A computational toolkit for assessing phonological specification in phonetic data: Discrete Cosine Transform, Micro-Prosodic Sampling, Bayesian Classification

Jason A. Shaw & Shigeto Kawahara

January 12, 2017

Abstract
Many previous studies have argued that phonology may leave some phonetic dimensions unspecified. Lacking a phonologically specified phonetic target, the phonetic signal in these cases can only be structured by surrounding segments or, more specifically, by linear interpolation between segments flanking the targetless element. Natural variability in the phonetic signal presents a challenge for linear interpolation as a phonological diagnostic—the phonetic signal is never completely linear. Given this, how certain can we be on the basis of quasi-linear interpolation in the phonetics that a segment contains or does not contain a phonologically specified target? In this paper, we present a new methodology that allows us to answer this question. We setup stochastic generators of competing phonological hypotheses and use them to compute, on a token-by-token basis, the likelihood that a phonetic signal is the consequence of linear interpolation or, alternatively, that it is structured by a phonological target. The empirical material used to demonstrate the approach comes from Electromagnetic Articulography (EMA) recordings of high vowel devoicing in Japanese. We use Discrete Cosine Transform (DCT) to express tongue dorsum movement trajectories as a small number of frequency components (cosines differing in frequency and amplitude) that correspond to linguistically meaningful signal modulations, i.e., articulatory gestures. Our stochastic generators operate over this frequency space, generating the micro-prosody of tongue dorsum movements with realistic variation according to the presence or absence of a lingual articulatory gesture for the devoiced vowel. Finally, a Bayesian classifier trained on simulations of competing hypotheses assigns posterior probabilities to the data.

1.0 Introduction
1.1 Linear interpolation as evidence for phonological underspecification

Early generative phonology assumed that every segment is specified for every distinctive feature and receives “a phonetic command” for all the phonetic dimensions represented by distinctive features (e.g., Chomsky & Halle, 1968:403-419). However, this assumption has given way to various proposals regarding underspecification (Archangeli, 1988; Keating, 1988). Building on the phonological theory of feature underspecification, Keating (1988) observed that some segments lack a particular “phonetic target” in some dimension. One example that is used to argue for the theory of phonetic underspecification is English /h/, which on spectrograms can look like a linear interpolation from the preceding segment to the following segment; the illustrative figure from Keating (1988) is reproduced here as in Figure 1-(i).

The idea of phonetic underspecification—or in more neutral terms, apparent phonetic “targetlessness”—plays a central role in a wide range of phonological analyses. In the realm of intonation, Pierrehumbert (1980) argues that modeling intonational contours of English can be best achieved by only sparsely specifying H(igh) and L(ow) targets, rather than specifying all syllables in an utterance for tone. Pierrehumbert and Beckman (1988) demonstrate that apparent H-tone spreading in Japanese unaccented words is better characterized with phonetic underspecification, which shows linear decline from a H-tone to the next L-tone. Building on these observations, sparse tonal specification has been extended to the tonal and intonational analysis of many languages (e.g., Beckman & Pierrehumbert, 1986; Myers, 1998), and now constitutes a fundamental assumption in the Autosegmental Metrical (AM) theory of intonation (Arvaniti & Ladd, 2015; Jun, 2014; c.f., Xu, Lee, Prom-on, & Liu, 2015). Another oft-discussed example is the comparison of nasality in English and French, developed by Cohn (1993). She compared the phonetics of nasalization in French, where
nasalization is contrastive in vowels, and English, where nasalization is not contrastive in vowels. Figure 1-(ii) reproduces the English nasal airflow data. On the basis of these measurements, she argues that in English, vowel nasalization before a nasal consonant involves an interpolation from [-nasal] to [+nasal], with the vowel being underspecified for [nasal].

Figure 1-(i): Formant properties of intervocalic /h/ in English, take from Keating (1988: 283). (ii) Nasal airflow patterns in English, taken from Cohn (1993: 60). Both of these patterns are used to argue for phonetic underspecification.

Linear interpolation is indeed the crux of many phonetic arguments for phonological underspecification. Consider an ABC sequence, where the phonological specification of B is at issue. Whether observed in the domain of intonation (Pierrehumbert & Beckman, 1988: 37-38), vowels (Browman & Goldstein, 1992) or consonants (Cohn, 1993; Keating, 1988) linear interpolation on some phonetic dimension between A and C constitutes an argument for the “targetlessness” of some phonological feature of B. But there is one question that is not addressed sufficiently in the previous literature: how linear is linear?

Because phonetic signals are never completely linear, how much deviation from linearity would justify us to reject the targetlessness hypothesis? Consider again Figure 1-(ii). Does (b) really involve linear interpolation? The contours could also be described as higher dimensional (e.g., quadratic) functions. Similarly, (c) looks as if there may be a [+nasal] target in the middle of the vowel, or it involves a logarithmic function, because there is a plateau of nasality during the second half of the vowel. On the other hand, perhaps these data are just noisy actuations of a pure linear trajectory, as Cohn (1993) argues. Since humans are biological organisms, not speaking machines, phonetic data always come with some natural variability. Importantly, the specific patterning of phonetic variability can reveal the phonological form that structures the signal (e.g., Best, 2015; Shaw, Gafos, Hoole, & Zeroual, 2011). Explicitly modelling how different phonological forms structure natural variation in the phonetic signal provides a way to assess the likelihood that observed deviation from linearity is attributable to the presence of a phonologically-specified target or, alternatively, to the absence of such specification. Returning to the case of ABC, appropriately leveraging phonetic data to assess phonological specification of B requires distinguishing between complete targetlessness of B from phonetic reduction of B due to, for example, susceptibility to coarticulation with surrounding segments (c.f., Recasens & Espinosa, 2009) or predictability in context (Cohen-Priva, 2015; Seyfarth, 2014). Although rigorous assessment of linear interpolation is a challenging problem, it is one that can greatly

1 Not all work on phonetic underspecification assumes linearity. Pierrehumbert (1980: 71), cited and discussed in Myers (1998: 368), argues that for H* accents in English "the F0 falls until it is time to start aiming for the next H* level".
enhance our understanding of phonetic underspecification, which continues to play a crucial role in many contemporary phonological analyses\(^2\).

