɪp[˞]]. The three p sounds here are, respectively, *aspirated* (followed by a burst of breath), *unaspirated*, and *unreleased* (meaning that there is no audible end to the temporary occlusion of airflow needed to make the sound). These three variants are allophones of the English phoneme /p/:

$$/p/ = \{[p^h], [p], [p^{\text{˞}}]\}. \tag{1}$$

Phonemes differ between languages: Thai speakers, for example, consider [p^h] and [p] to be manifestations of distinct phonemes /p^h/ and /p/, respectively. Allophones may be grouped into phonemes by examining the sounds that surround them (their phonological environment). Consider the frame /kæ_/ where /æ/ is the vowel phoneme in the word

⟨cat⟩. Inserting any of the allophones of /p/ would produce an utterance recognized by English speakers as the word ⟨cap⟩. However, inserting the sound [b] produces an utterance with a different meaning. We say that [b] is in *contrastive distribution* with [p]; if we switch the sounds in the frame, we change the meaning. Allophones of the same phoneme are generally not in contrastive distribution; they are (normally) in *complementary distribution*. This means that they are not found in the same immediate phonological environment. For example, [p^ɾ] can only occur in syllable-final consonant clusters and [p^h] only occurs either word-initially or at the beginning of a stressed syllable. Therefore, they never contrast, and given a phoneme we can predict which allophone will appear simply by knowing what other sounds surround it. The definition of a phoneme and its allophones applies identically to consonants and vowels. For example, an English vowel phoneme is /æ/, which occurs in ⟨bad⟩ and ⟨bat⟩. Here the allophone in ⟨bad⟩ is longer, transcribed [æː]. Another characteristic which distinguishes vowel allophones is *nasalization* as in ⟨ban⟩[bæ̃n].

Since the allophones of a phoneme typically exhibit quite subtle variations, the term *phoneme* is often used as if it referred to a single sound in the language. Doing so is no less consistent than using *allophone* in the same way. In reality, the units of sound uttered by speakers vary widely between and within individuals. Both phonemes and allophones are categories, with one being a subcategory of the other. We will exploit this idea when we come to explore allophonic variations in Sec. VII.

B. Vowel systems

As explained above, the first two formants, or equivalently the position of the tongue, are the primary determiners of vowel sounds, but other articulatory variations can be involved as well. After height and backness, the next most common is lip rounding [12]. There is, however, a strong cross-linguistic correlation between rounding and backness (94.0% of front vowels are unrounded, and 93.5% of back vowels are rounded), making the parameter redundant in most cases. Beyond lip rounding, many languages have separate series of vowels, each distinguished by some additional characteristic such as length (long vs short) or nasalization (nasalized vs oral). Often the sounds in both series occupy the same positions in vowel space. For example, in Mazatec [44] there are four oral vowels /i, e, a, o/ and four corresponding nasalized versions /ĩ, ě, ã, õ/. In this case, we say that Mazatec has four vowel *qualities*. This matching between series is the norm, so we can ignore additional characteristics and still provide a description of vowel systems which captures the essential properties of their structure for most languages. This will be our approach, until we consider allophonic variation in Sec. VII.

III. MODEL

Our model may be derived based on general assumptions about the language-learning process, and how this is affected by the *overcrowding* of sounds in acoustic space. It may also be derived as an approximation to an explicit model of

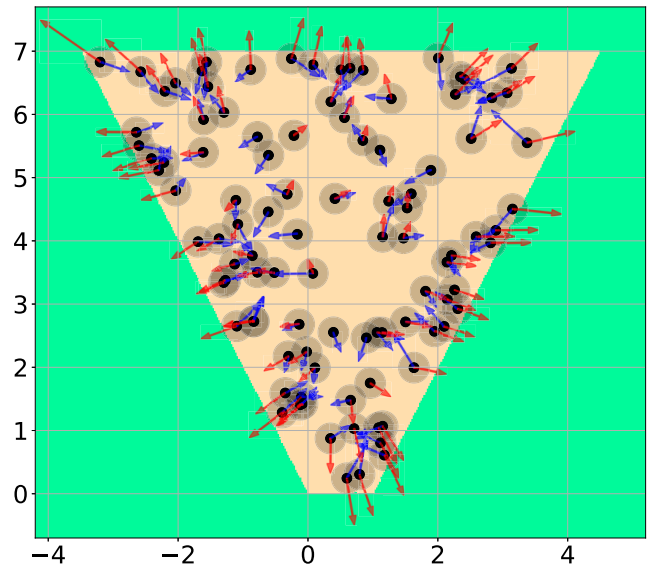


FIG. 2. Early stage configuration of 100 particles (words) in a chamber representing acoustic (F_1, F_2) vowel space. Range of attractive force $\alpha = 1$, cloud radius $\sigma = 1$, diffusion coefficient $D = 0.1$. Blue arrows show attractive forces, red show repulsive.

language learning in which utterances are either accepted or rejected as valid examples of a given sound. In this section we provide the first, nonexplicit, derivation.

We view words as particles in acoustic space, with positions determined by the vowel sounds they contain. A cluster of words then implies the existence of a frequently occurring sound in a language. We propose short-range attractive and long-range repulsive forces between words based on ideas from linguistics about the interactions between phonemes [13,17,45]. Before setting out the details, we describe how such forces can lead to the formation of a vowel system. Consider a large number of words distributed uniformly at random throughout a chamber representing vowel space as in Fig. 2. At first, particles near to each other will be drawn together, forming loose clusters whose typical size will be determined by the radius at which the attractive force becomes repulsive, as in Fig. 3. As time progresses, provided the volatility of the random component of their motion is not too large, these clusters will become tighter and more separated due to long-range repulsive effects. Depending upon the shape of the chamber, and the range of the short- and long-range components of the interaction force, we will obtain a number of different arrangements of particle clusters within the system, each of which represents a different inventory of sounds.

We now set out the details. For words with only one vowel, their position is unambiguous, but words with two or more vowels occupy multiple positions. We view such words as generating one or more phonological frames, or environments, for the vowels they contain. For example,

$$\langle \text{hoodwink} \rangle \rightarrow /h_dwɪŋk/ + /hʊdw_ɪŋk/. \quad (2)$$

Each such frame now has a unique position in the acoustic domain, determined by the sound which is placed in its gap.

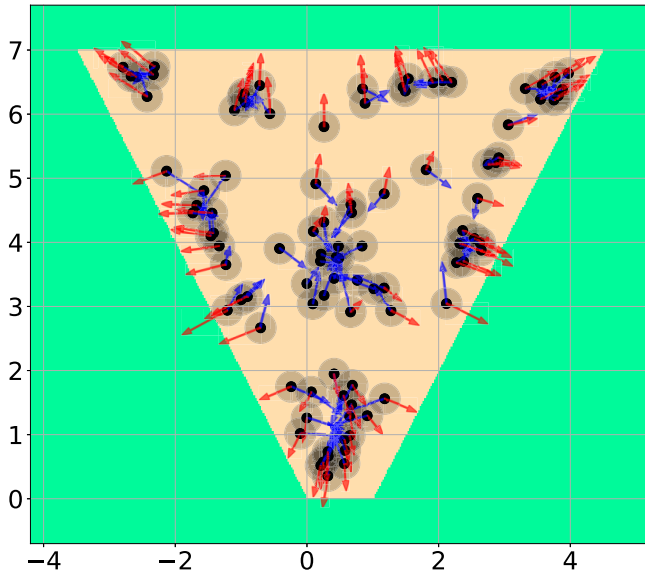


FIG. 3. Later stage configuration ($t = 0.5$) of 100 particles (words) in a chamber representing acoustic (F_1, F_2) vowel space. Range of attractive force $\alpha = 1$, cloud radius $\sigma = 1$, diffusion coefficient $D = 0.1$.

For simplicity, we assume that two frames from the same word interact in the same way as frames from different words.

We now consider a large group of speakers who are sufficiently socially connected to pronounce the words of their language in a roughly similar way. Acoustic experiments show that vowel sounds are subject to variation. Peterson and Barney [6] collected formant data for a group of 76 American speakers pronouncing a series of 10 words which differed only in their vowel, allowing F_1/F_2 values to be collected for 10 vowels. Because these utterances vary between speakers, the formant values form clouds in acoustic space. In Fig. 1 the large colored dots show the centroids of each vowel phoneme cloud. We have superimposed the standard IPA vowel quadrilateral [5] on the scatter plot, and we note that to a good approximation the centroids sit in the chart positions assigned by linguists over half a century earlier [46].

We let $\mathbf{x}_i(t) \in \mathbb{R}^2$ be the population average vowel sound uttered for frame i , and we refer to $\mathbf{x}_i(t)$ as the position of this frame. Because utterances of the vowel sound in each word form a cloud in acoustic space [6], the next utterance to be heard from frame i will be a random variable $\mathbf{X}_i(t)$. The speaker who produces this utterance must arrange her vocal apparatus into an appropriate configuration to create the desired sound. We denote by \mathbf{g} the function which maps points in articulatory space (mouth cavity shape, etc.) to points in acoustic space as illustrated in Fig. 4. To generate the desired sound, the speaker must have an intuitive knowledge of the inverse mapping \mathbf{g}^{-1} so that given the target sound \mathbf{x} she is able to form her mouth parts into the appropriate configuration $\mathbf{y} = \mathbf{g}^{-1}(\mathbf{x})$. We assume, for now, that on average the articulatory states \mathbf{Y}_i used by the population to generate sounds \mathbf{X}_i produce unbiased results in the sense that the average values of outputs over all speakers and utterances are equal to the frame position

$$\mathbb{E}[\mathbf{g}(\mathbf{Y}_i(t))] = \mathbb{E}[\mathbf{X}_i(t)] = \mathbf{x}_i(t). \quad (3)$$

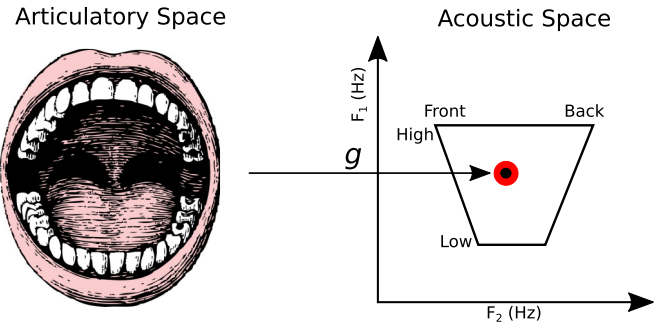


FIG. 4. The articulatory to acoustic map \mathbf{g} . In order to utter a target sound in acoustic space, a speaker must arrange her articulatory apparatus (tongue, lips, laryngeal structures, etc.) into the correct position \mathbf{y} . The uttered sound will be $\mathbf{x} = \mathbf{g}(\mathbf{y})$.

We also assume that articulatory states used to generate \mathbf{x}_i are normally distributed, having mean $\mathbf{g}^{-1}(\mathbf{x}_i)$. In this case, if the map \mathbf{g} is affine

$$\mathbf{x} = \mathbf{g}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{c}, \quad (4)$$

where $\mathbf{A} \in \mathbb{R}^{2 \times 2}$, then the utterances \mathbf{X}_i will also be normally distributed, having density function

$$\psi_i(\mathbf{x}) = \frac{1}{2\pi|\Sigma|} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T \Sigma^{-1}(\mathbf{x} - \mathbf{x}_i)\right). \quad (5)$$

Here, the covariance matrix Σ determines the shape of the cloud of utterances [6,47]. In the spherical case we set $\Sigma = \sigma \mathbf{I}$ where \mathbf{I} is the identity matrix, and σ is the *cloud radius*. Strictly speaking, we could have taken (5) as the definition of the shape of the utterance cloud in acoustic space, without needing to consider the map \mathbf{g} . However, experiments show that there are regions of acoustic space where \mathbf{g} is not affine [11] and, as we will see in Sec. VI, this affects both cloud shape and system dynamics.

We define the state of the vowel system of our language to be the set of all frame positions $\{\mathbf{x}_i(t)\}_{i=1}^n$, where n is the number of frames. Interactions between these frames are generated by the language-learning process. When a speaker learns how to pronounce the word which generated frame i , not only will utterances of the word itself provide templates for its vowel sound, but so will similar sounds in other words. If two frames contain vowel sounds which can be used as templates for one another, that is, speakers consider them to contain *the same sound*, we write

$$\mathbf{x}_i \stackrel{T}{=} \mathbf{x}_j, \quad (6)$$

where $\stackrel{T}{=}$ denotes *template equality*. We write S_i for the set of frames whose vowel sounds act as templates for frame i ,

$$S_i = \{j \text{ such that } \mathbf{x}_j \stackrel{T}{=} \mathbf{x}_i\}, \quad (7)$$

where time dependence is implicit. In this paper we use proximity in acoustic space to define template equality

$$\mathbf{x}_i \stackrel{T}{=} \mathbf{x}_j \equiv |\mathbf{x}_i - \mathbf{x}_j| < \alpha, \quad (8)$$

where $\alpha \geq 0$ is the *template range*. The set of templates for each frame will evolve over time, as frames change position. The overall density of utterances which may be used as

templates for learning the vowel sound in frame i is then defined

$$\widehat{\psi}_i(\mathbf{x}) := \sum_{j \in S_i} (1 + \omega \delta_{ij}) f_j \psi_j(\mathbf{x}), \tag{9}$$

where f_j is the relative frequency with which frame j is uttered, and $\omega \geq 0$ (the *self-focus*) is the extra weight placed on utterances of the frame as a template for its own vowel sound. We call $\widehat{\psi}_i(\mathbf{x})$ the *template density* for frame i . When we have a set of frames for which every frame is a template for every other, then we call this set a phoneme. Given the template density for a frame i we can compute the *template mean* for that frame

$$\hat{\mathbf{x}}_i = \frac{\int_{\mathbb{R}^2} \widehat{\psi}_i(\mathbf{x}) \mathbf{x} d\mathbf{x}}{\int_{\mathbb{R}^2} \widehat{\psi}_i(\mathbf{x}) d\mathbf{x}} \tag{10}$$

$$= \frac{1}{\mathcal{N}_i} \int_{\mathbb{R}^2} \widehat{\psi}_i(\mathbf{x}) \mathbf{x} d\mathbf{x} \tag{11}$$

$$= \frac{\sum_{j \in S_i} (1 + \omega \delta_{ij}) f_j \mathbf{x}_j}{\sum_{j \in S_i} (1 + \omega \delta_{ij}) f_j}, \tag{12}$$

where \mathcal{N}_i is the normalizing constant for the template density. The template mean $\hat{\mathbf{x}}_i$ is the mean value, weighted for self-focus, of all the utterances from the frames which a language learner uses when learning the sound in frame i . The linguistic environment of each new learner will be different, and they will inevitably introduce their own idiosyncrasies driven by learning mistakes, the desire to emulate certain individuals, and variations in their own physiology. However, on average we have no reason to expect anything other than unbiased variations around the template mean. Ignoring the effect of frames which are not templates, we therefore expect new speakers coming of age to use sounds which on average match the template mean. As older speakers with more archaic forms of speech die, the speech sounds of the population as a whole will move in the direction of the template mean. This behavior will induce an effective force on frame i which draws it toward the mean of its templates

$$\mathbf{f}_i^{\text{att}} := \hat{\mathbf{x}}_i - \mathbf{x}_i. \tag{13}$$

This is the simplest choice of interparticle force consistent with the above considerations, which are summarized visually in Fig. 5.

We now consider repulsive interactions, which are induced by disruptions to the learning process caused by frames which are nearby in acoustic space, but not sufficiently near to be templates. Given a frame i , the density of such *antitemplate* frames is

$$\widetilde{\psi}_i(\mathbf{x}) := \sum_{j \notin S_i} f_j \psi_j(\mathbf{x}). \tag{14}$$

We call $\widetilde{\psi}_i(\mathbf{x})$ the *antitemplate density*. Sounds from antitemplates will interfere with language learners' ability to recognize nearby template sounds, making the template sounds less likely to be copied [17,23,45]. To see how this might occur, consider a language that contains two acoustically similar vowel phonemes. Suppose that a language learner has noticed that these two sounds play two different roles in the language. We do not know how the developing mind achieves this but it

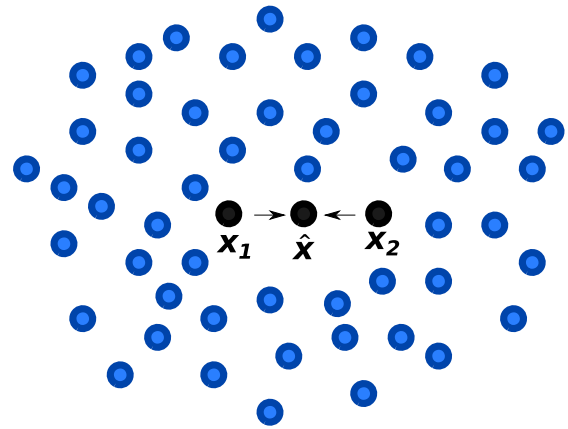


FIG. 5. The set of utterances for two different frames whose vowels are part of the same phoneme. The template mean is the average location of the frames, and since frames are attracted toward their template mean, they are attracted to other frames in the same phoneme.

is necessary in order to make sense of language. For example, there are *minimal pairs* of words such as ⟨pen⟩[p^hɛn] and ⟨pan⟩[p^hæn] which contrast only in a single vowel. Here, it is essential to be able to distinguish /ɛ/ from /æ/. When learning how these two phonemes should sound, there will be many occasions where words are uttered using a sound which, from a purely acoustic perspective, is hard to categorize as one or the other. Evidence for this is shown in Figs. 6 and 7 where, at least in American English, phonemes can be separated by as little as one standard deviation of their acoustic distribution. In this situation, an experienced speaker may be able to use a word's context, and the vowel's phonological environment, to efficiently identify what has been said. However, a younger speaker must build up this ability over time. Until then, there will be cases where the meaning of a word is ambiguous. Even if the word meaning is understood, the

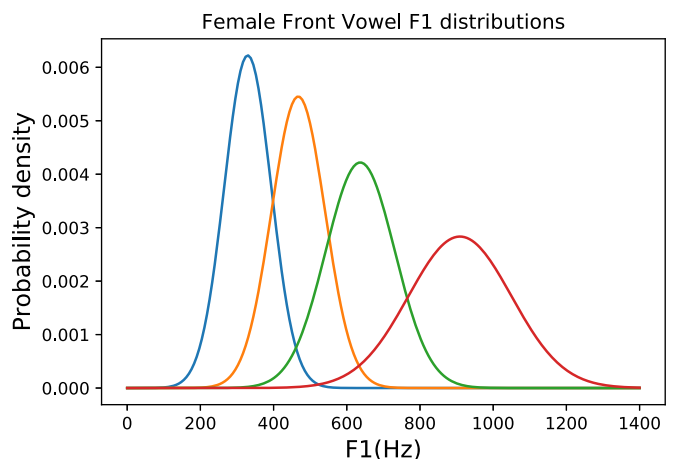


FIG. 6. Approximate distribution of F_1 values for female front vowels, derived from Peterson and Barney vowel data [6]. Curves are normal densities having means and standard deviations equal to those of the first formants of female front vowels. Color coding matches that in Fig. 1: blue i, orange ɪ, green ɛ, red æ.