The main aim of this paper is to develop a methodology for rigorous assessment of linear interpolation in phonetic signals. There are three main components to the approach. The first, Discrete Cosine Transform (DCT), we use to decompose high dimensional phonetic data into a low-dimensional frequency space that can be mapped to phonological form. In this space, we formulate competing phonological hypotheses, including the linear interpolation hypothesis. Next, we deploy stochastic sampling techniques developed for syllable micro-prosody (e.g., Shaw and Gafos, 2015). We estimate distributions over signal components in frequency space and sample from these distributions to convert competing phonological hypothesis into the real world spatial-temporal dimensions of the data. This step factors into the analysis the range of natural variability found in the phonetic data, allowing us to generate realistically noisy phonetic signals from phonological hypotheses. Finally, we train a Bayesian classifier on the data simulated from competing phonological hypotheses and use it to compute, on a token-to-token basis, the probability of linear interpolation given the phonetic data. Taken together, the suite of computational tools allows rigorous assessment of the “targetlessness” claim on the basis of phonetic data. Most generally put, we maintain that our toolkit instantiates a happy marriage between computational methodologies with actual phonetic data, from the viewpoint of abstract phonological theory.

### 1.2 Japanese high vowel devoicing

To illustrate our computational approach, we examine high vowel devoicing in Tokyo Japanese (Fujimoto, 2015; Kondo, 2005; Tsuchida, 1997). A classic description of this phenomenon is that “high vowels are devoiced between two voiceless consonants and after a voiceless consonant before a pause”. High vowel devoicing is well-studied, particularly in Tokyo Japanese, with research covering phonological conditions (e.g., Kondo, 2005; Tsuchida, 1997), acoustic and perceptual characteristics of devoiced vowels (Beckman & Shoji, 1984; Faber & Vance, 2000; Nielsen, 2015), and articulatory studies of the vocal folds (Fujimoto, Murano, Niimi, & Kiritani, 2002; Hirose, 1971). See Fujimoto (2015) for a recent informative overview. When vowels are devoiced, it is difficult to ascertain whether they maintain their lingual articulatory specification. While we now have a good understanding of many aspects of high vowel devoicing in Tokyo Japanese, the status of the lingual gestures of devoiced vowels is still poorly understood. Here, we entertain four hypotheses, stated in (1).

1. **Hypotheses about the status of lingual articulation in devoiced vowels**

   - **H1: full lingual targets**—the lingual articulation of devoiced vowels is the same as for voiced counterparts.
   - **H2: reduced lingual targets**—the lingual articulation of devoiced vowels is phonetically reduced relative to voiced counterparts.
   - **H3: targetless**—devoiced vowels have no lingual articulatory target.
   - **H4: optionally targetless**—devoiced vowels are sometimes targetless

Not all previous studies are directly framed to support one of these hypotheses, but some studies are relevant to one of these hypotheses. Kawakami (1971: 24-26) argues that vowels delete in some phonological environments (=H3) and devoice in others. Sometimes, vowels leave no trace of themselves but coarticulation on surrounding consonants, which has led some researchers to conclude that the vowels are entirely deleted (Beckman, 1982; Beckman & Shoji, 1984; Whang, 2014). If deletion is phonological, then the vowel should not exhibit a lingual gesture, predicting H3. Kondo (2001) argues that high vowel devoicing is actually deletion based on a phonological consideration.

\(^2\) Some researchers have proposed that underspecification is not restricted to surface phonological or phonetic representations, but also holds at the level of the mental lexicon (e.g., Lahiri & Reetz, 2002; Lahiri & Reetz, 2010).
Devoicing in consecutive syllables is often prohibited, and Kondo (2001) argues that this prohibition stems from a constraint against complex onset or complex coda (i.e. *CCC). Even if vowel devoicing is phonological, some studies show that its application is optional or variable (Fujimoto, 2015; Nielsen, 2015), suggesting H4.

On the other hand, Kawahara (2015) argues that bimoraic foot-based truncation (Poser, 1990) counts a voiceless vowel as one mora (e.g. [suto] from [sutoraiki] ‘strike’, *[stora]). If [u] was completely deleted losing its mora, the bimoraic truncation should result in *[stora], but in actuality devoiced vowels always count toward the templatic, bimoraic requirement. This sort of proposal implies that the lingual gesture of devoiced high vowels should be phonologically present, and predict either H1 or H2. In particular, H1 is what would be predicted by a ”gestural overlap theory” of high vowel devoicing (Faber & Vance, 2000; Jun & Beckman, 1993; Jun, Beckman, & Lee, 1998). In this theory, high vowel devoicing occurs when laryngeal abduction gestures of surrounding consonants overlap with the vowel. In this sense, high vowel devoicing processes in Japanese (and Korean) are “not…phonological rules, but as the result of extreme overlap and hiding of the vowel's glottal gesture by the consonant's gesture” (Jun and Beckman 1993: 4). This passive devoicing hypothesis would predict that lingual gestures remain intact (=H1). Even if devoiced high vowels are not phonologically deleted or otherwise targetless, it would not be too surprising if the lingual gestures of high vowels are reduced. Due to devoicing, the acoustic consequences of a reduced lingual gesture would not be particularly audible. Hence from the standpoint of an effort-distinctiveness tradeoff, we expect reduction of oral gestures in high devoiced vowels (=H2).

The central aim of this paper is to develop a general methodology that allows us to distinguish between the hypotheses in (1). Especially, distinguishing between H2 and H3 is a specific case of the general issue raised in section 1.1. How do we know that a phonetic signal lacks a phonological target (=H3), rather than being reduced (=H2)? Although the empirical material used to demonstrate our approach comes from Japanese high vowel devoicing, the question that we are addressing is much more general: how do we assess linear interpolation in order to confirm the targetlessness of a particular segment? Some possible extensions of our proposed toolkit are discussed in section 6.4.

A key tenet of our approach is to express abstract phonological hypotheses in the units of the phonetic data. Like snowflakes and fingerprints, no two phonetic signals are identical, even those that actuate identical phonological structures. This fact dictates that rigorous assessment of phonological hypotheses on the basis of phonetic data requires a probabilistic model of how phonological form maps to the phonetic signal. Following recent approaches to syllable Micro-prosody (Gafos, Charlow, Shaw, & Hoole, 2014; Shaw & Gafos, 2015), we seek to estimate distributions that relate low dimensional phonological hypotheses to high dimensional phonetic data. Accordingly, we construct stochastic phonetic models that are parameterized by our phonological hypothesis as well as by the level of variability naturally present in the phonetic data. Before moving to the details of the computational toolkit, we briefly introduce the data to be modelled.

1.3 Electromagnetic Articulography recordings of Japanese vowels

The phonetic data used to illustrate our computational approach were drawn from a larger experiment (Authors 2016) using Electromagnetic Articulography (EMA) to track the movement of fleshpoints on the tongue during the production of voiced and voiceless vowels in Tokyo Japanese. The experimental methods are reported in section 4, and further details can be found in Authors (2016). Here, we preview the raw data in order to make concrete the nature of the problem. Figure 2 provides representative data from one speaker producing 11 repetitions of the minimal pair /ɸusoku/ ‘shortage’~/uşoku/ ‘enclosure’ in the carrier phrase /okee _____ to itte/ “Ok, say ____” (here and throughout, all the analyses of EMA data and computational analyses have been performed using Matlab; illustrative graphs are also generated using Matlab). The movement trajectories span from the /e/ in /okeg/ to the /k/ in /ɸusoku/ and in /uşoku/. In line with descriptions of high vowel devoicing in contemporary Tokyo Japanese, this speaker produced voiced /u/ in /uşoku/ and devoiced /u/ in /ɸusoku/ (see Authors 2016). Of interest for our case study is whether the lingual gesture of the devoiced vowel in /ɸusoku/ has an articulatory target. The top panel of Figure 2 shows the height of the tongue
dorsum (TD) sensor (y-axis) over time (x-axis) with /ɸusoku/ (devoiced /u/) in red and /ɸuzoku/ (voiced /u/) in blue. The middle panels show movement of the tongue blade (TB) and the bottom panel shows the tongue tip (TT). For the portion of the figure corresponding to /u/, the TD is lower for devoiced /u/ (/ɸusoku/, red lines) than for voiced /u/ (/ɸuzoku/, blue lines). At the very least, this pattern indicates that the devoiced vowel is phonetically reduced in this speaker’s pronunciation.

In the remainder of this paper, we describe a computational approach to evaluating the four hypotheses in (1) on the basis of continuous phonetic data, such as that shown in Figure 2.

Figure 2: Lingual articulatory trajectories from a female speaker of Tokyo Japanese producing /ɸusoku/ (red lines) and /ɸuzoku/ (blue lines).

2.0 Discrete Cosine Transform and Micro-prosodic Sampling
This section introduces the first two of three computational tools: Discrete Cosine Transform (DCT) and Micro-prosodic sampling. The purpose of these tools is to estimate distributions that characterize the phonetic data so that we can simulate realistically variable trajectories actuating phonological hypothesis (including the targetless hypothesis).

As a starting point, we assume that the speech articulators follow direct paths between articulatory goals (c.f., Browman & Goldstein, 1992; Keating, 1988). The idealized movement trajectory corresponding to the targetless vowel hypothesis (H3) would therefore be a linear trajectory from V1 to V3 (Browman & Goldstein, 1992; Lammert, Goldstein, Ramanarayanan, & Narayanan, 2014). In real articulatory data, flesh-point trajectories are never straight lines. There are well-studied cases in which tongue trajectories are curved because of biomechanical factors even when the idealized movement based on phonological form would dictate a linear trajectory (Mooshammer, Hoole, & Kühnert, 1995; Perrier, Payan, Zandipour, & Perkell, 2003). To account for the numerous perturbations, biomechanical and otherwise, of linear trajectories between articulatory goals in speech production, we
take a stochastic, data-driven approach, modelling actual trajectories as noisy actuations of phonological goals (Shaw & Davidson, 2011; Shaw & Gafos, 2010; Shaw et al., 2009).

Consider again the data in Figure 2. Since we are interested in the presence of a vowel, we focus on tongue dorsum (TD) movement, which is the primary articulator for (non-front) vowels (Browman & Goldstein, 1992; Johnson, Ladefoged, & Lindau, 1993). The trajectories begin with the vowel /e/ of the carrier phrase preceding the target words /ɸusoku/ (red lines) and /ɸuzoku/ (blue lines). The TD starts out high for the vowel /e/. The vowel /u/, if it is present, would follow the /e/. Some tokens show a slight rise in TD height at the start of the trajectory, which is expected if the tongue dorsum rises in height from /e/ to /u/; many tokens, however, particularly those of /ɸusoku/, show a monotonic decrease in height from /e/ to /o/, which is expected if there is no lingual target for /u/. The question we ask with the modelling is whether the observed trajectory from /e/ to /o/ is reliably different from a realistically variable linear trajectory between /e/ and /o/. If so, this would support the phonetic reduction hypothesis (H2); if not, the result would support the "targetless" hypothesis (H3), at least for some tokens (H4).