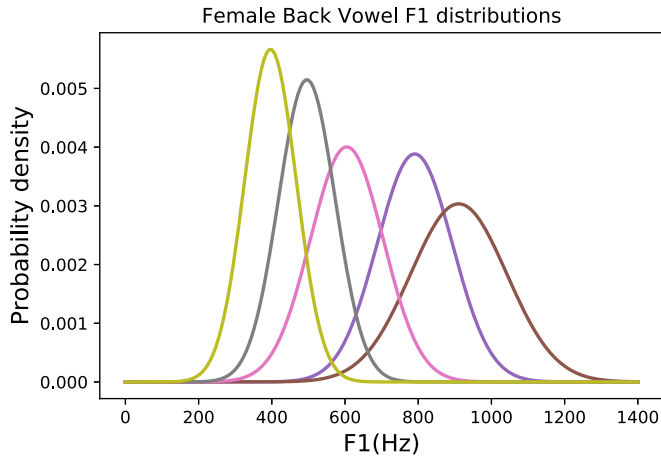


FIG. 7. Approximate distribution of F_1 values for female back vowels, derived from Peterson and Barney vowel data [6]. Curves are normal densities having means and standard deviations equal to those of the first formants of female back vowels. Color coding matches Fig. 1: light green u , gray v , pink o , purple Λ , brown α .

utterance may not be perceived as a legitimate pronunciation, if its phonemes are too close to other recognizable sounds. In both these cases the uttered sound is, we assume, less likely to influence the language development of the speaker. An argument similar to this is made by Labov [35] and is the basis of exemplar theory [23,24,45]. The effect of rejecting or otherwise reducing the importance of acoustically ambiguous phonemes during the learning process will be to push the sounds of the language away from regions of acoustic space which are overcrowded with phonemes. This will induce an effective repulsive force between frames that are not close enough to be considered as templates of one another.

To arrive at a plausible form for the repulsive effect on frame i from its antitemplate frames, consider a single frame $j \notin S_i$. A simple measure of the extent to which j crowds i is the ratio of the density of frame j utterances at \mathbf{x}_i , to the total template and antitemplate frame density in the same location, weighted for *functional load*, $\gamma \in [0, 1]$: the amount of work that individual vowel phonemes do in distinguishing words [48,49]

$$\frac{\gamma f_j \psi_j(\mathbf{x}_i)}{\widehat{\psi}_i(\mathbf{x}_i) + \gamma \widetilde{\psi}_i(\mathbf{x}_i)}. \quad (15)$$

As $\gamma \rightarrow 0$ sounds become irrelevant to word identification and so the effect of overcrowding becomes negligible. The effect of overcrowding by frame j will be proportional both to this overcrowding ratio and to its distance away from i : if more distant templates are rejected, the effect on the mean of the templates which are accepted will be larger. The total repulsive force on frame i consistent with these assumptions is then

$$\mathbf{f}_i^{\text{rep}} := \frac{\gamma}{\widehat{\psi}_i(\mathbf{x}_i) + \gamma \widetilde{\psi}_i(\mathbf{x}_i)} \sum_{j \notin S_i} f_j \psi_j(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j), \quad (16)$$

and the total force on frame i is $\mathbf{f}_i = \mathbf{f}_i^{\text{att}} + \mathbf{f}_i^{\text{rep}}$. To summarize: In the absence of phonemic overcrowding, the mean acoustic position of all the utterances of a frame made by new learners is equal to the mean position of all the templates for that

frame. These include the frame itself, and other frames containing sufficiently similar sounds. Phonemic overcrowding, realized as the close proximity of antitemplates, causes the rejection of templates in overcrowded regions, creating a bias away from these regions. Attraction toward the template mean generates a short-range attractive force between frames, and bias away from overcrowded regions generates a long-range repulsive force. The template range α determines the radius of the attractive region around a frame. Outside this radius, the strength and range of repulsive interframe forces are controlled, respectively, by the functional load γ and by the cloud radius σ .

The interactions defined by (13) and (16) are not symmetric or additive. More common frames exert a greater influence on their surroundings, and the effect of one frame on another depends on its relative rather than absolute density. Because each language learner in the community will be exposed to a different set of utterances, and different speakers may learn differently from what they hear, the evolution of frame position will not be deterministic. We capture this stochasticity in speaker behavior as a diffusion process with coefficient D , having a deterministic drift component given by the attractive and repulsive forces (13) and (16):

$$d\mathbf{x}_i(t) = \mathbf{f}_i dt + \sqrt{2D} d\mathbf{W}_i. \quad (17)$$

Here, \mathbf{W}_i is a two-dimensional Brownian motion [32]. We emphasize that this equation describes the evolution of the expected utterance $\mathbf{x}_i = \mathbb{E}[\mathbf{X}_i]$ for frame i , that is, the behavior of the community as a whole. This Itô stochastic differential equation (SDE) is equivalent to inertia-free Langevin dynamics [33]. The initial conditions of Eq. (17) depend on the problem we are interested in, but typically we will randomize the initial locations of frames within a bounded region representing the set of vowel sounds acoustically accessible to humans.

IV. DERIVATION AS AN EXEMPLAR MODEL

Our model (17) is phenomenological in the sense that it is motivated by empirical observation and theories of language learning, without being directly derived from an explicit model of this process. We now derive it as an approximate exemplar model, partly in order to connect it with recent theory (exemplar dynamics [22–24]) but also because it is useful (for Secs. VI and VIII and for future work) to have an explicit model of language learners accepting and rejecting templates. Exemplar dynamics in its purest form is an explicit computational model of a large population of sound units (exemplars) in the memories of speakers. Typically, these sounds are characterized by a single acoustic variable [23,24]. Exemplars for us are a subset of the utterances $\{\mathbf{X}_i\}_{i=1}^n$ (those which are not rejected by learners). Excepting the means $\{\mathbf{x}_i\}_{i=1}^n$, the distributions of the utterances \mathbf{X}_i are specified exogenous to the model. This simplification allows the dynamics of our language to be specified as a set of SDEs.

If an utterance $\mathbf{X}_j(t) \in S_i$ influences how a young speaker learns the vowel sound $\mathbf{x}_i(t)$, in frame i then it is an *exemplar* of that sound. Suppose that a template of frame i is uttered as sound \mathbf{x} and $p_i(\mathbf{x})$ is the probability that this sound is accepted as an exemplar of the frame. The *exemplar mean* for frame i

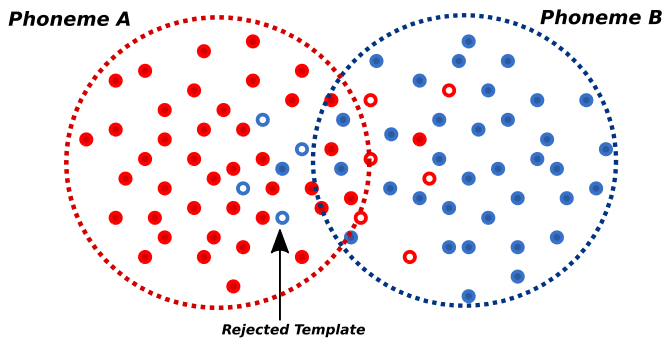


FIG. 8. The set of utterances for two different phonemes, each consisting of at least one frame. When two phonemes are near to each other in acoustic space, utterances of one may sound like the other increasing the likelihood that these utterances will not be accepted as valid templates by language learners. Such utterances are represented as open circles in the above diagram.

is then the expected value of its exemplars

$$\bar{\mathbf{x}}_i := \frac{\int \widehat{\psi}_i(\mathbf{x}) p_i(\mathbf{x}) \mathbf{x} d\mathbf{x}}{\int \widehat{\psi}_i(\mathbf{x}) p_i(\mathbf{x}) d\mathbf{x}}. \quad (18)$$

We may approximate the exemplar mean by expanding $p_i(\mathbf{x})$ to first order about the template mean. The template density (a superposition of Gaussian densities having overall mean $\hat{\mathbf{x}}_i$) is approximated as a Gaussian with mean $\hat{\mathbf{x}}_i$ [so $\nabla \widehat{\psi}_i(\hat{\mathbf{x}}_i) = 0$] and covariance $c_i \Sigma$. We have

$$\bar{\mathbf{x}}_i \approx \frac{\int \widehat{\psi}_i(\mathbf{x}) \mathbf{x} [p_i(\hat{\mathbf{x}}_i) + (\mathbf{x} - \hat{\mathbf{x}}_i) \cdot \nabla p_i(\hat{\mathbf{x}}_i)] d\mathbf{x}}{\int \widehat{\psi}_i(\mathbf{x}) [p_i(\hat{\mathbf{x}}_i) + (\mathbf{x} - \hat{\mathbf{x}}_i) \cdot \nabla p_i(\hat{\mathbf{x}}_i)] d\mathbf{x}} \quad (19)$$

$$= \frac{p_i(\hat{\mathbf{x}}_i) \mathcal{N}_i \hat{\mathbf{x}}_i + (\int \widehat{\psi}_i(\mathbf{x}) \mathbf{x} \otimes (\mathbf{x} - \hat{\mathbf{x}}_i) d\mathbf{x}) \nabla p_i(\hat{\mathbf{x}}_i)}{p_i(\hat{\mathbf{x}}_i) \mathcal{N}_i + (\int \widehat{\psi}_i(\mathbf{x}) (\mathbf{x} - \hat{\mathbf{x}}_i) d\mathbf{x}) \cdot \nabla p_i(\hat{\mathbf{x}}_i)} \quad (20)$$

$$= \hat{\mathbf{x}}_i + \frac{(\int \widehat{\psi}_i(\mathbf{x}) \mathbf{x} \otimes (\mathbf{x} - \hat{\mathbf{x}}_i) d\mathbf{x}) \nabla p_i(\hat{\mathbf{x}}_i)}{\mathcal{N}_i p_i(\hat{\mathbf{x}}_i)} \quad (21)$$

$$= \hat{\mathbf{x}}_i + \frac{c_i \Sigma \nabla p_i(\hat{\mathbf{x}}_i)}{p_i(\hat{\mathbf{x}}_i)}, \quad (22)$$

where \otimes is the outer product. This relation links the exemplar mean to the template mean via the acceptance probability. The number $c_i > 1$ is the ratio of the width of the template density to the width of the densities of individual frames. Assuming that frame clusters are tight compared to cloud size, then $c_i \approx 1$.

We now explicitly define the acceptance probability based on an overcrowding argument similar to that used in Sec. III. Consider frame i , located at \mathbf{x}_i , and also another point \mathbf{x} in vowel space. If \mathbf{x} is not too far from \mathbf{x}_i , and a large fraction of the sounds at \mathbf{x} come from templates of i , then these templates are unlikely to be confused with antitemplates of i near \mathbf{x} . However, if location \mathbf{x} is overcrowded with nearby antitemplates, then rejection becomes likely (see Fig. 8). The simplest choice consistent with these considerations is to let $p_i(\mathbf{x})$ be the relative density of i templates at \mathbf{x} , corrected for functional load

$$p_i(\mathbf{x}) = \frac{\widehat{\psi}_i(\mathbf{x}) \mathbf{1}_{|\mathbf{x} - \mathbf{x}_i| < R}}{\widehat{\psi}_i(\mathbf{x}) + \gamma \widetilde{\psi}_i(\mathbf{x})}, \quad (23)$$

where $R > 0$ is a cutoff radius beyond which sounds are rejected outright. In the absence of functional load ($\gamma = 0$), $p_i(\mathbf{x}) = \mathbf{1}_{|\mathbf{x} - \mathbf{x}_i| < R}$, so vowel sounds are rejected only when excessively distant from the current frame position, without reference to the crowding effects of other phonemes. Using (23) we have

$$\frac{\Sigma \nabla p_i(\hat{\mathbf{x}}_i)}{p_i(\hat{\mathbf{x}}_i)} = \frac{-\gamma \widehat{\psi}_i(\hat{\mathbf{x}}_i) \Sigma \nabla \widetilde{\psi}_i(\hat{\mathbf{x}}_i) (\widehat{\psi}_i(\hat{\mathbf{x}}_i) + \gamma \widetilde{\psi}_i(\hat{\mathbf{x}}_i))}{(\widehat{\psi}_i(\hat{\mathbf{x}}_i) + \gamma \widetilde{\psi}_i(\hat{\mathbf{x}}_i))^2 \widehat{\psi}_i(\hat{\mathbf{x}}_i)} \quad (24)$$

$$= -\frac{\Sigma \nabla \widetilde{\psi}_i(\hat{\mathbf{x}}_i)}{\gamma^{-1} \widehat{\psi}_i(\hat{\mathbf{x}}_i) + \widetilde{\psi}_i(\hat{\mathbf{x}}_i)} \quad (25)$$

$$= -\sum_{j \neq S_i} \frac{f_j \Sigma \nabla \psi_j(\hat{\mathbf{x}}_i)}{\gamma^{-1} \widehat{\psi}_i(\hat{\mathbf{x}}_i) + \widetilde{\psi}_i(\hat{\mathbf{x}}_i)} \quad (26)$$

$$= -\sum_{j \neq S_i} \frac{f_j \Sigma \Sigma^{-1}(\mathbf{x}_j - \hat{\mathbf{x}}_i) \psi_j(\hat{\mathbf{x}}_i)}{\gamma^{-1} \widehat{\psi}_i(\hat{\mathbf{x}}_i) + \widetilde{\psi}_i(\hat{\mathbf{x}}_i)} \quad (27)$$

$$= \sum_{j \neq S_i} \frac{f_j \psi_j(\hat{\mathbf{x}}_i) (\hat{\mathbf{x}}_i - \mathbf{x}_j)}{\gamma^{-1} \widehat{\psi}_i(\hat{\mathbf{x}}_i) + \widetilde{\psi}_i(\hat{\mathbf{x}}_i)} \quad (28)$$

yielding the following expression for the exemplar mean:

$$\bar{\mathbf{x}}_i \approx \hat{\mathbf{x}}_i + \sum_{j \neq S_i} \frac{f_j \psi_j(\hat{\mathbf{x}}_i) (\hat{\mathbf{x}}_i - \mathbf{x}_j)}{\gamma^{-1} \widehat{\psi}_i(\hat{\mathbf{x}}_i) + \widetilde{\psi}_i(\hat{\mathbf{x}}_i)}. \quad (29)$$

If acoustic differences between the mean vowel sounds of frames within one phoneme greatly exceed the difference between phonemes, or the cloud radius, then we can approximate the template means in the summand of (29) with \mathbf{x}_i . In this case we have

$$\bar{\mathbf{x}}_i - \mathbf{x}_i \approx \hat{\mathbf{x}}_i - \mathbf{x}_i + \sum_{j \neq S_i} \frac{f_j \psi_j(\mathbf{x}_i) (\mathbf{x}_i - \mathbf{x}_j)}{\gamma^{-1} \widehat{\psi}_i(\mathbf{x}_i) + \widetilde{\psi}_i(\mathbf{x}_i)} \quad (30)$$

$$= \mathbf{f}_i^{\text{att}} + \mathbf{f}_i^{\text{rep}}, \quad (31)$$

where $\mathbf{f}_i^{\text{att}}$ and $\mathbf{f}_i^{\text{rep}}$ are the phenomenological forces defined in Sec. III. If we consider a single frame, and assume that on average young speakers match the mean of the sounds they accept as exemplars for this frame, then the community's speech will evolve toward that mean. The simplest force consistent with this assumption is $\mathbf{f}_i = \bar{\mathbf{x}}_i - \mathbf{x}_i$, and we will show below that this force may be derived from a simple agent-based model. The above calculations showed that in the exemplar acceptance/rejection picture, this force decomposes into short-range attractive and long-range repulsive components given by (13) and (16).

To derive the SDE (17), we consider a community of N speakers whose evolution is driven by the replacement of older speakers by new speakers who learn from the community. We divide time into intervals of length $\delta t = N^{-1}$. At each interval, a speaker is selected uniformly at random from the population, "retired," and replaced with a new speaker whose language state is a random variable with expectation equal to the exemplar mean. We can write this new state $\bar{\mathbf{x}}_i + \boldsymbol{\epsilon}_i$ where $\boldsymbol{\epsilon}_i = s \mathbf{Z}_i$, with $\mathbf{Z}_i \sim \mathcal{N}(0, \mathbf{I})$, is a random variable which captures the stochasticity injected by the new speaker, perhaps by their own free will, physiological differences, or selective copying of certain individuals. This produces speakers who live for a geometrically distributed number of time intervals,

with mean N , giving an average lifespan of one time unit. To calculate the state of the community after a replacement, we first write the state of the speaker who is removed as

$$\mathbf{x}_i^{\text{dead}} = \mathbf{x}_i + \delta_i, \quad (32)$$

where δ_i is a zero-mean random variable with approximately the same statistical properties as ϵ_i . After a replacement the new frame position will be

$$\mathbf{x}_i(t + \delta t) = \mathbf{x}_i(t) - \frac{\mathbf{x}_i^{\text{dead}}(t)}{N} + \frac{\bar{\mathbf{x}}_i(t) + \epsilon_i}{N} \quad (33)$$

$$= \left(\frac{N-1}{N}\right)\mathbf{x}_i(t) + \frac{1}{N}[\bar{\mathbf{x}}_i(t) + \epsilon_i - \delta_i]. \quad (34)$$

Defining $\delta\mathbf{x}_i(t) = \mathbf{x}_i(t + \delta t) - \mathbf{x}_i(t)$, we have

$$\delta\mathbf{x}_i = (\bar{\mathbf{x}}_i - \mathbf{x}_i)\delta t + \sqrt{\frac{2s^2}{N}}\mathbf{Z}_i\sqrt{\delta t}. \quad (35)$$

In this simple model, stochasticity in the population as a whole is smaller in larger populations. However, an implicit assumption of the model is that new speakers are exposed to the whole community because their linguistic state depends on the exemplar mean. N is therefore not realistically equal to the number of speakers of an entire language because there are likely to be many smaller communities with dialects. It is possible to extend the above model to account for geographically or socially separated groups, but this is beyond the scope of this paper. Letting $\mathbf{W}_i \in \mathbb{R}^2$ be a standard two-dimensional Brownian motion then we have

$$\sqrt{\delta t}\mathbf{Z}_i \stackrel{d}{=} \mathbf{W}_i(t + \delta t) - \mathbf{W}_i(t) := \delta\mathbf{W}_i, \quad (36)$$

where $\stackrel{d}{=}$ denotes equality in distribution. We may therefore write our discrete SDE (35) as

$$\delta\mathbf{x}_i = (\bar{\mathbf{x}}_i - \mathbf{x}_i)\delta t + \sqrt{\frac{2s^2}{N}}\delta\mathbf{W}_i \quad (37)$$

$$= (\mathbf{f}_i^{\text{att}} + \mathbf{f}_i^{\text{rep}})\delta t + \sqrt{2D}\delta\mathbf{W}_i, \quad (38)$$

where

$$D = \frac{s^2}{N}. \quad (39)$$

From this we see that our phenomenological equation (17) is the continuous time equivalent of (38).