2.1 Discrete Cosine Transform
The TD trajectories in Figure 2 consist of 35 data points per trajectory, one data point for every 10 ms, due to the 100 Hz sampling rate used in the EMA recording (see section 4 for experimental methods). The data points are not statistically independent. Rather, the height of the TD at any point in time, τ, is closely related to the height of the TD at earlier, τ-1, or later time points, τ+1. We use Discrete Cosine Transform (DCT), the first computational tool in our toolkit, to capture dependencies between data points. Doing so allows data compression and sparse representation, which both simplifies subsequent computation and facilitates generalization to new data.

DCT represents the data as sums of cosines of different frequencies and amplitudes. In expressing spatial data in terms of harmonic components (i.e., frequency space), DCT is similar to Fast Fourier Transform (FFT) typically used to construct spectrograms from the acoustic signal. The main advantage of DCT, in particular for our purpose here, is that it represents the data with a small set of parameters, a general property of DCT (Jain, 1989: 151), each of which has a clear linguistic interpretation. Also important is that DCT has a known inverse function, which we use to simulate TD trajectories from DCT components. Each cosine component of a DCT has an amplitude coefficient that is fitted to the data. We interpret the amplitude of the cosines as the degree to which a corresponding gesture modulates the TD trajectory. DCT has been used in some previous phonetic studies, which have shown that phonetic signals, particularly vowel formants, can be represented quite well with a small number of cosine components (Watson & Harrington, 1999; Williams & Escudero, 2014).

We provide a mathematical expression of DCT transform in (2) and represent it visually with an example from the data in Figure 3. In the numerical expression, $y(k)$ is the amplitude of the $k^{th}$ cosine component. This is the output of the DCT. The other terms in the equation are as follows: $L$ is the length of the trajectory (i.e., the number of data points); $x(n)$ is the trajectory of the data being modelled; $w(k) = \frac{1}{\sqrt{L}}$ when $k = 1$ and $w(k) = \frac{\sqrt{2}}{\sqrt{L}}$ otherwise. The first DCT coefficient, $y(1)$, defines a straight line at a position above the average value of the data. This is because when $w(k) = 1$, the term of the cosine function is zero. This means that the first coefficient is equal to $\frac{\sum_{n=1}^{L} x(n) / \sqrt{L}}{\sqrt{L}}$, the sum of all data points in the trajectory divided by the square root of the number of data points (c.f., $\frac{\sum_{k=1}^{L} x(n)}{L}$, the average trajectory). Each subsequent DCT component defines a cosine of increasing frequency, as increases to $k$ linearly increase the term of the cosine function.
(2) Numerical expression of Discrete Cosine Transform

\[ y(k) = w(k) \sum_{n=1}^{L} \cos \left( \frac{\pi(2n - 1)(k - 1)}{2L} \right) \quad k = 1, 2, \ldots, L \]

where

\[ w(k) = \begin{cases} 
\frac{1}{\sqrt{L}} & \text{k = 1} \\
\sqrt{\frac{2}{L}} & 2 \leq k \leq L
\end{cases} \]

Figure 3 illustrates the DCT components of a TD trajectory. The top panel shows the trajectory, the vertical movement of the TD (y-axis) over time (x-axis). The first DCT coefficient defines a straight line at 14 mm (above the level of the occlusal plane, which is at 0 mm on the vertical axis). In the discussion below, we refer to this line as the baseline TD height. Subsequent coefficients describe deviations from the line as cosine-shaped modulations of increasing frequency. These subsequent components, i.e., the second to the fourth DCT components, are centered on zero. The second DCT coefficient captures the downward trend of the TD trajectory, ranging from +2 mm to -2 mm. Thus, the second DCT coefficient captures the fact that, in this data, the TD starts high and then lowers over time and that the range of this lowering motion covers a 4 mm span. The third coefficient adds another modulation to the trajectory. Towards the middle of the trajectory there is a rise. The third DCT coefficient indicates that this rise constitutes a modulation of the baseline trajectory on the order of ±2 mm. The effect of the fourth DCT coefficient is more subtle, specifying modulations that are less than ±0.5 mm.
Figure 3: An illustration of DCT components for a TD trajectory. The top panel shows the signal. The bottom four panels show individual DCT components contributing to the signal.

To evaluate how many cosine components are needed to represent movement trajectories in the EMA data, we re-simulated the TD data shown in Figure 2 using different numbers of DCT coefficients and evaluated the degree of precision representing the trajectory as a function of the number of DCT coefficients. The number of DCT coefficients was varied from one to ten. The right panel of Figure 4 shows how the correlation between the raw data and the simulated data increases with the number of DCT coefficients. The Pearson correlation is $r = .858$ with just two cosine coefficients. Increasing the number of parameters to four increases the correlation up to $r = .992$. The correlation with six DCT coefficients is $r = .998$. The left side of Figure 4 illustrates the goodness of fit token by token. The same set of eleven /fuzoku/ tokens from Figure 2 is re-displayed in green. The blue lines show trajectories that were simulated from DCT coefficients. With four coefficients, the blue and the green trajectories overlap almost completely, illustrating nearly lossless compression of the trajectories.
Figure 4: The left panels show the raw data (green lines) and simulated trajectories (blue lines) for 11 tokens of /ɸuzoku/. Simulations were based on two (top panel), four (middle panel) and six (bottom panel) DCT coefficients. The right panel shows the Pearson correlation ($r$) between raw data and simulated data as a function of the number of DCT coefficients employed in data representation.