Figure 9 illustrates the difference between the exemplar model and its approximate form [the phenomenological model (17)]. In Fig. 9 we compare the shift in the exemplar mean calculated exactly using the acceptance probability (23), and the approximate shift calculated by expanding $p_i(\mathbf{x})$ to first order about $\hat{\mathbf{x}}_i$. We have considered two phonemes each composed of a set of collocated frames, and set the template range to $\alpha = 0^+$, so we only see the repulsive component of the interaction. As expected, the approximation converges to the exact result as the phoneme separation tends to zero. For larger separations the approximate shift falls away more quickly, creating a shorter-range repulsive interaction. This difference in range appears because in (17) we are measuring interference effects on frame i using the relative density of antitemplates at the location of i , rather than in the outer reaches of its cloud, so antitemplate frames must be closer

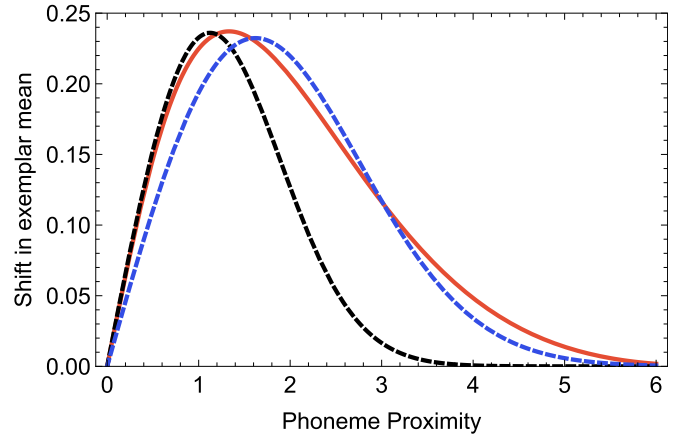


FIG. 9. Shift in the exemplar mean away from the frame position $\bar{\mathbf{x}}_i - \mathbf{x}_i$ for one phoneme (modeled by unit variance Gaussian) induced by another identically sized phoneme. Red: exact calculation using acceptance probability (23) when $\sigma = 1$, $\gamma = 0.5$, $R = 3$. Black: approximation with the same parameter values (30). Blue: approximation with larger clouds $\sigma = 1.5$ and lower functional load $\gamma = 0.3$.

to have an effect. By increasing the cloud radius relative to the exemplar model, and lowering the functional load, we can achieve a similar interaction force in both models. The two forms of the model are alternative but qualitatively similar ways of characterizing how phoneme overcrowding affects vowel sound evolution, based on the same underlying ideas. We work with the approximate model because of its simpler form which is efficient to simulate and to analyze mathematically. However, we take account of the effects above when selecting the cloud radius.

V. COMPARISON TO REAL SYSTEMS

We now explore the behavior of our model by simulation, and compare to real systems. The range of possible vowel sounds is constrained by the limits of the human vocal apparatus. In Sec. VA we define the shape and dimensions of this accessible space, and specify boundary conditions. In Sec. VB we use a large cross-linguistic sound inventory database to explore how the vowel sounds of real languages are distributed in this space. In Sec. VC we explore the distributions generated by our model, and compare them to the statistics of empirical distributions.

A. Defining vowel space

Although the traditional vowel chart is a wide-based trapezium, average formant values for vowel sounds suggest that the accessible region in (F_1, F_2) space is closer to triangular in shape, with /a/ forming the lowest apex. Figure 10 shows the mean formants of the *peripheral* (outermost) vowels of Northern Standard Dutch. In this particular language /o/ and /ɔ/ are very close, suggesting that the difference between them is captured by something other than their first two formants. In many other languages /o/ and /ɔ/ are not close, for example, in American English $\Delta F_1 = 226$ Hz [50]. This highlights the fact that the meanings of the phonetic

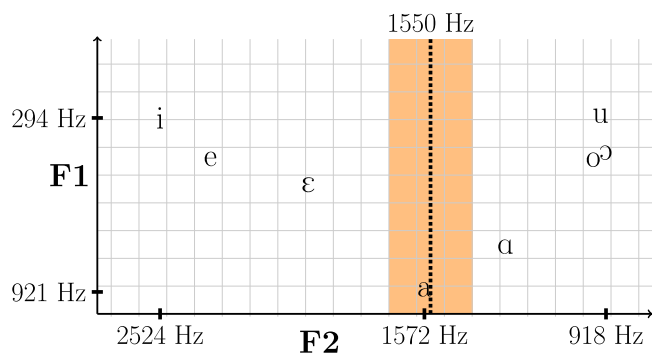


FIG. 10. Average formant values of the peripheral vowels $i, e, \epsilon, a, o, \circ, u$ for female speakers of Northern Standard Dutch [51]. Marked formant frequencies give the positions of the extreme vowels $/i, a, u/$. Orange region shows the typical frequency range $F_2 \in [1400, 1700]$ for the second subglottal resonance in female speakers [4] (see Sec. VI).

vowel symbols are not precisely defined in terms of any measurable quantity. Rather, they are a tool for describing the general structure of the phonemic systems of languages. The approximate correspondence between typical formant values and the traditional IPA vowel chart is all the more remarkable for this. A particularly notable discrepancy is the relative heights of $/a/$ and $/\alpha/$, which are identical in the standard IPA vowel chart but in reality appear to differ in their first formant.

Motivated by the above considerations, and to avoid introducing unjustified complexity into the model, we opt for a symmetric trapezoidal (approximately triangular) space, as is used to tabulate differences between vowel systems in large inventories [12,52]. Figure 11 illustrates this and also shows *representative* positions for the nine most common

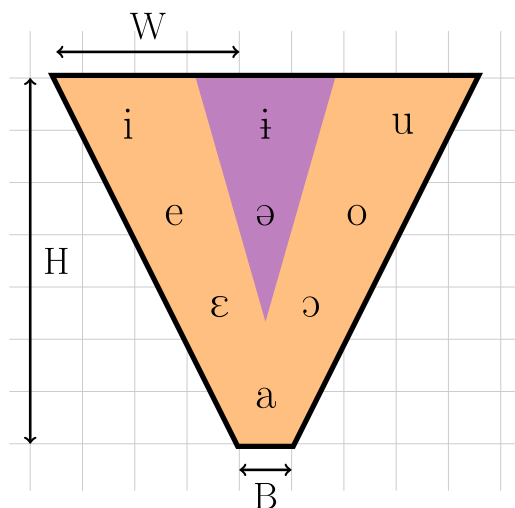


FIG. 11. Symmetric trapezoidal vowel space used in our model, together with the IPA symbols and approximate positions of the nine most common short vowels (see Table I). Orange region is the *periphery* of vowel space, purple is the *interior*. Speakers of English may be surprised by the absence of low back vowels $/\alpha/$ (7%) or $/\upsilon/$ (4%) which appear in words like $\langle \text{pot} \rangle$. These are rather rare worldwide.

TABLE I. The nine most common short vowels and their long versions. Frequency column shows the percentage of languages in the PHOIBLE database [53] which contains each vowel. Frequencies (and IPA symbols) of long forms are in brackets. GenAM and RP stand for “general American” and “received pronunciation.”

IPA symbol	Example word	Frequency
i (i:)	GenAm USA: $\langle \text{bee} \rangle$ [bi:]	92% (32%)
u (u:)	GenAm USA: $\langle \text{shoe} \rangle$ [ʃu:]	88% (29%)
a (a:)	RP UK: $\langle \text{now} \rangle$ [naʊ]	86% (30%)
e (e:)	RP UK: $\langle \text{bay} \rangle$ [beɪ]	61% (21%)
o (o:)	GenAm USA: $\langle \text{go} \rangle$ [gou]	60% (21%)
ϵ (ε:)	GenAm USA: $\langle \text{bet} \rangle$ [bet]	37% (11%)
\circ (ɔ:)	RP UK: $\langle \text{bore} \rangle$ [bɔə]	35% (10%)
ə (ə:)	RP UK: $\langle \text{bear} \rangle$ [beə]	22% (4%)
i (i:)	S. African Eng: $\langle \text{lip} \rangle$ [lip]	16% (1%)

short vowels [53], whose relative frequencies, together with example words, are given in Table I. The locations in Fig. 11 have been chosen to lie approximately at the centers of the regions of vowel space used in the simplified classification scheme which we describe in Secs. VB and VC. They should not be interpreted as empirically measured average formant positions, although the typical formant values of these vowels will have a similar arrangement to that shown in Fig. 11. Letting the bottom left vertex of vowel space define the origin of coordinates, then the locations are

$$x = \frac{B}{2} \pm \frac{kW}{4}, \tag{40}$$

$$y = \frac{H}{8} + \frac{kH}{4} \tag{41}$$

with $k \in \{0, 1, 2, 3\}$, with interior vowels $/\text{ə}, i/$ at $(B/2, 5H/8)$ and $(B/2, 7H/8)$.

In our model, repulsion between phonemes occurs as a result of cloud overlap, rather than from any notion of *contrast*. We therefore use formant cloud shapes to estimate the dimensions of our trapezium. It is clear both from Fig. 1 and formant inventories for different languages [47] that clusters of formant data representing different phonemes within a single language, and the same phoneme across different languages, vary in size and shape. We characterize shape and position as follows. Given a large formant data set $\{F_{1i}, F_{2i}\}_{i=1}^N$, we let $S(X)$ denote the set of formants which were uttered for phoneme X . We define the F_1 mean and radius (standard deviation) of phoneme X to be

$$\mu_1(X) = \frac{1}{|S(X)|} \sum_{i \in S(X)} F_{1i}, \tag{42}$$

$$\sigma_1(X) = \left(\frac{1}{|S(X)|} \sum_{i \in S(X)} F_{1i}^2 \right) - \mu_1^2(X) \tag{43}$$

with similar expressions for the F_2 mean and radius. Because of vocal tract size, average formant values for male and female speakers differ systematically. This elongates formant clouds at the population level. Since listeners subconsciously normalize for such differences [25], we consider only one sex: female. In Fig. 6 the F_1 radius of phoneme

clusters increases (approximately linearly) with increasing F_1 . If $\sigma_1(X) = \sigma_2(X)$, we say that X is spherical. Provided phonemes which are close enough to interact experience approximately the same systematic shape variation with position in formant space, then a transformation of this space which makes all phonemes spherical will have little effect on their overlaps. This assumption means we can model all frame clouds as spherical with the same cloud radius in this transformed space.

When frame and phoneme clouds are spherical, vowel arrangements which are near equilibrium with respect to repulsive forces are strongly affected by the aspect ratio of the transformed vowel space

$$AR = \frac{\text{Max width}}{\text{Height}}. \tag{44}$$

We may estimate this ratio using the formant data in Fig. 1. We define a *standardized distance* between the high vowels /i/ and /u/, which is approximately the same in both formant and transformed space

$$\Delta(u,i) := \frac{\mu_2(u) - \mu_2(i)}{\bar{\sigma}_2} \tag{45}$$

$$= \frac{1891}{227} \tag{46}$$

$$\approx 8.3, \tag{47}$$

where $\bar{\sigma}_2$ is the average radius of u and i. A similar calculation for front vowels gives

$$\Delta(\ae,i) = \frac{\mu_1(\ae) - \mu_1(i)}{\bar{\sigma}_1} \tag{48}$$

$$= \frac{580}{93} \tag{49}$$

$$\approx 6.2, \tag{50}$$

where $\bar{\sigma}_2$ is the average radius of the front vowels /i, ɪ, ε/, and /æ/. Since /æ/ is not the lowest vowel, but lies between /a/ and /ε/, the true standardized height of vowel space is greater than this. According to the standard vowel chart /æ/ lies midway between /a/ and /ε/ so we approximate $\Delta(a,u) = 6\Delta(\ae,u)/5$. Using our two standardized distances we can estimate the aspect ratio

$$AR \approx \frac{\Delta(u,i)}{\Delta(a,i)} = \frac{8.3}{7.4} \approx \frac{8}{7}. \tag{51}$$

We take $H = 7$ and $B + 2W = 8$ so that a unit phoneme cloud radius would be consistent with the Fig. 1 data. We set $B = 1$ to accommodate the lowest vowel, giving the vowel space dimensions shown in Fig. 11 which we use for all simulations. Because the phenomenological model generates shortened interaction range we set the frame cloud radius to be $\sigma = 1.5$ in simulations, which generates interactions of comparable range to the exemplar version with unit radius (see Fig. 9). We model the effect of vowel space boundaries using a repulsive force perpendicular to each boundary, of magnitude

$$|\mathbf{f}_i^{\text{bou}}| = \frac{E}{2} \left[1 + \tanh \left(\frac{|\Delta \mathbf{x}_i|}{w} \right) \right], \tag{52}$$

TABLE II. The nine vowel categories used in our analysis. “Frequency of representation” is the percentage of surveys in the PHOIBLE repository [53] containing the category representative (after stripping modifications). “Frequency of category” is the percentage of surveys containing at least one category member. Categories’ members were selected based on their typical formant value proximity to the category representative.

Category repr.	Freq. of repr.	Category description	Members of cat.	Frequ. of cat.
i	96.1%	High front	i,ɪ,y,ʏ	99.4%
u	92.4%	High back	u,ʊ	99.0%
a	91.4%	Low	a,ɑ,æ,ɶ,ɛ	94.5%
e	74.2%	Upper mid front	e,ø	74.8%
o	74.1%	Upper mid back	o	74.1%
ɛ	39.0%	Lower mid front	ɛ,ɛ̃	40.7%
ɔ	37.5%	Lower mid back	ɔ,ʌ,ɒ	40.1%
ə	23.9%	Central	ə,ɜ,ɝ,ɞ,ɟ	29.1%
ɨ	17.4%	High mid	ɨ,ʉ,ɤ,ø	26.9%

where $|\Delta \mathbf{x}|$ is the distance from the boundary, E is the maximum magnitude, and w is the width of the boundary. We set $E = 10$, $w = 0.2$ in all simulations.

B. Properties of real vowel systems

Our sources of statistical information on vowel system properties are PHOIBLE [53], an online repository of phonological inventory data (containing 2186 languages); Maddieson’s “Patterns of Sounds” [12], based on a representative sample of the world’s languages, contained in the UCLA Phonological Segment Database (UPSID), and Crothers’ vowel system typology [52], which sought to find a simplified classification of vowel systems.

Both Crothers and Maddieson use the idea of vowel quality, allowing them to identify sets of similar sounds as equivalent. We mimic this approach by identifying every vowel sound as a member of one of nine categories, each represented by one of the most common sounds shown in Fig. 11. All of the most common systems with fewer than 10 vowels identified by Crothers also used only these nine sounds [52]. The phonetic alphabet is equipped with an extensive notation for recording subtle modifications of the 28 symbols on the standard IPA vowel chart. For example, surveys recorded in the PHOIBLE repository together contain 1094 symbols representing vowels. We first removed all modifying marks (producing 28 symbols) then used the mapping in Table II to assign category membership. For example, /æ:/ is first stripped of its length mark, and then assigned to the category /a/. Our category assignment was based on estimated overlap in formant space, using typical formant values for the 28 IPA vowels. The frequencies in Table II also illustrate the percentage representation of each category amongst the surveys in the PHOIBLE repository. It is interesting to note the very high degree of front-back symmetry in these representations, which is reflected in our choice of system geometry.

Our categorization scheme generates a *typology*: a classification of the world’s languages into groups according (in this case) to the structure of their vowel system. Of the $2^9 = 512$

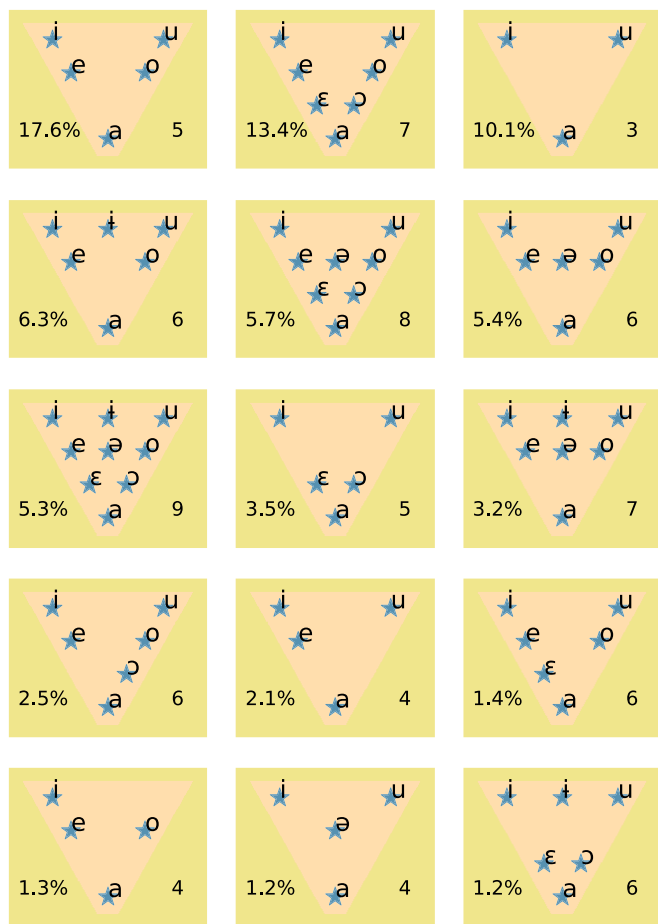


FIG. 12. The 15 most common vowel system types following categorization of PHOIBLE repository vowel data. Percentages give frequency amongst 3020 surveys, and integers give system size.

possible system types in our scheme, only 106 are attested out of 3020 surveys (note: the number of surveys exceeds the number of languages due to dialectal variation [54]). This suggests that the dynamical processes which generate vowel systems are attracted to certain equilibria, and that many possible types are inherently unstable or unattractive to speakers. Figure 12 shows the 15 most common vowel systems in the world’s languages, accounting for over 80% of the surveys. Eleven have perfect front-back symmetry, and of the four that don’t, two form a symmetrical pair. Similar results were obtained by Crothers [52], with eight of the eleven most common systems in his typology being symmetric.