By using DCT, we can effectively reduce the dimensionality of the data without loss of precision. The results in Figure 4 indicate that 4 DCT coefficients are sufficient to retain a detailed phonetic representation of the TD signal, for the Japanese case under discussion. More importantly, the 4 DCT coefficients have clear linguistic interpretations, on which we now elaborate.

Figure 5 re-displays the /ɸusoku/~/ɸuzoku/ data from Figure 2 along with mean DCT components fit to the data. The left column shows the voiced /u/ in /ɸuzoku/; the right column shows the devoiced /u/ in /ɸusoku/. The top panels show the raw data (green lines) together with the average trajectory (black line) and a linear trajectory between /e/ and /o/ vowel (red line). The average trajectory was computed by averaging DCT coefficients. It can be observed that the average trajectory (black line) is closer to the linear trajectory for /ɸusoku/ (right) than for /ɸuzoku/ (left). More to the point, each of the DCT coefficients has a clear linguistic interpretation, which helps to isolate the difference in trajectory between voiced and voiceless vowels. The first DCT coefficient represents baseline tongue height, as discussed above. The second DCT coefficient (3rd panel from the top) indicates a fall in TD height from /e/ to /o/. This component represents the vowel-to-vowel transition, which is similar for both words. Vowel to vowel intervals have long provided building blocks for speech production models and have a privileged theoretical status in some phonological models (Carré & Chennoukh, 1995; Mrayati, Carré, & Guérin, 1988; Ohman, 1966; Smith, 1995). The third DCT coefficient represents an increase in TD height for the vowel /u/. This rise is present for both /ɸusoku/ and /ɸuzoku/ but the magnitude of the rise is greater for the voiced vowel in /ɸuzoku/ than for the devoiced vowel in /ɸusoku/. Thus, the third DCT coefficient isolates the difference between these words observed in Figure 2. Finally, the fourth DCT coefficient adds a subtle (< .5 mm) modulation to the TD trajectory. The time course of this modulation is consistent with coarticulatory effects of coronal consonants /s/ and /z/ on TD height.

It is worth emphasizing here that the choice of using four DCT coefficients for this Japanese data was not determined a priori but arrived at through a combination of empirical and theoretical considerations. As we have illustrated above, (1) four DCT components provides a very precise representation of the data and (2) each component has a linguistic interpretation. In the general case, we advise that both of these criteria are deployed in determining the appropriate number of DCT components for a given data set.
Figure 5: Average DCT components for /ɸuzoku/ (left) and /ɸusoku/ (right). The raw data is displayed in the top panel. Average DCT components are plotted below.

To summarize, the first computational step in our approach is to express TD trajectories over /VCuCV/ sequences as the sum of four DCT components. The first of these components expresses a baseline TD position (akin to the intercept in regression analysis) while the second, third and fourth DCT components capture linguistically meaningful modulation of vocal tract constrictions. We next use these compressed representations of phonetic detail to estimate distributions characterizing the phonetic expression of phonological form, including the targetless hypothesis, H3 in (1).

2.2 Micro-prosodic sampling

We refer to the next computational tool as “Micro-prosodic sampling” because it borrows from recent stochastic approaches to modelling prosodic structure in terms of gestural timing (Shaw et al., 2011; Shaw & Gafos, 2010, 2015; Shaw et al., 2009). These studies estimate distributions over spatio-temporally defined gestural landmarks (Gafos, 2002) and sample from the distributions under different conditions. The parameters of such stochastic generators can be varied to test specific hypotheses about the phonological structure of the data, including the presence or absence of gestures (Shaw & Davidson, 2011) or the syllabic affiliation of the segments (Shaw & Gafos, 2010, 2015). We proceed here by defining Gaussian distributions over DCT coefficients instead of gestural landmarks.

Gestural landmarks and DCT coefficients both offer a sparse representation of detailed phonetic data. For the current case, an advantage of using DCT coefficients is that it is not necessary to parse specific gestural landmarks from the signal. Parsing gestural landmarks often relies on heuristic use of movement velocity profiles (Gafos, Hoole, Roon, & Zeroual, 2010; Shaw et al., 2009). In the trajectories shown in Figure 2, however, it is not possible to identify clear velocity peaks corresponding to the different vowel gestures. Rather, the TD moves smoothly with more or less constant velocity from one vowel to the next, a pattern also reported in other kinematic data sets (e.g., Browman &
Goldstein, 1992; Ohman, 1966). In data such as these, selecting a single point in time that corresponds to the vowel is arbitrary (Mücke, Grice, & Cho, 2014). Our solution here is to model the entire trajectory, but in the compact and linguistically relevant form of DCT coefficients.

To formulate the targetless hypothesis in terms of DCT coefficients, we first fit a straight line from $V_1$ to $V_3$ in $V_1CuCV_3$ sequences. If there is no independent TD height target for /u/, i.e., $H_3$ in (1), then the tongue dorsum position should follow a straight path from $V_1$ to $V_3$. To formulate a stochastic version of this targetless trajectory, we coerced the linear interpolation between vowels into frequency space by fitting four DCT coefficients to the straight line from $V_1$ to $V_3$. We then defined distributions over those DCT coefficients. The shape of the distributions was guided by analysis of the data. We chose Normal Gaussian distributions, since the DCT coefficients fit to our data did not significantly depart from Normality, according to Shapiro-Wilk tests. For the targetless hypothesis, the means of the distributions were the DCT coefficients fit to the linear interpolation between vowels. The standard deviation of the distributions was set to the standard deviation of DCT coefficients fit to the corresponding data. This ensures that we inject reasonable quantities of variation into the targetless trajectory. By expressing the interpolation line in the same stochastic four-parameter DCT space, we imbibe linear interpolation with variability along the same dimensions as variation in the data. Formalized as distributions over four DCT coefficients, corresponding to the bottom four panels of Figure 5, the targetless hypothesis thus has the same degrees of freedom as the full vowel hypothesis, meaning that it varies in the same dimensions and to the same degree as the raw data that we model.