The relative frequency of different system sizes (cardinalities) has also been of interest to linguists [12,14,52]. According to both Crothers and Maddieson, five vowel systems are more common than any other, with /a,e,i,o,u/ the most common of all. Our typology reproduces this result. However, there is no unique solution to the problem of determining the relative frequencies of different vowel system sizes. Human societies in their “natural” preindustrial state have low geographical connectivity, supporting many closely related languages and dialects in relatively small areas. A nation state destroys this variation, replacing it with a *national language*. Regions such as Papua New Guinea, Central and West Africa, and Australia therefore contribute a disproportionately high

TABLE III. Percentage of languages by their vowel system sizes (number of qualities) estimated by Maddieson [12], Crothers [52], and using our typology, based on PHOIBLE [53].

Num. vowels	Maddieson	Crothers	PHOIBLE
2			0.2%
3	5.4%	11.1%	10.3%
4	8.5%	10.6%	6.8%
5	30.9%	30.8%	22.7%
6	18.9%	19.2%	16.5%
7	14.8%	13.5%	16.8%
8	5.4%	4.3%	7.5%
9	7.9%	7.2%	8.5%
10, 11, ...	8.1%	3.4%	10.6%

number of languages. This problem is typically addressed by sampling uniformly from language family groups [12] and from geographical areas [52]. Ambiguities can also result from the need to define vowel quality. Despite this, in Table III we see that our results (all PHOIBLE surveys with no adjustments for language size or family) are in broad agreement with those of Maddieson and Crothers.

The final statistical property that we consider is correlation between different sounds (two-point functions, in statistical mechanics). Given a category X , we define the indicator function that it is present in survey ℓ :

$$S_X(\ell) = \begin{cases} 1 & \text{if } X \in \ell, \\ 0 & \text{if } X \notin \ell. \end{cases} \tag{53}$$

The indicator is a binary variable, and the correlation between two categories is

$$\phi_{XY} := \frac{\langle S_X S_Y \rangle - \langle S_X \rangle \langle S_Y \rangle}{\sqrt{(\langle S_Y^2 \rangle - \langle S_X \rangle^2)(\langle S_Y^2 \rangle - \langle S_Y \rangle^2)}}, \tag{54}$$

where $\langle \cdot \rangle$ denotes the average over all surveys. In statistics this correlation is called the *phi coefficient*, but it is also equal to the Pearson correlation coefficient. Table IV shows these correlations calculated from PHOIBLE, as well as simulated values. The strongest correlations are between front and back vowels of the same height: that is, if we have one of a front-back pair at given height, then we are likely to have the

TABLE IV. The 10 largest correlation (ϕ) coefficients between categories in the PHOIBLE database and the model. Starred pairs appear in both lists. Model parameters $\sigma = 1.5$, $\gamma = 0.5$, $D = 0.025$, $n = 150$.

Rank	Pair	ϕ (PHOIBLE)	Pair	ϕ (Model)
1	ɔ, ε	0.74	ɪ, o	0.47
2	*e, o	0.71	ɪ, e	0.42
3	i, u	0.28	*e, o	0.32
4	*ə, ɪ	0.19	*ə, ε	0.32
5	ə, ɔ	0.18	e, ɔ	0.29
6	*ə, ε	0.15	*ε, o	0.23
7	ə, e	0.15	ε, ɪ	0.23
8	u, ɔ	0.15	a, e	0.20
9	*ə, o	0.15	*ə, o	0.16
10	*o, ε	0.15	*ə, ɪ	0.15

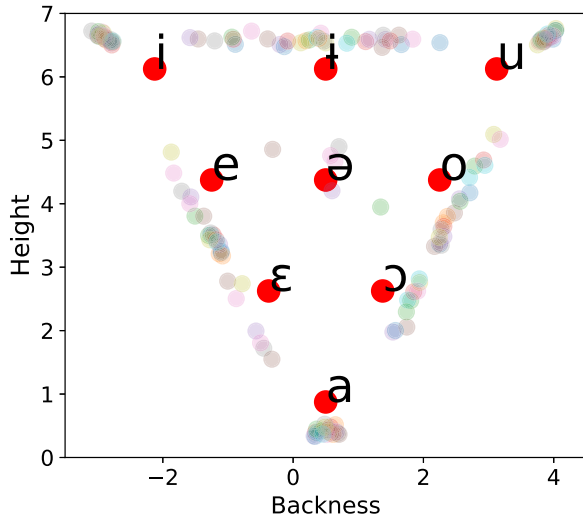


FIG. 13. Red dots show positions of standard vowels used for classification. Each colored dot shows the position of a single phoneme obtained by mean shift cluster analysis of quasistable state of $n = 150$ frame system. Parameter values: $D = 0.025$, $\sigma = 1.5$, $\gamma = 0.5$, $\alpha = 1$. Plot superimposes 20 different simulated vowel systems.

other. After these, there are weaker correlations between the mid central category and its surroundings, and in particular between the high and mid central categories i and ϵ . We will discuss the significance of this relationship in Sec. VI.

C. Model behavior and cross-linguistic comparison

To investigate the behavior of our model, and compare to real systems, we generate frame distributions beginning from randomized starting configurations where frames form a constant intensity Poisson point process [55] within vowel space. We then run the model for sufficient time (25 lifetimes) for a set of phoneme clusters to form, and to settle into a quasistationary state. Stabilization consists of frame movement in response to repulsive interactions, and occasional *mergers* between phonemes. More rarely, we see *splitting* events, where groups of frames spontaneously break away from a phoneme. When only one or two frames break away, we call this *evaporation*. In order to determine what vowel qualities our model has produced, we perform a mean shift clustering [56] of the locations of frames once a stable state is reached. The mean shift algorithm locates the peaks of a kernel density estimate [57] of the distribution of frames within vowel space. Peaks of this density represent the centers of phonemic clusters. Having identified the locations of our phonemes, we then assign each to one of our nine categories (Table II) by proximity to the positions of their representative sounds, shown in Fig. 11. In some cases we find that two simulated peaks are assigned to the same quality: we interpret this as a primitive form of allophonic variation, and count both peaks as a single phoneme. Figure 13 shows a superposition of the phoneme locations computed by this process. There are dense clusters of phonemes at the vertices of the space, corresponding to the three most common vowel qualities $/a/$, $/i/$, $/u/$. We also see clusters at the peripheries of vowel

space aligned with the boundaries, corresponding to the front, back, and high central vowels. We note that these locations are averaged both over the frames within each phoneme, and over the cloud of formant values for each frame. For this reason we would expect individual experimental formant measurements to be much more dispersed in acoustic space than the clusters in Fig. 13. Whether the mean formant values of the phonemes of real languages *line up* along vowel space boundaries to the extent seen in Fig. 13 is a difficult question to answer empirically. While cross-linguistic formant data sets do exist [43], the number of speakers involved in each individual study is typically small (less than 10). For this reason, the average formant values are subject to substantial noise. We also note that the shape of vowel space (the acoustically accessible region) will be different for each speaker, so the distance of an individual’s peripheral vowels to the boundary of their own vowel space may vary between speakers. In this work we have represented vowel space boundaries using short-range repulsion, consistent with the earlier maximal contrast models [13], but we note that at least one recent (one-dimensional) exemplar model [23] allows for a system-wide bias toward less extreme sounds, known as *lenition*.

The parameter with greatest influence on how many vowels form is the template range α . We can obtain an approximate lower bound on this parameter by considering the threshold for (first) formant frequency discrimination in normal speech, which is $\Delta F_1 \approx 50 \pm 10$ (Hz) [58], where $F_1 \in [235, 850]$ [59]. In standardized coordinates (dividing by $\bar{\sigma}_1$), this corresponds to a template range $\alpha_{\min} \approx 0.5$. It is likely that the acoustic proximity that defines two sounds as equivalent will vary between languages because phoneme clouds and their separations vary between languages [47]. Evidence for this is provided by vowel-to-vowel coarticulation [60] where the articulatory requirements (configuration of the vocal articulators) for a phoneme are anticipated during the production of a previous phoneme. Such interactions between phonemes at different positions in a word lead to greater variability in how the same phoneme is pronounced in different words. For example, in [60], utterances of the form $/apV/$ where V is a vowel which gives the *context* of $/a/$ were analyzed to discover the extent to which the choice of V altered the average formants of $/a/$ in three Bantu languages: Ndebele, Shona, and Sotho. Whereas $/a/$ is relatively isolated in the vowel systems of Ndebele and Shona ($/i,e,a,o,u/$), it is relatively crowded in Sotho ($/i,e,\epsilon,a,\text{ɔ},o,u/$). Formant experiments showed that coarticulatory effects on $/a/$ were considerably greater in Ndebele and Shona than in Sotho. That is, a greater range of sounds were used as if they were the same phoneme in the languages with fewer vowels. In our model, such languages would have a greater template range. Additional factors used for contrast beyond the first two formants may also make precision in these formants less important, thereby altering α . We explored this effect (Fig. 14) by generating the mean number of phonemes for a series of template ranges. From this we see that our estimate for α_{\min} is consistent with the observation that very few (3%) of all languages have more than nine vowel qualities [52].

At high template range we find that most systems have only three vowels, typically $/i/$, $/a/$, $/u/$, consistent with

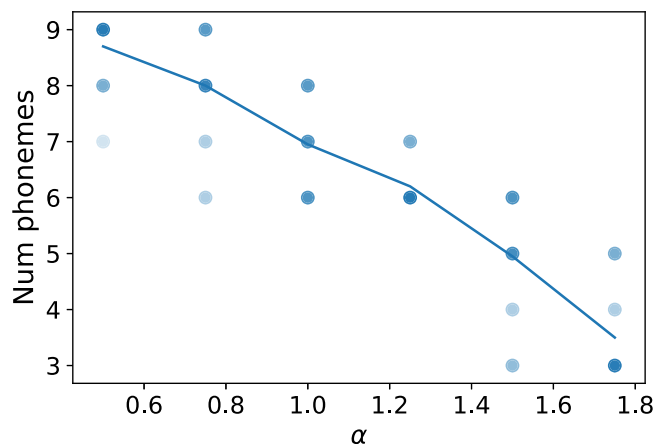


FIG. 14. Average number of spontaneously formed phonemes vs template range, for a system of $n = 150$ frames. Parameter values: $D = 0.025$, $\sigma = 1.5$, $\gamma = 0.5$. Dots show underlying data with darker shading indicating repeated data points.

observed systems. Almost no real languages have fewer than three vowel qualities (Table III).

From Fig. 14 we see that with a template range in the interval $1 \leq \alpha \leq 1.5$ we obtain vowel system sizes in the range $\{3, \dots, 8\}$, capturing $\approx 84\%$ of the size variability across world languages [12]. However, α is defined exogenous to the model, and we do not have means to estimate its distribution, other than by comparison to the distribution of vowel system sizes. In order to compare the relative frequencies with which different vowel phonemes appear in our model to their frequencies in real languages, we use Maddieson’s data in Table III and [12]. We tabulate all 120 simulated vowel systems used to create Fig. 14, and then compute the empirical probability mass function over the nine vowels for each size of system. We can then compute relative vowel frequencies over all system sizes as an average over our fixed-size mass functions, weighted by the frequencies of each mass function in the world’s languages. The results are given in Table V. The predictions of the model typically lie within $\approx 10\%$ of the observed values except for the high and central vowels $/i/$ and $/\partial/$. As with other dispersion theories, our

TABLE V. Predictions (compared to data from Table II) for the frequencies of the nine most common vowel qualities, with and without quantal effects. Simulation parameters: $n = 150$, $\sigma = 1.5$, $\gamma = 0.5$, $D = 0.025$. In the quantal case we have parameters $Q = 2$, $\omega_x = 2$, $\omega_y = 1$.

Category	Obs. Freq.	Pred. Freq.	Quantal pred.
i	99%	98%	93%
u	99%	97%	93%
a	94%	94%	87%
e	75%	64%	67%
o	74%	56%	68%
ɛ	41%	36%	39%
ɔ	40%	44%	36%
ə	29%	11%	1%
ɨ	27%	64%	36%

TABLE VI. The most common vowel systems of each size in PHOIBLE [53], Crothers [52], and the model. Frequency (%) columns show the percentage of vowel systems of the given size which have the given form. Simulation parameters: $n = 150$, $\sigma = 1.5$, $\gamma = 0.5$, $D = 0.025$.

No.	Rank	PHOIBLE	%	Crothers	%	Model	%
3	1	a,i,u	93	a,i,u	100	a,i,u	62
3	2	a,i,o	2			ɛ,i,u	17
3	3	a,e,o	2			ɔ,i,u	12
4	1	a,e,i,u	28	a,ɛ,i,u	59	a,e,i,u	24
4	2	a,e,i,o	17	a,i,u,ɨ	41	a,o,i,u	24
4	3	a,i,ə,u	16			a,ɛ,i,u	16
5	1	a,e,i,o,u	67	a,ɛ,i,ɔ,u	86	a,ɛ,i,o,u	28
5	2	a,ɛ,i,ɔ,u	13	a,ɛ,i,o,ɨ	8	a,ɛ,i,o,u	17
5	3	a,ɛ,i,o,u	4			a,i,o,u,ɨ	13
6	1	a,e,i,o,u,ɨ	31	a,ɛ,i,ɔ,u,ɨ	73	a,ɛ,i,o,u,ɨ	55
6	2	a,e,i,o,u,ə	27	e,i,o,u,ɛ,ɔ	18	a,ɛ,i,ɔ,u,ɨ	20
6	3	a,e,i,o,u,ɔ	12			a,ɛ,i,o,u,ɨ	18
7	1	a,e,i,o,u,ɛ,ɔ	64	a,ɛ,i,o,u,ə,ɨ	50	a,ɛ,i,o,u,ɔ,ɨ	41
7	2	a,ɛ,i,o,u,ə,ɨ	15	a,ɛ,i,o,u,ɛ,ɔ	46	a,ɛ,i,o,u,ɛ,ɨ	36
7	3	a,e,i,o,ɔ,ə	4			a,ɛ,i,ɔ,u,ə,ɨ	8

model predicts that a high central vowel should appear with high probability, when in real languages it is rather rare. A possible explanation for this rarity is the existence of a natural resonance in the human vocal tract which occurs between the front and back of vowel space, creating a discontinuity in the relationship between articulatory and acoustic parameters [11,43,61]. We return to this point in Sec. VI. The fact that interior vowels, represented by $/\partial/$ in our typology, are rare in our model may be understood by noting that their existence relies on a balance of repulsive forces from other sounds at the boundaries of vowel space. If stochastic effects cause two of these peripheral sounds to merge, then this can destabilize the central vowel, which moves out to the boundary to fill the gap. The reverse of this mechanism is a split in a peripheral vowel, creating an overcrowded boundary, forcing a phoneme into the interior. Unlike mergers, splits typically require some mechanism beyond the simple learning model we have defined, such as the conversion of two allophones into distinct phonemes, or borrowing of sounds from some exterior source like another language or dialect [35]. An extension to the model which allows allophonic variation is discussed in Sec. VII. As mentioned in Sec. IV, the model can also be extended to allow for interacting communities, or linguistic systems, but this is beyond the scope of this paper.

Table VI shows that apart from small height variations in the mid vowels $/e, \epsilon/$ and $/o, \text{ɔ}/$, our model, the data from PHOIBLE, and the Crothers study agree on the most common vowel systems. As we noted earlier, these height variations are subject to the interpretation of individual linguists, and should not be thought of as corresponding to a precisely defined formant interval. In less common systems there is more disagreement between the three typologies. In particular, the seven and eight vowel systems generated by the model are unrealistically likely to possess a high central vowel. We conclude that while the model matches the broad characteristics of real systems, there are details which it fails

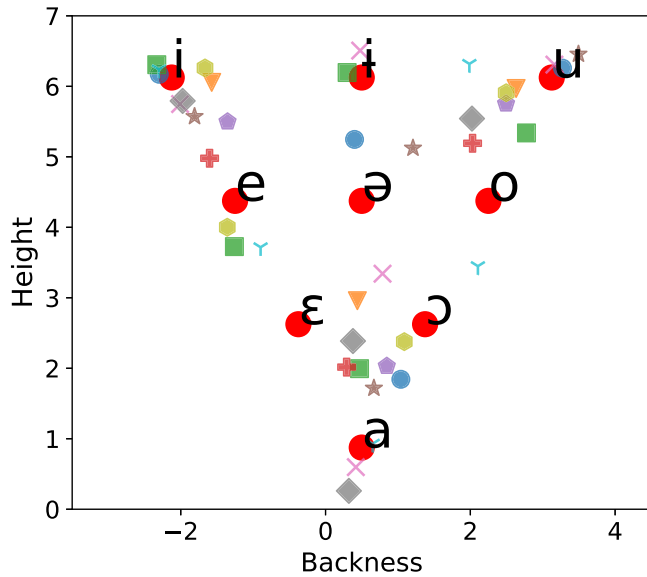


FIG. 15. Phoneme locations for 10 simulations of $n = 150$ using low functional load ($\gamma = 0.1$). Other parameter values: $\sigma = 1.5$, $\alpha = 1.25$, $D = 0.05$. Each set of symbols represent the phonemes from a single simulation.

to match. At the same time, the question of which typology is correct may not have an answer.

We now consider the effect of functional load. The *functional load hypothesis*, first proposed by Gilliéron [48,62], maintains that the probability of phoneme loss is inversely related to the amount of “work” done by the phoneme in identifying words. One simple measure of this work is the number of minimal pairs that a phoneme distinguishes, and a recent cross-linguistic corpus study [48] has shown that phonemes which define more minimal pairs are less likely to merge. In our model, phoneme merger can occur if the peripheries of two clusters are closer than the template range. Peripheral frames from the two clusters are attracted, pulling the clusters progressively closer. Merger is more likely if interphoneme forces are weaker, and from the definition of the repulsive force (16), we see that reducing functional load γ weakens them (see Fig. 15). Since vowel system formation consists of sequential cluster merging (and splitting), we expect low functional load to result in smaller phoneme inventories. Figure 16 shows the strength of this effect in our model for two levels of stochastic noise $D \in \{0.025, 0.05\}$. We find that the effects on system size (and therefore merger probability) are systematic but weak. With lower noise, the extremes of functional load produce an average difference of one phoneme, and for higher noise the difference is larger, but seen only at very low functional load where weak repulsive forces between phonemes allow some new systems to form (Fig. 15). For example, we have three different four vowel systems containing a mid central vowel, matching the 14th most common system in PHOIBLE (Fig. 12). We also see that doubling the stochastic diffusion D has a subtle but systematic effect on system size. Stochastic effects are required to bring two phonemes within merger range, so higher diffusivity results in more mergers, and therefore smaller systems.