The computational method described above expresses the targetless hypothesis in the phonetic dimensions of the data, specifically the height of the tongue dorsum over time, as phonological control of the vocal tract passes from one vowel to the next. Table 1 provides a specific example. The top two rows show the DCT distributions of the raw data. The mean value of each coefficient is shown with standard deviation given in parenthesis. The bottom two rows provide the parameters for the targetless hypothesis for the same data. The mean parameters come from a four parameter DCT of the straight line trajectory, left-delimited by the mean target of $V_1$ and right-delimited by the mean target of $V_3$. Note that the third coefficient, Co3, is zero, for the targetless hypothesis, indicating no rise from the trajectory defined by Co2 (see also Figure 5 for reference). The standard deviation of the targetless hypothesis is identical to the raw data because the level of variability in the targetless hypothesis is set to the level of variability in the data.

---

3 Four DCT coefficients are sufficient to capture a linear trajectory with a similar degree of precision ($r > .99$) as they capture dynamic movement trajectories. The illustrative figure, which can be read in the same way as Figure 4, is provided below as Figure-α.

---

Figure-α: Fitting a linear line with 4 DCT components.
Having defined distributions over DCT coefficients, we can sample from the DCT coefficients to simulate trajectories corresponding to the vowel deletion trajectory. The sampled DCT coefficients can then be used to specify the TD trajectory by applying the inverse DCT function to the coefficients. The formula for simulating trajectories by applying inverse Discrete Cosine Transform (iDCT) is given in (3). As with the DCT expression in (2), \( L \) indicates the length of the trajectory; \( x(n) \) is the trajectory, this time on the left side of the equation; \( y(k) \) represents the \( k \)th DCT coefficient; and \( w \) is a constant. We simulated trajectories that were equal to the mean duration of the \( V_1 \) to \( V_3 \) signal with \( k = 4 \) DCT Coefficients.

(3) Numerical expression of inverse Discrete Cosine Transform

\[
x(n) = \sum_{n=1}^{L} w(k) y(k) \cos \left( \frac{\pi(2n-1)(k-1)}{2L} \right) \quad n = 1, 2, \ldots L
\]

where

\[
\begin{align*}
\frac{1}{\sqrt{L}} & \quad k = 1 \\
\frac{\sqrt{2}}{L} & \quad 2 \leq k \leq L
\end{align*}
\]

Figure 6 illustrates the simulations graphically. For reference, the top panels re-plot the data from Figure 2 in green lines. However, in this figure only the portion of the trajectory beginning with the /e/ from the carrier phrase and ending with the /o/ is shown. The TD trajectories from /usoku/ are shown on the left; the trajectories from /usoku/ are on the right. The ‘x’’s denote vowel targets, the mean height of \( V_1 \), /e/, and \( V_3 \), /o/. The red line connects the means of the vowels and defines the linear interpolation trajectory. Comparison of the left and right panels reveals that TD height in /usoku/ tends to be closer to the line than does TD height in /usoku/, essentially the same observation we made in Figure 2 (but this time with reference to the linear interpolation trajectory). The bottom panels show simulated TD trajectories, sampled from the distributions of DCT coefficients given in Table 1. For reference, the red line denoting the linear interpolation trajectory is drawn in the lower panels as well. Note that, even though the mean of the DCT coefficients is based on the straight line, the stochastic simulations are non-linear. This is because the distributions over DCT coefficients define the same range of variability as is present in the TD trajectories, which are also not linear.

<table>
<thead>
<tr>
<th></th>
<th>Co1</th>
<th>Co2</th>
<th>Co3</th>
<th>Co4</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw data</td>
<td>/usoku/ (raw)</td>
<td>54.7 (3.27)</td>
<td>7.18 (1.80)</td>
<td>0.263 (0.886)</td>
</tr>
<tr>
<td></td>
<td>/usoku/ (raw)</td>
<td>64.4 (5.15)</td>
<td>6.25 (2.31)</td>
<td>-1.44 (1.95)</td>
</tr>
<tr>
<td>targetless hypothesis</td>
<td>/usoku/ (simulated)</td>
<td>53.3 (3.27)</td>
<td>7.29 (1.80)</td>
<td>0.00 (0.886)</td>
</tr>
<tr>
<td></td>
<td>/usoku/ (simulated)</td>
<td>60.2 (5.15)</td>
<td>7.96 (2.31)</td>
<td>0.00 (1.95)</td>
</tr>
</tbody>
</table>

Table 1: Mean and standard deviation (in parentheses) of DCT coefficients.
Figure 6: Top panels show actual TD data for /ɸuzoku/ (left) and /ɸusoku/ (right). The bottom panels show simulated trajectories from the targetless hypothesis (H3 in (1)).

Focusing on the bottom panels of Figure 6, observe the “accidental” vowels. That is, some of the trajectories belonging to the targetless hypothesis have an increase in height in the middle of the trajectory. If observed in isolation, these tokens could be misinterpreted as arising from active high vowel constrictions. The presence of “accidental” vowels underscores an important point about evaluating phonological hypothesis on the basis of phonetic data and the role of stochastic modelling. It is crucial to consider the level of variability in the data. In the case at hand, we find that amongst tokens actuating linear interpolation, there are some that show a rise in tongue dorsum height at approximately the point in time that we would expect the vocal tract to be under control of a vowel gesture; however, such “accidental vowels” simply come from the noise in the data, which should not be confused with phonologically-specified targets. The presence of accidental vowels indicates that the level of normal variation that characterizes fluent production of a native language is on the order of magnitude of the presence/absence of a vowel in one or two out of a dozen or so tokens.

At this point, it is now possible to statistically adjudicate between three of our four hypotheses in (1). In asking whether the vowel is targetless in the upper right panel of Figure 6, we are essentially asking whether the rise of the TD in the middle of the trajectory is greater than can be expected by chance. We have defined chance for /ɸusoku/ as our targetless trajectory in the bottom right corner of Figure 6. Using sparse representation of the data, as in Table 1, we can statistically compare /ɸusoku/ to /ɸuzoku/ to examine whether the TD trajectory in the devoiced vowel is significantly different from the TD trajectory in the voiced vowel. This constitutes a direct statistical test of H1, the hypothesis that
devoiced vowels are the same as the voiced counterparts. A significant difference would falsify H1, leaving us with H2 and H3, i.e., that the lingual gesture in devoiced vowels is either reduced (H2) or deleted (H3). Further, we can compare the TD trajectory in /husoku/ to the simulated targetless trajectory to test whether the data are significantly different from linear interpolation. This constitutes a statistical test of H3. A significant difference would leave us with H2, as the only viable alternative. However, this method does not allow us to test H4, the variable targetlessness hypothesis. The reason is that in evaluating statistical significance in this way, we are testing whether the tokens as a group are different, which involves the implicit assumption of phonological homogeneity across tokens of a word (see Bayles, Kaplan, & Kaplan, 2016 for a recent, relevant discussion). The next computational tool we introduce alleviates this assumption. We note, however, that if it can be ensured that a phonological process is not optional, then DCT together with Micro-prosodic sampling should suffice to provide a rigorous assessment of linear interpolation.

3.0 Bayesian Classification

A number of phonological processes are optional. Capturing the variability requires a probabilistic phonological model (Anttila, 1997; Boersma & Hayes, 2001; Coetzee & Kawahara, 2013). In these models, phonetic reduction and variable deletion (or even variable targetlessness, as in our case) are completely different scenarios. The latter requires stochastic interpretations of constraint rankings (or rules); the former requires continuous phonological representations of some sort (e.g., Smolensky, Goldrick, & Mathis, 2014). We therefore incorporate a third component to our computational toolkit, a naïve Bayesian classifier, which will allow us to analyze the data token-by-token without committing to the assumption that the surface phonological form of a word is uniform and invariant.

The Bayesian classifier assigns the probability of category membership. Importantly, it does so for each test token separately. For the case at hand, we use the DCT representation (four coefficients) as input to the classifier. The output is the probability of whether the articulatory target in that token comes from the “target present” category or the “targetless” category. The DCT coefficients that describe the data are statistically independent, which makes them appropriate dimensions for the naïve Bayesian classifier. The basic formula is provided in (4). The output of the formula, i.e. \( \prod_{i=1}^{n} p(T|C_{oi}) \), is the posterior probability of targetlessness, which designates the probability of a targetless articulation, given the four DCT coefficients. The alternative to a targetless articulation is that there is a full vowel target present. The posterior probability of a vowel target is calculated from a prior probability of targetlessness and the probability of the DCT values given the category. The prior probability of targetlessness is the term \( p(T) \). The probability of the DCT values given the category is the term \( \prod_{i=1}^{n} p(C_{oi}|T) \). This is calculated on the basis of the training data and is normalized by a third term, the probability of the DCT coefficients in the whole data set: \( \prod_{i=1}^{n} p(C_{oi}) \).\(^4\)

(4) Formula for Naïve Bayesian Classifier

\[
p(T|C_{o1}, ..., C_{on}) = \frac{p(T) \prod_{i=1}^{n} p(C_{oi}|T)}{\prod_{i=1}^{n} p(C_{oi})}
\]

where \( C_{oi} \) is \( i \)-th DCT coefficient (and \( n = 4 \) for the case at hand)

In this particular case, we are concerned with assessing the four hypotheses in (1) on the basis of the phonetic signal. To give each hypothesis equal weight, we assign equal prior probabilities to the

\(^4\)The denominator guarantees that the posterior is a well-formed probability, i.e., falls between 0 and 1, but since the denominator does not depend on \( T \), it does not influence separation between categories and, for this reason, is sometimes left out to simplify the equation.
categories target present, H1 in (1), and target absent, H3 in (1) hypotheses. Thus, \( p(T) \) is set to .5.\(^5\) The other hypotheses, vowel reduction (H2) and variable targetlessness (H4), can also be evaluated on the basis of posterior probability patterns, as we illustrate below (Figure 8).

We trained the Bayesian classifier on two sets of four DCT coefficients. The “target present” data came from DCT coefficients fit to tokens of /ɸuzoku/. The “target absent” data came from DCT coefficients fit to the linear interpolation trajectory from /e/ to /o/ (as in Figure 6, bottom). Since we have set the prior probability to even odds of targetlessness, it is the probability of each DCT coefficient given the presence/absence of a TD height target (the term \( \prod_{i=1}^{n} p(C_i|T) \)) that dictates posterior probabilities. To visualize this factor, Figure 7 compares Probability Density Functions (PDF)’s across hypotheses, target present vs. target absent, for each DCT coefficient. The black lines show PDFs over the “targetless” hypothesis; the red lines show the “target present” hypothesis. As can be seen from Figure 7, the PDF’s of the 1\(^{st}\), 2\(^{nd}\), and 4\(^{th}\) DCT coefficients are heavily overlapped. The main difference between the presence and absence of a vowel is found in the PDF of the 3\(^{rd}\) DCT coefficient. This is expected since the 3\(^{rd}\) coefficient dictates the magnitude of the TD rise between /e/ and /o/ vowels (see Figure 5). Thus, the parameters of the Bayesian classifier for /ɸusoku/ give quantitative probabilistic form to the observations we have already made about the data.