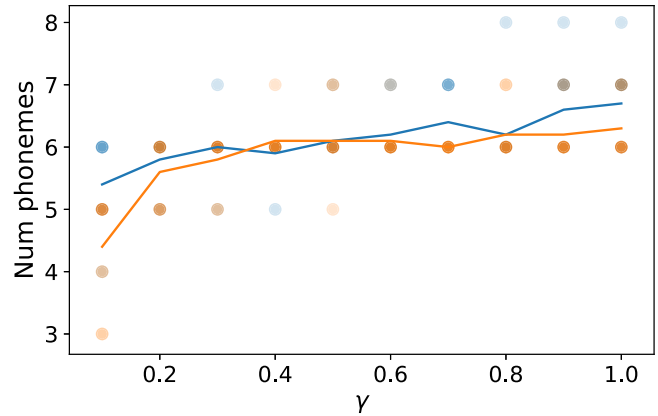


FIG. 16. Average number of spontaneously formed phonemes vs functional load, for a system of $n = 150$ frames. Parameter values: $\sigma = 1.5$, $\alpha = 1.25$. Blue curve $D = 0.025$. Orange curve $D = 0.05$.

VI. EFFECT OF THE SUBGLOTTAL RESONANCE

Our model, in its simplest form, may be viewed as a form of dispersion theory [13,19]. It explains the placement of vowel phonemes in terms of a force which acts to maximize the acoustic distances between them. Patterns of sounds for which these forces are in equilibrium, or near it, are more likely to be observed in the model, helping us understand why certain vowel sounds, and combinations of sounds, are more common than others. An alternative explanation, *quantal theory* [4,11,43,63], begins from the observation that the relationships between articulatory configurations and acoustic outputs of the human vocal apparatus contain pronounced nonlinearities where small changes in articulatory parameters can generate relatively large changes in acoustic output (Fig. 17). The central idea of quantal theory is that these nonlinearities quantize acoustic space into separate stable regions where the effects of changing articulatory parameters are small and predictable. Within these regions, speakers can more reliably produce a desired output, increasing the efficiency of communication. It is argued that these *quantal*

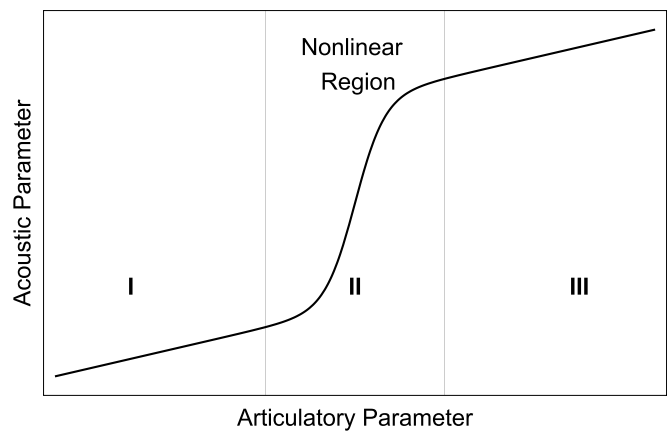


FIG. 17. Nonlinear relationship between articulatory and acoustic parameters. The nonlinear region II divides acoustic/articulatory space into two *quantal regions* I and III, where speech output is less sensitive to changes in articulation.

regions define the inventory of sounds used in human languages. Of particular interest to us is the relation between F_2 and tongue “backness” which has been experimentally observed in the acoustic signals of diphthongs, continuous sounds which begin as one vowel and end as another. When pronouncing back-front diphthongs (for example, /ai/), F_2 increases with time as the tongue moves forward. At ≈ 1400 Hz, F_2 jumps rapidly by around 50–300 Hz [63]. This is caused by a coupling between the oral and subglottal cavities, and occurs near the second subglottal resonance (see Fig. 10). The typical magnitude and location of the jump is predictable using a simple two-tube acoustic model of the cavities [63].

Because quantal effects arise via the map \mathbf{g} from articulatory to acoustic space, we temporarily switch our attention to the distribution of utterances in articulatory space. We consider a two-dimensional articulatory domain, with two dimensions *backness* (y_1) and *height* (y_2). It appears that listeners mentally compensate for the effects on formant values of age and sex so that they perceive utterances of the same phoneme by a small and a large person as essentially the same sound [64]. We therefore assume that both articulatory coordinates $\mathbf{y} = (y_1, y_2)$ and their acoustic counterparts $\mathbf{x} = (x_1, x_2)$ are normalized, so we can think of all speakers as being the same size. Because the utterances of each frame in acoustic space form a cloud $[\psi_i(\mathbf{x})]$, so must the articulatory parameters which generated them. In the definition of our model (Sec. III), we assumed that when uttering the sound in frame i , speakers on average used articulatory parameters

$$\mathbf{y}_i := \mathbf{g}^{-1}(\mathbf{x}_i) \tag{55}$$

with normally distributed variations. Here, \mathbf{g}^{-1} is the inverse of the articulatory to acoustic map. We write the *articulatory cloud*

$$\phi_i(\mathbf{y}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_1 - y_{i1})^2 + (y_2 - y_{i2})^2}{2\sigma^2}\right) \tag{56}$$

$$= \frac{\exp\left(-\frac{(y_1 - y_{i1})^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(y_2 - y_{i2})^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \tag{57}$$

$$:= \phi_{i1}(y_1)\phi_{i2}(y_2), \tag{58}$$

where we have assumed that the units of articulatory parameters are chosen so that the cloud radii σ in the two spaces are the same in those (quantal) regions where the map \mathbf{g} is affine. Since the jump in F_2 is generated by front-back movement of the tongue, parametrized by y_1 , we can write the map \mathbf{g} as

$$\mathbf{g}(\mathbf{y}) = \begin{bmatrix} g_1(y_1) \\ a + y_2 \end{bmatrix}, \tag{59}$$

where $g_1(y_1)$ is a nonlinear function which captures the effect of the subglottal resonance. To determine the shape of the acoustic cloud corresponding to $\phi_i(\mathbf{y})$, we let (Y_1, Y_2) be random variables drawn from ϕ_i . The corresponding acoustic variables are then

$$X_1 = g_1(Y_1), \tag{60}$$

$$X_2 = a + Y_2. \tag{61}$$

Because we have assumed that the articulatory cloud is spherical, having covariance matrix $\Sigma = \sigma\mathbf{I}$, then $Y_1 \perp Y_2$ and also

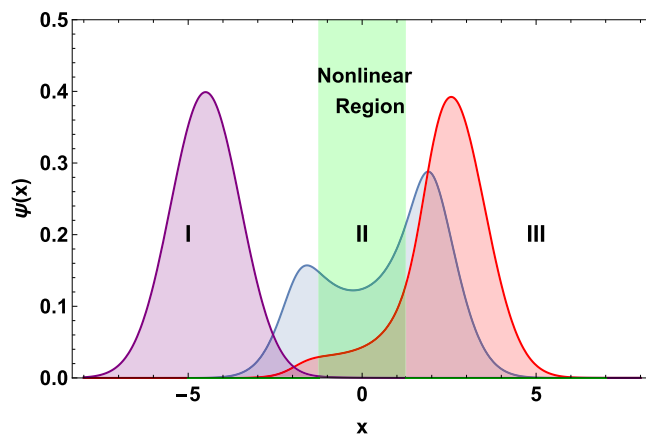


FIG. 18. Distortion of the acoustic frame/phoneme cloud as a subglottal resonance at the origin is approached (red and green distributions). Each cloud is normal in the articulatory parameter (with $\sigma = 1$), and away from the nonlinear region it remains normal (purple cloud). The parameters of g_1 , given by Eq. (65), are $j = 1$, $w = 0.5$.

$X_1 \perp X_2$ so the cloud density in acoustic space also factorizes

$$\psi_i(\mathbf{x}) = \psi_{i1}(x_1)\psi_{i2}(x_2). \tag{62}$$

Because the map \mathbf{g} is linear in the height variable, the acoustic height distribution is simply

$$\psi_{i2}(x_2) = \frac{\exp\left(-\frac{(x_2 - x_{i2})^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}. \tag{63}$$

Changing random variables [55] in the nonlinear case gives the acoustic backness distribution

$$\psi_{i1}(x_1) = \left| \frac{d}{dx} g_1^{-1}(x_1) \right| \phi_{i1}(g_1^{-1}(x_1)). \tag{64}$$

In order to compute this distribution, we use the following explicit form for g_1 (also used to generate Fig. 17):

$$g_1(y_1) = y_1 + j \tanh(x/w), \tag{65}$$

which is a linear function throughout most of its domain, but with a step of height $2j$ and width w , centered at the origin. As $w \rightarrow 0$, the step becomes a sharp discontinuity. Figure 18 shows the effect of the resonance on the acoustic backness distribution. In the quantal regions, away from the nonlinearity, the acoustic cloud remains normal. However, as the nonlinear region is approached, some utterances have articulatory parameters which cross the jump point, creating sounds with extreme acoustic characteristics compared to the average acoustic value of the phoneme. If unusual sounds are rejected by learners, as we have assumed when defining the acceptance probability (23), then the mean of the set of accepted sounds will shift. For a single phoneme composed of colocated frames, assuming that repulsion effects have had sufficient time to isolate it, the acceptance probability for one of its frames i becomes approximately $p_i(\mathbf{x}) = \mathbf{1}_{|\mathbf{x} - \mathbf{x}_i| < R}$, where R is the threshold beyond which sounds are rejected for being too unusual. The backness component of the exemplar

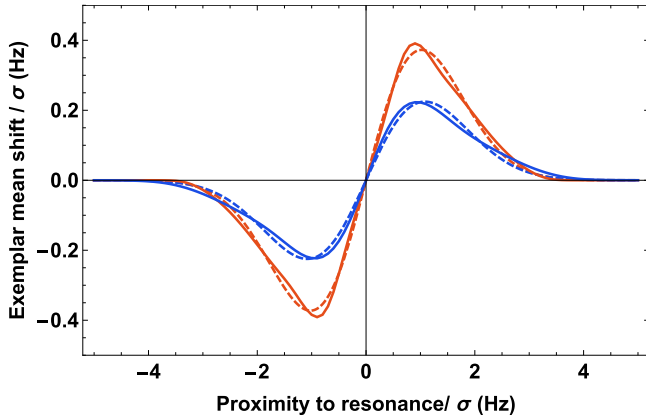


FIG. 19. Solid blue curve shows shift in exemplar mean with threshold $R = 2\sigma$, generated by jump in g_1 of magnitude $j = 1$ and width $w = 0.5$, defined in Eq. (65). Solid red curve shows exemplar shift for a sharper discontinuity ($w = 0.2$). Dashed curves show least-squares fits of approximate quantal force function (68). Fitting parameters: blue, $Q = 0.34$, $\omega_x = 1.53$; red, $Q = 0.60$, $\omega_x = 1.45$. In both cases we have assumed we are at the top of vowel space of $x_2 = x_2^*$.

mean is therefore

$$\bar{x}_{i1} = \frac{\int_{-R}^R x_1 \psi_{i1}(x_1) dx_1}{\int_{-R}^R \psi_{i1}(x_1) dx_1}. \quad (66)$$

The shift of the exemplar mean away from the mean position of the phoneme defines an additional *quantal force*

$$\mathbf{f}_i^{\text{qua}} = \begin{bmatrix} \bar{x}_{i1} - x_{i1} \\ 0 \end{bmatrix}, \quad (67)$$

the backness component of which is plotted in Fig. 19. Here, we see that the discontinuity drives phonemes away from the subglottal resonance. We have also fitted an approximate quantal force curve with backness component

$$f^{\text{qua}} = Q(x_1 - x_1^*) \exp\left(-\frac{(x_1 - x_1^*)^2}{\omega_x^2} - \frac{|x_2 - x_2^*|}{\omega_y}\right), \quad (68)$$

where x_1^* is the position of the resonance in acoustic space and x_2^* is highest point in acoustic space. The height-dependent term in the exponent accounts for the fact that the quantal force requires a wide range of backness ($\approx 6\sigma$) in order to operate (Fig. 18), and so we only expect to see its effects near the top of vowel space. From Fig. 19 we see that this force curve gives a close match to the numerical calculation of the force. Since we do not know the exact form of the nonlinear map g_1 , and because computing the distorted clouds and exemplar mean requires numerical integration, we take (68) as a phenomenological definition of the force induced by the subglottal resonance. Combining this force with the attractive and repulsive components of our phenomenological model we obtain the quantal model with interframe force

$$\mathbf{f}_i = \mathbf{f}_i^{\text{att}} + \mathbf{f}_i^{\text{rep}} + \mathbf{f}_i^{\text{qua}}. \quad (69)$$

In the exemplar dynamics picture, the interframe forces will also be altered by the nonlinear map \mathbf{g} because clouds become stretched as they approach the resonance, increasing the ef-

fective interaction range across it. Because the repulsive effect of the resonance will push phonemes away until their clouds no longer cross it, we assume that enhanced cross-resonance phonemic repulsion may be self-consistently neglected in the steady state.

To examine the effects of the subglottal resonance we reestimate the relative frequencies of each vowel quality in our model when a quantal force is present. In order to estimate the parameters of this force, we note from [63] that the F_2 jump can be up to 300 Hz, which in standard coordinates gives $j \approx 1.5$. The width w of the jump region depends on the properties of the vocal tract. A more sudden jump increases the magnitude Q of the quantal force. Taking $j \in [1, 1.5]$ and $w \in [0.01, 0.5]$ gives force parameters $Q \in [0.5, 2.5]$ and $\omega_x \in [1.3, 2]$ when $\sigma = 1.5$. We have selected $Q = 2$, $\omega_x = 2$. Table V shows the predicted frequencies of each vowel when quantal forces are included. Because quantal forces drive phonemes away from the high central position /i/, we find that the frequency of this phoneme reduces to approximately half its nonquantal value, close to its frequency in real systems. However, the quantal force also appears to remove the mid central vowel in nearly all the systems generated by our sample. This happens because, although initially many systems have a central vowel, it is only stable if repulsive forces from boundary vowels are in balance. If the high central vowel is removed, stability is lost, and the mid central vowel migrates up and out to the system edge. In real vowel systems, the pair /i,ə/ have the fourth highest correlation (Table IV), suggesting that the existence of one creates conditions which make the other more likely to persist. However, the question of what mechanism gives rise to stable central vowels is unresolved by our model. One possibility is that these sounds have low functional load, and are therefore less strongly affected by repulsion. Alternatively, some languages exhibit vowel systems which are in flux [65], and are therefore not equilibria of our dynamics. We address the question of how stable systems can spontaneously change in Sec. VIII.

The origins of quantal theory lie in a theory of sound patterns, developed by Chomsky, Halle, and others [66], where each unit (segment) of speech is characterized by the presence or absence of a set of features [67]. For example, the height of a vowel is described by the two features [\pm high] and [\pm low], where \pm denotes the presence or absence of the feature. The back mid vowel /ɔ/ includes among its features [$-$ high, $-$ low, $+$ back]. Phoneticians were motivated to search for a physical mechanism which could explain why it was possible to construct a successful phonological theory based on features like [\pm back] when, phonetically, backness appears to be a continuous variable [61]. The subglottal resonance provides a phonetic basis for the feature [\pm back] because it divides vowel space into two stable regions. The relative scarcity of high and mid central vowels in real systems is consistent with this explanation, and formant-based studies have shown that the resonance provides a reliable boundary between front and back vowels [61]. Within our model, the second most common vowel system seen in real languages (Fig. 12), of which Italian and Yoruba are examples, requires quantal effects for its long-term stability (Fig. 20). Without this force, the upper mid vowels /e,o/ are only marginally stable, and if stochastic effects cause one of them to migrate

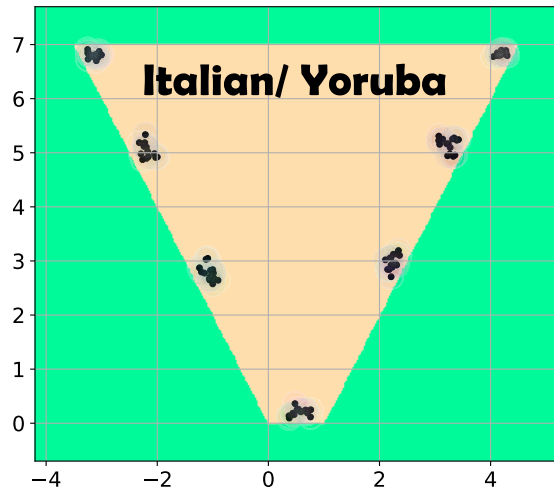


FIG. 20. A phonemic configuration matching the Yoruba and Italian languages [47], which is highly stable in the presence of a quantal force, but only marginally stable without. This is the second most common vowel system cross linguistically, accounting for 13.4% of language surveys in the PHOIBLE database [53]. Parameter values $\alpha = 1$, $\sigma = 1.5$, $D = 0.025$, $\gamma = 0.5$, $Q = 2$, $\omega_x = 2$, $\omega_y = 2$.

away from the boundary of vowel space, then repulsion from the remaining boundary vowels will push it upward to become /i/.

VII. ALLOPHONIC VARIATION

We now consider an extension to the model which allows us to capture allophonic variation within a set of phonemes. A phoneme is a set of sounds which are never *contrastive*. That is, if two sounds are part of the same phoneme, then exchanging them within a frame cannot change the meaning of that frame. The allophones of a phoneme are subcategories of that phoneme which are acoustically distinguished from one another, and are used predictably in different contexts. That is, allophones are in *complementary* distribution with one another. Acoustic parameters are not intrinsically allophonic or phonemic. For example, in Australian English there are minimal pairs which differ only in vowel length as in the words ⟨cut⟩[kʌt] and ⟨cart⟩[kɑːt] implying that /ɜ/ and /ɜː/ are different phonemes, whereas in English received pronunciation (RP) lengthening is an allophonic variation which occurs, for example, when a vowel is followed by a voiced consonant.

We model allophonic variation by defining each acoustic parameter to be either phonemic or allophonic, noting that these definitions may spontaneously change. For simplicity, we consider two parameters: height x (phonemic) and length z (allophonic). The set of vowels sounds which are phonemically equivalent to the sound in frame i are then the templates $S_i = \{k \text{ s.t. } |x_i - x_k| < \alpha\}$. We also define another, similar set based on the allophonic parameter $P_i = \{k \text{ s.t. } |z_i - z_k| < \beta\}$ where β is an allophonic template range. The set of frames which are in both S_i and P_i are then allophonically equivalent

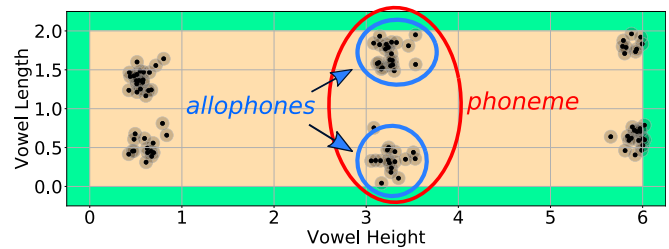


FIG. 21. Simulation of a vertical vowel system where vowel length is the allophonic parameter and $\alpha = 1$, $\beta = 0.5$, $\sigma_x = 1$, $\sigma_z = 0.5$, $D = 0.02$, $\gamma = 1$.

to i . We write this set

$$A_i = S_i \cap P_i. \quad (70)$$

When frame i is uttered, the probability density for the phonemic parameter is, as before, $\psi_i(x)$, whereas we write the density of the allophonic parameter $\chi_i(z)$ which we can assume is Gaussian but with a different variance to ψ_i . The density of phonemic templates for frame i is then defined exactly as in the standard version of the model

$$\hat{\psi}_i(x) := \sum_{j \in S_i} (1 + \omega \delta_{ij}) f_j \psi_j(x), \quad (71)$$

whereas the density of allophonic templates is

$$\hat{\chi}_i(z) := \sum_{j \in A_i} (1 + \omega \delta_{ij}) f_j \chi_j(x). \quad (72)$$

Phonemic and allophonic antitemplate densities $\tilde{\psi}_i$, $\tilde{\chi}_i$ are defined as sums of the summands above, but over S_i^c and $A_i^c \cap S_i$. As with the purely phonemic version of the model, we expect frames to be attracted to the mean of their phonemic templates \hat{x}_i and repelled from phonemic antitemplates. Similar arguments apply to interactions in the allophonic dimension, producing forces of identical form to (13) and (16), with the replacements $\psi_i(x) \rightarrow \chi_i(z)$ and $(S_i, S_i^c) \rightarrow (A_i, A_i^c \cap S_i)$. As a result, in this model we will not see interactions between allophones of different phonemes: if a long allophone of /u/ gets longer, this will not affect a long allophone of /i/.

An example simulation is shown in Fig. 21. Each set of allophones behave as their own subphonemic vowel system within a subspace parametrized by z_i . Using this example, it is possible to see how a phonemic split might occur. Consider the central phoneme in Fig. 21. The existence of the two allophones implies that there is some conditioning factor in the language (e.g., the voicing of a following consonant) which determines the phonemic environments in which each allophone should be used. If this factor disappears from the language, then it will no longer be possible to predict which allophone to use in a given frame. The two allophones are then merely two different sounds in the language. More importantly, after the conditioning factor is lost, their functional load will increase because the contrast brought by the condition factor has gone. They may even distinguish minimal pairs. For this reason, the allophonic parameter switches to being phonemic and we are left with two crowded phonemes, which will mutually repel, generating a phonemic split.

VIII. CHAIN SHIFTS, MOMENTUM, AND WORD-FREQUENCY EFFECTS

So far we have explored the equilibrium behavior of our model. We now show how it provides a simplified view of various processes of sound change observed in real languages. We also provide an extension which allows us to model *self-actuating* changes in vowel systems: significant shifts in state which would be unlikely to result from diffusive dynamics alone.

Beginning from a randomized initial state, our model will evolve over time into a stable system in which repulsive forces between phonemes are near equilibrium. Due to stochasticity, it is in principle possible for one or more phonemes to move sufficiently far away from this equilibrium so that the system enters the basin of attraction of some other equilibrium configuration. A series of further phoneme movements will then ensue, until the new equilibrium is reached. Spontaneous sound changes from one stable state to another also occur in real languages, and if they involve multiple interacting sounds, they are referred to as *chain shifts* [35]. This terminology arises because multiple-vowel linguistic changes often occur in sequence, with the movement of one vowel inducing another nearby to move, and so on. Chain shifts are traditionally divided into two classes: *push chains* and *pull chains*, discussed below. Our model appears unrealistically stable when compared to real languages, generating changes too infrequently. We will show below how sensitivity to an *age vector* [40–42] in the linguistic community can generate more realistic change processes.

A. Push and pull chains

One of the best known examples of a chain shift is *The Great Vowel Shift* which occurred in the English language between the time of Chaucer (1343-1400) and Shakespeare (1564-1616) [68]. The shift affected the front and back long vowels of English, which moved upward in vowel space, with the high vowels /i:/ and /u:/ shifting inward and becoming diphthongs [68]

$$/i:/ \rightarrow /əɪ/, \tag{73}$$

$$/u:/ \rightarrow /əʊ/, \tag{74}$$

where the notation $A \rightarrow B$ indicates that after the shift, the frames which originally contained vowel A , contain vowel B . Two theoretical explanations exist for the mechanism which allowed this to happen. We use these alternatives to illustrate the difference between a push chain and a pull chain without commenting on which is more likely [17]. For simplicity, let us consider only the front vowels. A simplified summary of their movement is as follows:

$$/a:/ \rightarrow /ɛ:/ \rightarrow /e:/ \rightarrow /i:/ \rightarrow /əɪ/. \tag{75}$$

This shift also involved a merger, so the final state of the frames using phonemes /e:/ and /ɛ:/ was /i:/. The *pull chain* explanation is that /i:/ changed first, creating a gap at the top of vowel space, into which the other lower vowels moved. That is, the *leading edge* of the chain moves first. The *push chain* explanation is that the back of the chain shifts first,

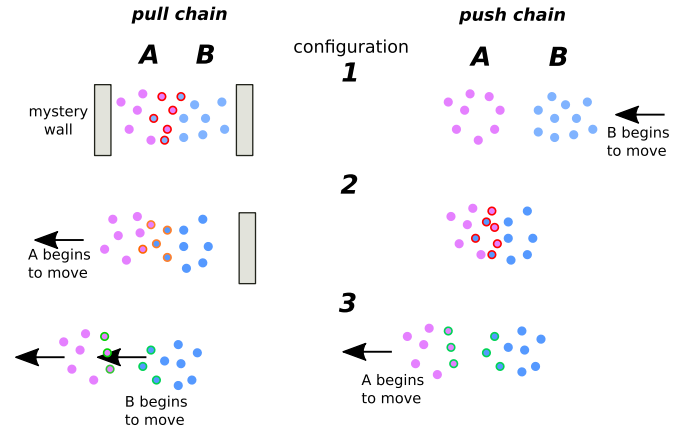


FIG. 22. The mechanics of push and pull chains, adapted from Labov [17]. Pink and blue dots are utterances of two different phonemes, A and B . Dots circled in red have a high probability of rejection as tokens used for learning. Orange circling indicates a lower probability of rejection. Green circle dots are unlikely to be rejected. Mystery walls are required to hold the phonemes together before the push chain begins.

overcrowding the sounds in front, and pushing them forward in sequence.

A learning-based explanation for push and pull (or *drag*) chains has been outlined by Labov [17], and an adapted form of this is summarized in Fig. 22. Consider first the pull chain. In [17], the initial locations of the phonemes A and B (configuration 1) are described as stable, without need for the “mystery walls” that we have added. These walls represent other phonemes or the boundaries of vowel space. In our model the initial configuration would not be stable without these confining elements because the rejection of peripheral tokens would lead to a repulsion effect between A and B . If the wall confining A is removed, then it will move away from B and B tokens that were previously rejected start to be accepted, causing B to begin motion. The fact that the walls are needed in our model reveals a difference between it and Labov’s qualitative description of a pull chain. We require a *release of confinement* in order to produce a pull chain. That is, we must begin with a configuration where vowels are compressed; for example, if there are four front vowels /e,ɛ,e,i/ held in place by repulsive forces from back vowels. If one of these, say /i/, “pops out” into the interior of vowel space, then the repulsive forces between the remaining sounds will push them upward to fill the front positions until interphonemic forces are in balance.

We now consider the push chain. The explanation in [17] begins with configuration 2 of Fig. 22. As we have noted, this configuration is not stable in our model. We therefore assume that our two phonemes are initially separate and form part of some larger stable system. We then assume that something causes phoneme B to start moving (we propose a mechanism below). When B is sufficiently close to A , token rejection effects (repulsion) cause phoneme A to start moving away. The phoneme B is then said to have *pushed* phoneme A .

Within our description of vowel dynamics, the distinction between push and pull chains is to some extent unnecessary. Both processes may be understood as the effects of repulsive

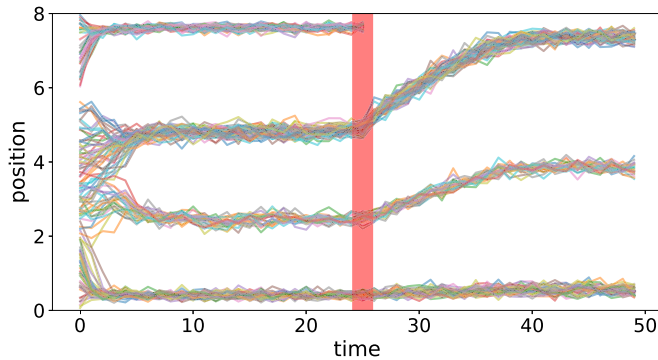


FIG. 23. An example of what is commonly called a pull or “drag” chain. Starting from randomized frame locations, four phonemes spontaneously formed. The highest is removed from the system at $t = 25$. Parameter values $\alpha = 1$, $\sigma = 1.5$, $\gamma = 1$, $D = 0.025$. One time unit = a single speaker’s expected lifetime. No momentum.

interaction forces, but in different contexts: one in which compression is released, and one in which it is applied. In our view there is really no such thing as a pull chain; vowel shifts are all caused by pushing, and it is the start of a shift that determines whether it is referred to as a push or a pull.

In Fig. 23 we have illustrated a pull chain in a one-dimensional vowel space, which may be viewed as a toy model of the consequence of Labov’s *upper exit principle* [65], which we paraphrase as “in chain shifting high peripheral vowels become nonperipheral.” In other words, high front vowels (/i,i:/) have a tendency to pop out of the top left corner of vowel space, as perhaps happened at the start of the great vowel shift. We may view the one-dimensional space in Fig. 23 as a simple model of the front of two-dimensional vowel space, in which phonemes are confined by the repulsion of other nonfront vowels (see Fig. 13). The four initial vowels in this space formed spontaneously, starting from randomized initial frame positions. At $t = 25$ lifetimes, we removed the highest vowel, mimicking a spontaneous shift of the form /i:/ \rightarrow /e:/ . Repulsive forces between the remaining three then pushed them upward to fill out the empty space. The traditional interpretation would be that the movement of the top vowel *dragged* the lower vowels upward.

B. Momentum

Stochasticity of frame trajectories derives in our model from unpredictability in the language-learning process, which at the population level is realized as a diffusion process for frame location. We will show in Sec. IX that beyond a critical level of diffusivity D_c , phonemic clusters cannot form. Assuming $D < D_c$, the motion of the centroid of a phoneme consisting of N frames will undergo a diffusion process with coefficient $D^{\text{pho}} \propto N^{-1}$. The magnitude of D^{pho} also depends on the distribution of relative frame frequencies within the phoneme, with heavier tailed distributions leading to greater diffusivity, due to the dominance a smaller number of very common words. This purely diffusive behavior is problematic from the point of view of empirical observations of sound change, which often self-actuate before progressing monotonically (or nearly so) over a sustained period [42].

One explanation is that language learners are sensitive to the direction of change. This direction is observable from differences in language use between older and younger speakers, often referred to as the *age vector* of a linguistic feature, or its *momentum*. A number of studies have sought to model such effects [40–42] in terms of the relative frequency of a linguistic feature. Although their mathematical details differ, the essential idea is that when (new) speakers select their linguistic state, they are biased in the direction of the age vector.

It is straightforward to incorporate the momentum effect into our model. Working with a one-dimensional acoustic space, where the position of the i th frame is given by $x_i(t)$, we define the linguistic memory of the community for frame i as a time average over its history

$$m_i(t) = \frac{1}{\tau} \int_{-\infty}^t x_i(s) e^{(s-t)/\tau} ds. \quad (76)$$

When $\tau = 1$ this memory is an average over the historical states of the community when each speaker was born. Differentiating with respect to time we obtain

$$\dot{m}_i(t) = \frac{1}{\tau} [x_i(t) - m_i(t)]. \quad (77)$$

We define the difference between the current state and the memory as the *age vector*

$$\Delta_i(t) := x_i(t) - m_i(t). \quad (78)$$

We then define an additional *momentum response*

$$\psi(\Delta) = \theta \tanh\left(\frac{b\Delta}{\theta}\right) \quad (79)$$

which is added to the attractive (13) and repulsive (16) forces already in our model. This is the shift in the average sound learned by a new speaker, based on their tendency to emphasize “younger” forms of speech. An analogous response function appears in the frequency-based models [40,41], where it is termed a *prediction function* [40] or generates a *perceived frequency* [41]. The parameter b in (79) is the *momentum sensitivity* and θ is the *cutoff*, giving the maximum possible magnitude of the momentum driven rate of change. For $\Delta \ll \theta$ the momentum response is approximately linear $\psi(\Delta) \sim b\Delta$ as $\Delta \rightarrow 0$. The momentum model for a single frame may be written

$$dx_i = [f_i^{\text{att}} + f_i^{\text{rep}} + \psi(x_i - m_i)]dt + \sqrt{2D} dW_i, \quad (80)$$

$$dm_i = \frac{1}{\tau}(x_i - m_i)dt. \quad (81)$$

We now analyze the stability of a single isolated phoneme consisting of N frames of equal frequency with self-focus $\omega = 0$. Formally, we consider the limit $\alpha \rightarrow \infty$, so the phoneme does not lose frames through evaporation. The template mean and the *template memory mean* are

$$\hat{x}(t) = \frac{1}{N} \sum_{i=1}^N x_i(t) =: \langle x_i \rangle, \quad (82)$$

$$\hat{m}(t) = \frac{1}{N} \sum_{i=1}^N m_i(t) =: \langle m_i \rangle, \quad (83)$$

where $\langle \cdot \rangle$ denotes the average over frames (the *phoneme average*). From the definition of the momentum model we have

$$d\hat{x} = \frac{1}{N} \sum_{i=1}^N [(\hat{x} - x_i)dt + \psi(x_i - m_i)dt + \sqrt{2D}dW_i] \quad (84)$$

$$= \langle \psi(\Delta_i) \rangle dt + \sqrt{\frac{2D}{N}} d\hat{W}, \quad (85)$$

where \hat{W} is a standard Brownian motion. The analogous equation for the template memory mean is

$$d\hat{m} = \frac{1}{\tau}(\hat{x} - \hat{m})dt. \quad (86)$$

Subtracting (86) from (85) we obtain the age vector dynamics

$$d\hat{\Delta} = \left[\langle \psi(\Delta_i) \rangle - \frac{\hat{\Delta}}{\tau} \right] dt + \sqrt{\frac{2D}{N}} d\hat{W}. \quad (87)$$

Expanding the momentum-response function about $\Delta_i = 0$ we obtain to order $\langle \Delta_i^3 \rangle$

$$d\hat{\Delta} = \left[\left(b - \frac{1}{\tau} \right) \langle \Delta_i \rangle - \frac{b^3}{3\theta^2} \langle \Delta_i^3 \rangle \right] dt + \sqrt{\frac{2D}{N}} d\hat{W}. \quad (88)$$

Writing $\Delta_i = \hat{\Delta} + \epsilon_i$ where ϵ_i is a zero-mean random variable we have, if ϵ_i is also symmetric (so $\langle \epsilon_i^3 \rangle \approx 0$),

$$\langle \Delta_i^3 \rangle = \hat{\Delta}^3 + 3\hat{\Delta} \langle \epsilon_i^2 \rangle \quad (89)$$

so

$$d\hat{\Delta} = \left[\left(b - \frac{1}{\tau} - \frac{b^3}{\theta^2} \langle \epsilon_i^2 \rangle \right) \hat{\Delta} - \frac{b^3}{3\theta^2} \hat{\Delta}^3 \right] dt + \sqrt{\frac{2D}{N}} d\hat{W}. \quad (90)$$

For a phoneme with a large number of frames the drift term dominates the dynamics and we see that provided

$$b - \frac{1}{\tau} - \frac{b^3}{\theta^2} \langle \epsilon_i^2 \rangle < 0, \quad (91)$$

then the age vector has a stable fixed point at $\hat{\Delta} = 0$ so the phoneme will be subject only to a weak diffusion with $D^{\text{pho}} = D/N$. In this case, we will not see any sustained movement in one direction. At a critical value of age vector sensitivity, the fixed point destabilizes and two stable fixed points appear at

$$\hat{\Delta} = \pm \sqrt{\frac{3\theta^2}{b^3} \left(b - \frac{1}{\tau} - \frac{b^3}{\theta^2} \langle \epsilon_i^2 \rangle \right)}. \quad (92)$$

The phoneme will select one of these at random, and then execute a sustained movement in that direction, until noise

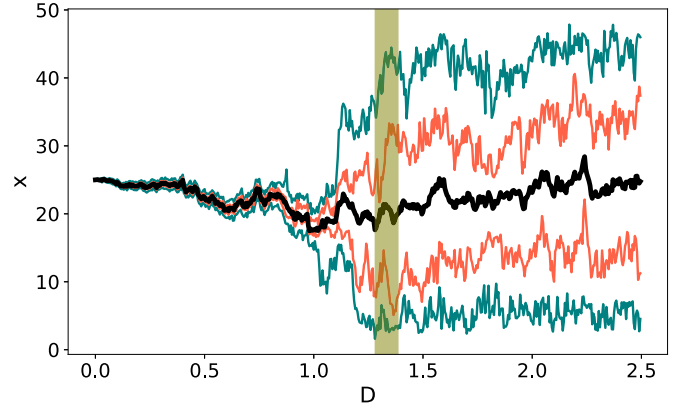


FIG. 24. Black line shows mean location of 100 frames, starting from position $x_i = 15$, as diffusion coefficient is increased from zero over $T = 5000$ lifetimes. System size $L = 30$. Red and blue-green lines show 70th (30th) and 90th (10th) percentile frame locations, which indicate distribution of frames within the system. Vertical olive band shows theoretical location of critical diffusion coefficient $D_c = \alpha^2/3$ where phoneme disintegrates. Other parameter values $\alpha = 2$, $\sigma = 1$, $\gamma = 0$.

effects, interactions with other phonemes, or with the boundaries of vowel space cause it to change direction or return to stability. If sensitivity to the age vector fluctuates with time, or if its value is near the threshold for stability, then the phoneme may switch between periods of stability and instability. We note that it is in principle possible to measure the age vector within a speech community by considering differences in the vowel systems of old and young speakers, and to measure sensitivity, given sufficient longitudinal data.

To illustrate the effect of momentum we first consider the behavior of a single phoneme driven entirely by diffusive dynamics ($b = 0$), without any repulsive effects ($\gamma = 0$). We consider a one-dimensional vowel space $x_i \in [0, L]$, beginning with all frames located at the same location $x = L/2$. Starting from zero diffusion coefficient, we gradually increase D over a long time interval ($T = 5 \times 10^3$ lifetimes), yielding the behavior shown in Fig. 24. While the diffusion coefficient is less than the critical value D_c , the phoneme remains intact and of finite size. Diffusive changes in position during this phase of evolution generate a cumulative shift of ≈ 5 cloud radii over the first 2×10^3 lifetimes. Such a shift corresponds approximately to a vowel changing from “high” to “low” (e.g., /i/ \rightarrow /æ/). Taking a single lifetime as 50 years then this shift would take around 1×10^4 years to complete. This timescale is unrealistic. For example, the great vowel shift took place over ≈ 300 years [68]. Moreover, the template range used in Fig. 24 is at the upper end of realistic, allowing larger diffusion coefficients to be reached before phonemic destruction. As the destruction point is approached, the phoneme expands in size, before disintegrating entirely, leaving frames distributed approximately uniformly over the system. We conclude that diffusive dynamics alone is not sufficient to describe the pace of realistic sound changes.

We now consider the effect of momentum. In Fig. 25 we have simulated a two-vowel system, with $\tau = 1$, where momentum sensitivity fluctuates over time. We model these

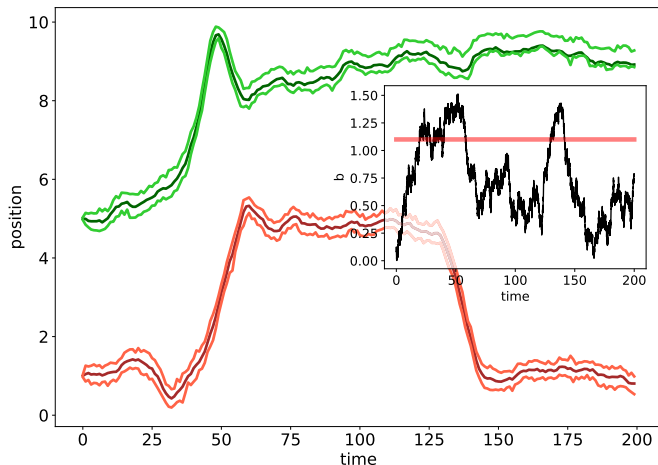


FIG. 25. Darker green and red lines show locations of two phonemes starting from positions $x = 1, 5$ in a system of size $L = 10$. Lighter lines give 90th (10th) percentile frame locations for these two phonemes, as an indication of size. Inset plot shows age vector sensitivity as a function of time. Horizontal red line shows estimated critical sensitivity $b \approx 1.1$. Model parameter values $\tau = 1$, $\alpha = 1$, $\sigma = 1$, $\gamma = 1$, $D = 0.025$, $\theta = 0.5$. Parameters for mean reversion (Ornstein-Uhlenbeck) process followed by momentum sensitivity are $b^* = 0.9$, $a = 0.05$, $\sigma_b = 0.1$.

fluctuations using Ornstein-Uhlenbeck dynamics [32]

$$db_t = a(b^* - b_t)dt + \sigma_b dW_t, \quad (93)$$

where b^* is the long run mean sensitivity, a is the reversion rate toward to this mean, and σ_b is the volatility of the sensitivity. We set $b^* = 0.9$, which is below the threshold for spontaneous sustained shifts. Initially, the two phonemes are four cloud radii apart, producing very weak interactions. When the sensitivity crosses the critical threshold, the lower vowel spontaneously begins rising, starting a push chain. The motion of the green vowel is halted by the boundary of vowel space, and the vowels enter a temporary equilibrium while the sensitivity again becomes subcritical. A final move of the lower vowel is generated by a short-lived supercritical period of sensitivity. We calculate the threshold b_c by first numerically estimating the variance of the age vector

$$\langle \epsilon_i \rangle \approx 0.07 \quad (94)$$

and then setting the right-hand side of (91) to zero, and solving for $b_c \approx 1.1$. The two major shifts generated in this simulation took ≈ 10 lifetimes, corresponding to ≈ 500 years, in line with the great vowel shift.

The one-dimensional momentum model we have defined here is too simple to be directly comparable to the great vowel shift because, in its current form, it does not describe different lengths of vowel or diphthongs. However, we can reproduce the essential properties of observed systems, serving a starting point for more sophisticated models which could be used to test hypotheses about the nature of historical vowel shifts, and to predict future changes.

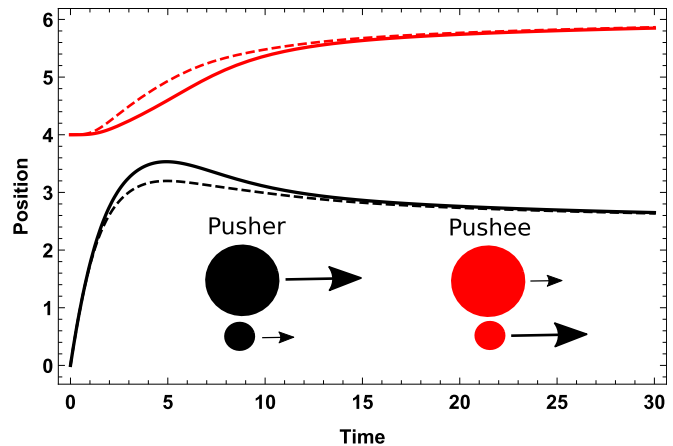


FIG. 26. The interaction of two phonemes (black and red), each with two frames. Solid lines: high-frequency frames ($f = 0.45$); dashed lines: low frequency ($f = 0.05$). Parameter values $\sigma = 1$, $\omega = 5$, $\gamma = 1$, $\tau = 1$. Initial age vector of black frame $\Delta = 2$. Momentum-response parameters $b = 0.7$, $\theta \rightarrow \infty$.

C. Word-frequency effects

Exemplar theory and empirical studies suggest [24] that high-frequency frames change more slowly than low-frequency frames when pushed by the movement of another vowel (a *push chain* [17]). The opposite occurs in the frames which are being pushed. We consider a system of two phonemes, each consisting of one high- and one low-frequency frame. Initially, the phonemes are separated by 4σ with one phoneme having an age vector $\Delta > 0$ in the direction of the other, with subcritical momentum sensitivity. This creates a simple two-frame push chain. Figure 26 shows the dynamics of the high- and low-frequency frames as they approach, in the noiseless limit $D \rightarrow 0$. We see that in the incident frame (the “pusher”) the high-frequency frame moves faster, whereas in the “pushee” frame, the high-frequency frame is slower to react. The effects are quite subtle, as they are in empirical data [69], and we note that the differing response of high- and low-frequency frames requires positive self-focus. Without this, the data used by learners are the same for all frames in a phoneme, so the dynamics of each frame is statistically identical. The magnitude of word-frequency effects therefore provides a mechanism to infer the extra weight that listeners place on words as templates for the vowel sounds they contain, as compared to words which contain similar sounds. Finally, we note an intuitive physical analogy. When $\omega > 0$, higher-frequency frames behave as more massive particles which are less strongly influenced by the proximity of others.

IX. CONTINUUM LIMIT

So far we have considered languages with relatively small numbers of frames, each of which can be explicitly simulated. We now consider the many-word limit where system dynamics may be described in terms of a continuum frame density. Let N be the total number of frames in our language and consider the limit $N \rightarrow \infty$ in one dimension, in which case

we can define a frame density

$$\rho(x) = \lim_{\delta \rightarrow 0} (1/\delta) \sum_i f_i \mathbf{1}_{|x-x_i| < \delta/2}. \quad (95)$$

We suppress the time dependence of ρ for compactness. The set of frames which are templates for a frame located at x lie in the *template interval* $T_x := [x - \alpha, x + \alpha]$. In one dimension the total force on frame i , given the positions of the other frames, may be written

$$\mu(x_i | \{x_j\}_{j \neq i}) = \hat{x}_i - x_i + \sum_{j \notin S_i} \frac{f_j \psi_j(x_i)(x_i - x_j)}{\gamma^{-1} \hat{\psi}_i(x_i) + \tilde{\psi}_i(x_i)}, \quad (96)$$

where we have used the symbol μ to avoid confusion with the frequency of frame i . In the limit $N \rightarrow \infty$ we have, for the attractive component of this force,

$$\hat{x}_i - x_i \rightarrow \frac{\int_{T_{x_i}} (u - x_i) \rho(u) du}{\int_{T_{x_i}} \rho(u) du}, \quad (97)$$

and for the repulsive component

$$\begin{aligned} & \sum_{j \notin S_i} \frac{f_j \psi_j(x_i)(x_i - x_j)}{\gamma^{-1} \hat{\psi}_i(x_i) + \tilde{\psi}_i(x_i)} \\ & \rightarrow \frac{\rho(x) \int_{T_{x_i}^c} (x_i - u) \psi_u(x_i) \rho(u) du}{\gamma^{-1} \int_{T_{x_i}^c} \psi_u(x_i) \rho(u) du + \int_{T_{x_i}^c} \psi_u(x_i) \rho(u) du}, \end{aligned} \quad (98)$$

where $\psi_u(x)$ is the $\mathcal{N}(u, \sigma^2)$ density; we consider the case $\sigma = 1$ and set the self-focus $\omega = 0$ for simplicity. In the limit $N \rightarrow \infty$ we write the force on a frame at x as $\mu(x)$, with the dependence on ρ implicit. In this limit the density evolves deterministically provided the relative frequencies of all frames tend to zero as $N \rightarrow \infty$. Conditional on the density field, the location of frame i obeys the stochastic differential equation

$$dx_i = \mu(x_i) dt + \sqrt{2D} dW_i \quad (99)$$

and therefore the probability density function, $p_i(x, t)$, for its location obeys the Fokker-Planck equation [33]

$$\partial_t p_i(x, t) = -\partial_x [\mu(x) p_i(x, t)] + D \partial_x^2 p_i(x, t). \quad (100)$$

To find the density field $\rho(x)$ we note that because it is deterministic

$$\rho(x) = \mathbb{E}(\rho(x)) \quad (101)$$

$$= \lim_{\delta \rightarrow 0} (1/\delta) \sum_i f_i \mathbb{E}(\mathbf{1}_{|x-x_i| < \delta/2}) \quad (102)$$

$$= \lim_{\delta \rightarrow 0} (1/\delta) \sum_i f_i p_i(x, t) \delta \quad (103)$$

$$= \sum_i f_i p_i(x, t). \quad (104)$$

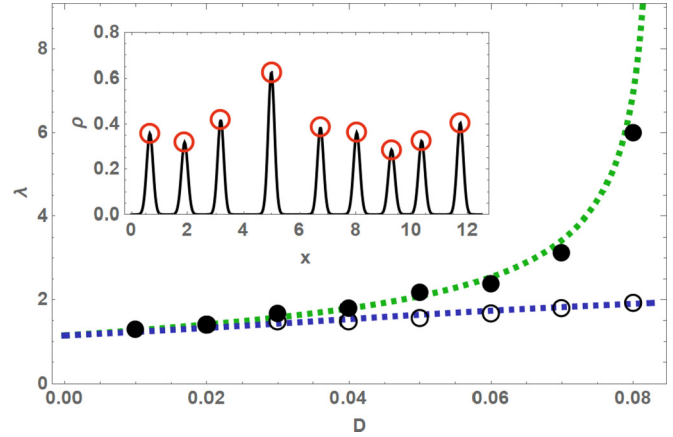


FIG. 27. Dots: estimated λ values (obtained by automated peak identification) when $\alpha = 0.5$, $\gamma = 0$. Circles: $\gamma = 1$. Initial fluctuations given by high-frequency sine wave. Dashed lines: analytical predictions. Inset: solution to (105) for $\gamma = 0.5$, $D = 0.02$, giving $\lambda = 1.39$ (cf. prediction $\lambda^* = 1.33$).

Multiplying both sides of (100) by f_i and summing over i we obtain

$$\begin{aligned} \partial_t \rho(x) = & D \partial_x^2 \rho(x) - \partial_x \left(\frac{\rho(x) \int_{T_x} (u - x) \rho(u) du}{\int_{T_x} \rho(u) du} \right) \\ & - \partial_x \left(\frac{\rho(x) \int_{T_x^c} (x - u) \psi_u(x) \rho(u) du}{\gamma^{-1} \int_{T_x^c} \psi_u(x) \rho(u) du + \int_{T_x^c} \psi_u(x) \rho(u) du} \right). \end{aligned} \quad (105)$$

This is the continuum evolution equation which defines our model in the limit of large numbers of words. We have made use of the fact that the contribution of individual frames to the density field is negligible so $\mu(x)$ is the same function for all frames.

Beginning from small fluctuations in frame density, solutions to Eq. (105) take the form of regularly positioned peaks representing distinct vowels (inset Fig. 27). Expanding $\rho(u)$ to second order in the attractive term, and comparing to the magnitude of the diffusive flux, we obtain the lower bound $\alpha > \sqrt{3D}$ on template range for vowel formation. This relationship was tested in Fig. 24.

We can calculate the typical number of spontaneously formed vowels in a system by finding the wave number at which density fluctuations are most susceptible to cluster formation. To achieve this, we write Eq. (105) in terms of diffusive, attractive, and repulsive fluxes

$$\partial_t \rho(x) = -\partial_x [j_D(x) + j_A(x) + j_R(x)] \quad (106)$$

and assume that our vowel system begins from a primordial state where the distribution of frames is approximately uniform with fluctuations which are spatially correlated only over very short ranges (no clusters of significant size). We write the state of the system at this early stage

$$\begin{aligned} \rho(x) &= c + \epsilon \eta(x, t) \\ &= c + \frac{\epsilon}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\eta}(\omega, t) e^{i\omega x} d\omega, \end{aligned} \quad (107)$$

$$= c + \frac{\epsilon}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\eta}(\omega, t) e^{i\omega x} d\omega, \quad (108)$$

where $c > 0$ is a constant and $\eta(x, t)$ is a zero-mean fluctuating field with Fourier transform $\hat{\eta}(\omega, t)$ and $\epsilon \ll 1$ is a small parameter measuring the magnitude of fluctuations. At small times t , the spatial correlation function

$$C(z, t) = \int_{-\infty}^{\infty} \eta(x, t)\eta(x+z, t)dx \quad (109)$$

decays rapidly with increasing $|z|$, and therefore has an energy spectrum $\sqrt{2\pi}|\hat{\eta}(\omega)|^2$ whose dominant contributions are from high wave numbers ω [33]. As the system evolves and phonemic clusters form, the energy spectrum will begin to concentrate at lower wave numbers, and spatial correlations will decay more slowly. The peak ω^* of the energy spectrum gives the typical separation of clusters as

$$\lambda^* = \frac{2\pi}{\omega^*}. \quad (110)$$

To estimate the location of the peak we consider the behavior of the diffusive, attractive, and repulsive fluxes as the clusters begin to form. To lowest order in ϵ we have, from (105), as $\epsilon \rightarrow 0$

$$j_D(x) \sim -\epsilon D \eta'(x), \quad (111)$$

$$j_A(x) \sim \frac{\epsilon}{2\alpha} \int_{T_x} (u-x)\eta(u)du, \quad (112)$$

$$j_R(x) \sim \frac{\epsilon\gamma \int_{T_x^c} (x-u)\phi(x-u)\eta(u)du}{(1-\gamma)\text{erf}(\alpha/\sqrt{2}) + \gamma}, \quad (113)$$

where ϕ is the standard normal probability density function. Substituting the Fourier representation of η into these fluxes allows us to write the total flux $j = j_D + j_A + j_R$ as

$$j(x) = \frac{i\epsilon}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\eta}(\omega, t) f(\omega) e^{i\omega x} d\omega, \quad (114)$$

where

$$f(\omega) = -D\omega + \frac{\sin(\alpha\omega) - \alpha\omega \cos(\alpha\omega)}{\alpha\omega^2} - \frac{\gamma}{(1-\gamma)\text{erf}(\frac{\alpha}{\sqrt{2}}) + \gamma} \int_{T_0^c} v\phi(v) \sin(v\omega)dv. \quad (115)$$

The final integral term may be evaluated in terms of error functions of a complex argument [70], but the integral representation is more compact and easier to interpret. Substituting expression (114) for the flux into the continuum evolution equation (106) we obtain the following ordinary differential equation for the transform of the fluctuations

$$\partial_t \hat{\eta}(\omega, t) = \omega f(\omega) \hat{\eta}(\omega, t) \quad (116)$$

which has solution

$$\hat{\eta}(\omega, t) = \hat{\eta}(\omega, 0) \exp[\omega f(\omega)t]. \quad (117)$$

The decoupling of Fourier modes which allowed this solution [71] is a consequence of the linearization of the fluxes in (111), (112), and (113), so the solution is only valid for small fluctuations. However, because the locations and sizes of clusters (peaks in ρ) are decided early in the evolution of the system, the wavelength λ^* corresponding to the fastest

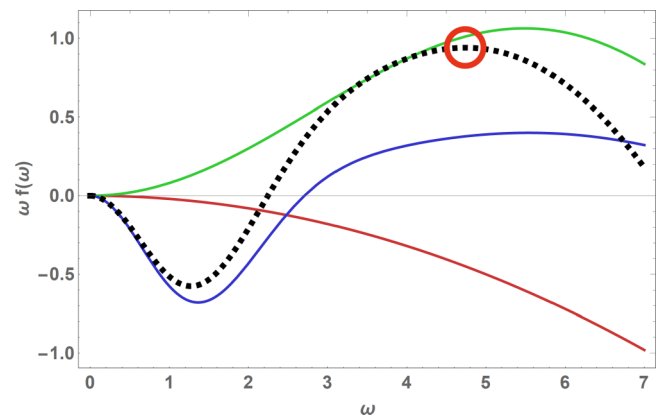


FIG. 28. Contributions to the net mode growth rate from diffusion (red), attraction (green), and repulsion (blue), when $D = 0.02$, $\alpha = 0.5$, $\gamma = 1$. Black dotted curve shows overall mode growth rate $\omega f(\omega)$ with maximum circled.

growing wave number ω^* gives an accurate approximation to the wavelength of the final peak distribution. From (117) we see that fluctuations with wave numbers for which $f(\omega) < 0$ decay over time, whereas those for which $f(\omega) > 0$ grow, with the fastest growing wave number ω^* corresponding to the maximum of $\omega f(\omega)$:

$$\omega^* = \arg \max_{\omega} \omega f(\omega). \quad (118)$$

This prediction is compared to simulations in Fig. 27. In the small noise limit $D \rightarrow 0$ we have $\lambda \rightarrow 2.29\alpha$, so we expect vowels to be separated by just over twice the maximum separation which speakers consider to be phonemically equivalent. The threshold for first formant frequency discrimination in normal speech is $\Delta F_1 \approx 50 \pm 10$ (Hz) [58], where $F_1 \in [235, 850]$ [59]. Taking ΔF_1 as a lower bound on α gives a theoretical maximum of six spontaneously formed vowel heights, consistent with observed systems [13].

In Fig. 28 we have plotted the separate contributions from the diffusive, attractive, and repulsive fluxes to the overall growth rate $\omega f(\omega)$ of Fourier modes. Here, we see that attractive forces cause all modes to grow, corresponding to increasing amplitude fluctuations at all length scales as frames are pulled together producing higher density peaks. The diffusive flux reduces (smoothes out) fluctuations at all length scales, but has greatest effect on high-frequency (short length scale) fluctuations because it penalizes the second derivative of the density distribution. The behavior of the repulsive flux is more subtle: for low wave-number fluctuations (large clusters) it has a negative effect: if the typical repulsion range is smaller than the size of a cluster, then it will break up the cluster into smaller parts. At larger wave numbers (corresponding to smaller clusters), repulsion has a positive effect on growth: when the clusters are smaller, the dominant repulsion effect is a squeeze from near neighbors, increasing the density of all clusters.

X. DISCUSSION

We have presented a model of the evolution of vowel systems in which words behave as interacting particles

diffusing in vowel space, with positions determined by the vowel sounds they contain. The interactions between particles are modeled mathematically as physical forces, but are derived based on our understanding of the language-learning process. Our work builds on an already substantial collection of models of vowel system evolution [13,14,18–24,27,30], so it is important to make clear what we are adding. Because the mechanisms which control the evolution of sound inventories are not fully understood, existing models have invoked a considerable range of mechanisms in order to explain observations. Many of these mechanisms induce a combination of attractive and repulsive interactions between vowel sounds. Since the precise linguistic, social, and cognitive processes which drive sound change remain unknown, it is difficult to justify anything other than a phenomenological approach, that is, one which is consistent with existing theories but does not rely on a precise scientific hypothesis regarding the underlying mechanisms. Under these conditions, simplicity is beneficial. By defining our model using the language of physics (forces and diffusion), we simplify the conceptual picture of vowel system dynamics, and allow for a more transparent and thorough analysis of how the system parameters control predictions. Moreover, these parameters may in some cases be directly measured or bounded (cloud radius, template range) or their values inferred from data (diffusion coefficient, momentum sensitivity).

The simplicity of our model definition has allowed us to combine quantal and dispersion theory into a single framework [11,43,61] to capture allophonic variation and elucidate the mechanism of phonemic splitting [35], to explore the effects of functional load [48], model self-actuating sound changes [40–42,65], provide a simplified picture of chain shifting in which the distinction between drag and push chains becomes redundant [17], explain empirically observed word-frequency effects in sound change [24], predict the critical level of stochasticity at which phonemes disintegrate, predict the maximum number of vowel heights a language can contain [12,19], and provide a simple picture of the process of phonemic merger. Its simplicity also allows for efficient simulation and mathematical analysis. Because it is derivable by considering the behavior of individual speakers, it is in principle straightforward to extend the definition to interacting communities and social networks. Because the atomic constituents of the model are individual words, it is also in principle possible to use our approach to model specific cross-linguistic interactions such as borrowing, where new phonemes are created by the inclusion of foreign words into a language (e.g., from French to English [35]).

Our scientific conclusions are as follows. In its simplest form, our model may be seen as a *dynamical* dispersion theory [31], where repulsive interphoneme forces drive vowel systems toward configurations which maximize contrast. By comparison to a large database of phonemic inventories, and by defining of our own system typology, we have shown that the model captures cross-linguistic relative frequencies of different sounds to within $\approx 10\%$, with the exception of the high central vowel /i/. Moreover, for vowel systems up to cardinality six, the most common systems generated by the model are consistent with our typology and that of Crothers [52]. The over-representation of the high central vowel is seen

in other (nondynamic) dispersion theories [13], and attempts have been made in the past to correct these predictions by defining sophisticated perceptual distance metrics [19] which effectively warp the shape and structure of vowel space. By their nature these metrics are difficult to rigorously derive and test. In contrast, the dynamics in our model is driven by phonemic overlap in acoustic space, which may be quantitatively measured using formants, fixing the aspect ratio of vowel space unambiguously.

Quantal theory provides an alternative to dispersion theory for understanding the structure phonetic inventories. We have shown that the two theories combine in our simple framework, with quantal effects entering via the articulatory to acoustic map [11,39,43] which exhibits a pronounced nonlinearity at the second subglottal resonance. We have shown that this induces a repulsive force away from the center of vowel space, reducing the relative frequency of /i/, and stabilizing the second most common vowel system (the seven-vowel system of Italian and Yoruba). However, the quantal force also destabilizes the mid central vowel, and we can only speculate as to the mechanisms which could counteract this effect. We note, however, that in real languages /i/ and /ə/ are positively correlated, suggesting that the presence of one may increase the stability of the other.

Moving beyond the overpreponderance of /i/, examination of cross-linguistic correlations between phonemes (Table IV), and of the most common empirically observed vowel systems (Fig. 12) reveals that vowels tend to occur in front-back pairs of the same height. This symmetry is present, but to a lesser extent, in our predictions. In dispersion theory, positive correlations between vowels of the same height occur because such configurations happen to maximize contrast. Our results suggest in real vowel systems there may be some additional mechanism which imposes this symmetry. We will address this point in further work.

Sporadic sound changes [72], often involving multiple vowels, are well documented [35,65]. However, the long-standing *actuation problem* remains unresolved: “Why do changes in a structural feature take place in a particular language at a given time, but not in other languages with the same feature, or in the same language at other times” [73]? One possible resolution to the problem is *momentum-based selection* [40–42,74], according to which speakers react to features whose relative frequency has risen in the recent past, by further emphasizing the use of these features. In the context of sound change, the analog of an increase in frequency is a net direction of change with time, which may be realized in *apparent time* [35] as a difference between the behavior of old and young speakers. Inspired by frequency-based models [40–42] we have incorporated momentum-based change in our model, showing that when sensitivity to this change exceeds a critical threshold, long-term sustained shifts in vowel systems can take place. We find that without such an effect, natural variations in speaker’s articulatory behavior, which can be substantial [75], and are captured by our diffusion coefficient D , are not sufficient to actuate sustained shifts over realistic timescales, while still preserving the integrity of phonemes (Fig. 24). Our model of momentum-based change generates shifts over $O(10^1)$ lifetimes, with variations depending on the level of sensitivity, consistent with the durations

of observed changes such as the Great English Vowel Shift [68]. In the deterministic setting ($D = 0$), artificially activated push chains, described by a system of ordinary differential equations, show that high-frequency words behave as more massive particles, responding more slowly to the encroachment of a phoneme (Fig. 26), provided that the self-focus is positive; that is, when learning how to pronounce the vowel sound in a word, learners place more emphasis on the word itself than on other words containing similar vowel sounds.

Our model is by no means a perfect description of vowel system structure and dynamics. It is best described as a *toy model*. However, due to its simplicity and flexibility we have been able to use it to study and understand a considerable

variety of linguistic processes, providing a simple mathematical picture which we hope may be a useful tool for understanding the evolution of the sounds of languages. The model may provide a framework for future predictive modeling approaches which use large formant data sets of phonemic inventories to directly calibrate dynamical models.

ACKNOWLEDGMENT

The authors are grateful to the Royal Society for an APEX award (2018–2020), Grant No. APX\R1\180117, supported by the Leverhulme trust.

-
- [1] P. Roach, *English Phonetics and Phonology* (Cambridge University Press, Cambridge, 2009).
- [2] D. Fogerty and D. Kewley-Port, Perceptual contributions of the consonant-vowel boundary to sentence intelligibility, *J. Acoust. Soc. Am.* **126**, 847 (2009).
- [3] F. Chen, L. L. N. Wong, and E. Y. W. Wong, Assessing the perceptual contributions of vowels and consonants to mandarin sentence intelligibility, *J. Acoust. Soc. Am.* **134**, EL178 (2013).
- [4] K. Stevens, *Acoustic Phonetics* (MIT Press, Cambridge, MA, 2012).
- [5] <http://www.internationalphoneticassociation.org/content/ipa-chart>.
- [6] G. E. Peterson and H. Barney, Control methods used in a study of the vowels, *J. Acoust. Soc. Am.* **24**, 175 (1952).
- [7] T. Arai, Education in acoustics and speech science using vocal-tract models, *J. Acoust. Soc. Am.* **131**, 2444 (2012).
- [8] J. Wang, J. R. Green, A. Samal, and Y. Yunusov, Articulatory distinctiveness of vowels and consonants: A data-driven approach, *J. Speech Lang. Hear. Res.* **56**, 1539 (2013).
- [9] J. Lee, S. Shaiman, and G. Weismer, Relationship between tongue positions and formant frequencies in female speakers, *J. Acoust. Soc. Am.* **139**, 426 (2016).
- [10] G. Birkoff and S. MacLane, *A Survey of Modern Algebra* (A K Peters, Massachusetts, 1997).
- [11] K. N. Stevens and S. J. Keyser, Quantal theory, enhancement and overlap, *J. Phonetics* **38**, 10 (2010).
- [12] I. Maddieson, *Patterns of Sounds* (Cambridge University Press, Cambridge, 1984).
- [13] J. Liljencrants and B. Lindblom, Numerical simulation of vowel quality systems: The role of perceptual contrast, *Language* **48**, 839 (1972).
- [14] B. de Boer, Evolution and self-organisation in vowel systems, *Evol. Commun.* **3**, 79 (1999).
- [15] J. Burridge, Spatial Evolution of Human Dialects, *Phys. Rev. X* **7**, 031008 (2017).
- [16] J. Milroy and L. Milroy, Linguistic change, social network and speaker innovation, *J. Linguist.* **21**, 339 (1985).
- [17] W. Labov, *Principles of Linguistic Change: Cognitive and Cultural Factors, Volume 3* (Wiley, Chichester, 2010).
- [18] N. Vallee, The weight of phonetic substance in the structure of sound inventories, *ZAS Papers Linguist.* **28**, 145 (2002).
- [19] B. Lindblom, Phonetic universals in vowel systems, in *Experimental Phonology*, edited by J. Ohala and J. J. Jaeger (Academic, Orlando, 1986), pp. 13–44.
- [20] R. Caree, Prediction of vowel systems using a deductive approach, in *Proceeding of Fourth International Conference on Spoken Language Processing, ICSLP '96* (IEEE, Piscataway, NJ, 1996), pp. 1593.
- [21] B. de Boer, Self-organization in vowel systems, *J. Phonetics* **28**, 441 (2000).
- [22] J. Pierrehumbert, Speech perception without speaker normalization: An exemplar model, in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullenix (Academic, New York, 1997), pp. 145–165.
- [23] P. F. Tupper, Exemplar dynamics and sound merger in language, *SIAM J. Appl. Math.* **75**, 1469 (2015).
- [24] S. Todda, J. B. Pierrehumbert, and J. Hay, Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model, *Cognition* **185**, 1 (2019).
- [25] K. J. B. Johnson, Speech perception without speaker normalization: An exemplar model, in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullenix (Academic, New York, 1997), pp. 145–165.
- [26] W. G. Moulton, Dialect geography and the concept of phonological space, *Word* **18**, 23 (1962).
- [27] J.-L. Schwartz, L.-J. Boe, N. Vallee, and C. Abry, The dispersion-focalization theory of vowel systems, *J. Phonetics* **25**, 255 (1997).
- [28] J. Pierrehumbert, Exemplar dynamics: Word frequency, lenition and contrast, in *Frequency and the Emergence of Linguistic Structure*, edited by J. L. Bybee and P. J. Hopper (Benjamins, Amsterdam, 2001), p. 137.
- [29] R. M. Nosofsky, Attention, similarity, and the identification–categorization relationship, *J. Exp. Psychol.: Gen.* **115**, 39 (1986).
- [30] B. Goodman and P. F. Tupper, Stability and fluctuations in a simple model of phonetic category change, *SIAM J. Appl. Dyn. Syst.* **17**, 2332 (2018).
- [31] B. Vaux and B. Samuels, Explaining vowel systems: dispersion theory vs natural selection, *Linguist. Rev.* **32**, 573 (2015).
- [32] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications* (Springer, Berlin, 2010).
- [33] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences* (Springer, Berlin, 2009).
- [34] L. D. Landau and E. M. Lifshitz, *A Course of Theoretical Physics: Statistical Physics: Volume 5* (Butterworth-Heinemann, Oxford, 1980).

- [35] W. Labov, *Principles of Linguistic Change: Internal Factors, Volume 1* (Wiley, Chichester, 1994).
- [36] R. J. Hunter, *Foundations of Colloid Science* (Oxford University Press, Oxford, 2001).
- [37] A. Stradner, H. Sedgwick, F. Cardinaux, W. C. K. Poon, S. U. Egelhaaf, and P. Schurtenberger, Equilibrium cluster formation in concentrated protein solutions and colloids, *Nature (London)* **432**, 492 (2004).
- [38] S. Mossa, F. Sciortino, P. Tartaglia, and E. Zaccarelli, Ground-state clusters for short-range attractive and long-range repulsive potentials, *Langmuir* **20**, 10756 (2004).
- [39] Y. Liu and Y. Xi, Colloidal systems with a short-range attraction and long-range repulsion: Phase diagrams, structures, and dynamics, *Curr. Opin. Colloid Interface Sci.* **39**, 123 (2019).
- [40] W. G. Mitchener, A mathematical model of prediction-driven instability: How social structure can drive language change, *J. Logic Lang. Inf.* **20**, 385 (2011).
- [41] K. Stadler, R. A. Blythe, K. Smith, and S. Kirby, Momentum in language change, *Lang. Dyn. Change* **6**, 171 (2016).
- [42] W. G. Mitchener, A stochastic model of language change through social structure and prediction driven instability, *SIAM J. Appl. Math.* **77**, 2272 (2017).
- [43] M. Sonderegger, Subglottal coupling and vowel space : An investigation in quantal theory, Ph.D. thesis, Massachusetts Institute of Technology, 2004.
- [44] K. L. Pike and E. V. Pike, Immediate constituents of Mazateco symbols, *Int. J. Am. Linguist.* **13**, 78 (1947).
- [45] Pierrehumbert, Probabilistic phonology: Discrimination and robustness, in *Probabilistic Linguist.*, edited by R. Bod, J. Hay, and S. Jannedy (MIT Press, Cambridge, MA, 2003), pp. 225–226.
- [46] Association phonétique internationale, Exposé des principes de l'association phontique internationale, *Le Maître Phonétique* **15**, 1 (1900).
- [47] S. Disner, Vowel quality: The relation between universal and language-specific factors, *UCLA Work. Pap. Phonetics* **58** (1983).
- [48] A. Wedel, A. Kaplan, and S. Jackson, High functional load inhibits phonological contrast loss: A corpus study, *Cognition* **128**, 179 (2013).
- [49] A. Martinet, Function, structure, and sound change, *Word* **8**, 1 (1952).
- [50] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, Acoustic characteristics of American English vowels, *J. Acoust. Soc. Am.* **97**, 3099 (1995).
- [51] P. Adank, R. van Hout, and R. Smits, An acoustic description of the vowels of northern and southern standard dutch, *J. Acoust. Soc. Am.* **116**, 1729 (2004).
- [52] J Crothers, Phonetic universals in vowel systems, in *Universals of human language. Vol. 2: Phonology*, edited by J. H. Greenberg, C. A. Ferguson, and Moravcsik (Stanford University Press, Stanford, 1978), pp. 93–152.
- [53] *PHOIBLE 2.0*, edited by Steven Moran and Daniel McCloy (Max Planck Institute for the Science of Human History, Jena, 2019).
- [54] J. K. Chambers and Peter Trudgill, *Dialectology* (Cambridge University Press, Cambridge, 1998).
- [55] G. Grimmett and D. Stirzaker, *Probability And Random Processes* (Oxford University Press, Oxford, 2001).
- [56] K. Fukunaga and L. Hoststler, The estimation of the gradient of a density function, with applications in pattern recognition, *IEEE Trans. Inf. Theory* **21**, 32 (1975).
- [57] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2009)
- [58] C. Liu and D. Kewley-Port, Vowel formant discrimination for high-fidelity speech, *J. Acoust. Soc. Am.* **116**, 1224 (2004).
- [59] J.C. Catford, *A Practical Introduction to Phonetics* (Oxford University Press, Oxford, 1988),
- [60] S. Y. Manuel, The role of contrast in limiting vowel-to-vowel coarticulation in different languages, *J. Acoust. Soc. Am.* **88**, 1286 (1990).
- [61] S. Lulich, Subglottal resonances and distinctive features, *J. Phonetics* **38**, 20 (2010).
- [62] J. Gillieron, *Genealogie des Mots Qui Designent l'Abeille d'après l'Atlas Linguistique de la France* (E. Champion, Paris, 1918).
- [63] X. Chi and M. Sonderegger, Subglottal coupling and its influence on vowel formants, *J. Acoust. Soc. Am.* **122**, 1735 (2007).
- [64] M. J. Sjerps and R. Smiljanic, Compensation for vocal tract characteristics across native and non-native languages, *J. Phonetics* **41**, 145 (2013).
- [65] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English: Phonetics, Phonology and Sound Change* (Mouton de Gruyter, Berlin, 2006).
- [66] N. Chomsky and M. Halle, *The Sound Pattern of English* (Harper and Row, New York, 1968).
- [67] R. Jakobson and M. Halle, *Fundamentals of Language* (Mouton, The Hague, 1971).
- [68] C. M. Millward, *A Biography of the English Language* (Holt Rinehart Wilson, Fort Worth, 1989).
- [69] J. B. Hay, J. B. Pierrehumber, A. J. Walker, and P. LaShell, Tracking word frequency effects through 130 years of sound change, *Cognition* **139**, 83 (2015).
- [70] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970).
- [71] J. Mathews and R. L. Walker, *Mathematical Methods of Physics* (Benjamin, New York, 1964).
- [72] W. Labov, *Principles of Linguistic Change: Social Factors, Volume 2* (Wiley, Chichester, 1994).
- [73] U. Weinreich, W. Labov, and M. T. Herzog, Empirical foundations for a theory of language change, in *Directions for Historical Linguist.: A Symposium*, edited by W. Lehmann and Y. Malkiel (University of Texas Press, Austin Texas, 1968), pp. 96–195.
- [74] T. M. Gureckis and R. L. Goldstone, How you named your child: Understanding the relationship between individual decision making and collective outcomes, *Top. Cognit. Sci.* **1**, 651 (2009).
- [75] L. Ellis and W. J. Hardcastle, Categorical and gradient properties of assimilation in alveolar to velar sequences: evidence from EPG and EMA data, *J. Phonetics* **30**, 373 (2002).