Figure 7: The probability distribution functions (PDFs) for DCT coefficients given the targetless hypothesis and the alternative target present hypothesis based on the data in Figure 9.

Four possible patterns, displayed as histograms over posterior deletion probabilities, are illustrated in Figure 8. These hypothetical results correspond to the four hypotheses laid out in (1). The

\(^5\) Just as the DCT analysis is flexible enough to allow us to use any number of DCT coefficients, \( p(T) \) is also flexible; it need not be set to 0.5 (equal probabilities of each hypothesis), if we have reason to set it otherwise (e.g. if we have a theory that prefers a presence of a vowel target in general, then \( p(T) \) can be set to be lower than 0.5). This is flexibility that is inherent in the Bayesian framework. See Gallistel (2009) for a general introduction to Bayesian analysis for cognitive scientists.
The histogram in the top left panel was obtained by submitting /fuZoku/ tokens (from 6 speakers) to the Bayesian classifier. As expected, most of these tokens have greater than .95 probability of containing a vowel, although there are a few tokens with lower probabilities. This pattern corresponds to H1, that the lingual gestures for voiced vowels are the same as for voiceless vowels. The histogram in the bottom left was obtained by submitting the same number of simulated “vowel absent” trajectories to the classifier. Again, as expected, most tokens have a .95 or greater probability of targetlessness, although there are a few tokens with lower probabilities and one “accidental vowel” which has a low-ish (.25) probability of targetlessness. This pattern corresponds to H3, that the lingual gestures of devoiced vowels are targetless. The third pattern, illustrated in the top right panel shows posterior probabilities for reduced vowels, H2. These were generated by stochastic sampling of DCT coefficients that were averaged between “target present” (H1) and “target absent” (H3) values. Thus, quite literally, the reduced vowel cases are intermediate trajectories between the fully articulated vowel and the targetless vowel. The fourth pattern, representing H4, is the variable deletion pattern. We created this pattern by sampling at random from distributions characterizing the target present data and the target absent data. In the following sections, we report methods for data collection and results of applying the tools described in this section to data collected from 6 speakers of Tokyo Japanese.

Figure 8: Four hypothetical posterior probability patterns. The vertical axis of each histogram shows posterior probabilities generated by the Bayesian classifier summarized in Figure 7. See the Appendix for different instantiations of H3 which assume different degrees of variability.

4.0 Electromagnetic Articulography Experiment
The experiment used Electromagnetic Articulography (EMA) (e.g., Hoole & Zierdt, 2010; Perkell et al., 1992) to track the articulatory kinematics of devoiced high vowels in Japanese. The full report can be found in Authors (2016).

4.1 Speakers
Six native speakers of Tokyo Japanese (3 male) participated. Participants were aged between 19 and 22 years at the time of the study. They were all born in Tokyo, lived there at the time of their participation in the study, and had spent no more than 3 months outside of the Tokyo region.
4.2 Materials
The stimuli in the experiment consisted of words presented in the carrier phrase: *ooke* __________ to itte* ‘Ok, say __________’. The preceding word /okee/ ending with [e] was selected so that the tongue would be in a non-high position at the start of the target word. A rise in tongue position from [e] to [u] would therefore suggest the presence of vowel target for [u]. To illustrate the computational approach, we focus on the four words in Table 2. In these words, the target vowel [u] occurs in either a devoicing environment (left column) or a voiced environment (right column). In both contexts, /u/ is unaccented. These words were randomized in a list of 16 other words, 10 of which did not contain high vowels in a devoicing context. All words were randomly displayed within the carrier phrase in normal Japanese script. Participants were instructed to speak as if they were making a request of a friend. Each participant produced a total of 10-15 repetitions of each target word.

<table>
<thead>
<tr>
<th>Devoiced vowels</th>
<th>Voiced vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>futaisei 主体性 ‘willingness’</td>
<td>judaika 主題歌 ‘theme song’</td>
</tr>
<tr>
<td>ɸusoku 不足 ‘shortage’</td>
<td>ɸuzoku 付属 ‘enclosed’</td>
</tr>
</tbody>
</table>

Table 2: (A subset of) stimulus items in the experiment used in Authors (2016). These two dyads are used in this paper to illustrate our computational method. See Authors (2016) for the analysis of other dyads, which contain non-initial devoiced high vowels.

4.3 Equipment
We used an NDI Wave EMA system sampling at 100 Hz to capture articulatory movement. NDI wave 5DoF sensors were attached to three locations on the sagittal midline of the tongue, and on the lips, jaw (below the lower incisor), nasion and left/right mastoids. The tongue dorsum (TD) sensor is the focus of our analysis here, as we are interested in the articulation of (non-front) vowels (Browman & Goldstein, 1992; Johnson et al., 1993; Wood, 1979). It was the most posterior of the three sensors on the tongue, attached as far back as was comfortable for the participant. Acoustic data were recorded simultaneously at 22 KHz with a Schoeps MK 41S supercardioid microphone.

4.4 Post-processing
Following the main recording session, we also recorded the bite plane of each participant by having them hold a rigid object, with three 5DoF sensors attached to it, between their teeth. Head movements were corrected computationally after data collection with reference to three sensors on the head, the left/right mastoid and nasion sensors, and the three sensors on the bite plane. The head corrected data was rotated so that the origin of the spatial coordinates corresponds to the occlusal plane at the front teeth.

5.0 Results
5.1 Analysis based on DCT and Micro-prosodic sampling
Figure 9 shows the TD height trajectory in /ɸusoku/ and /ɸuzoku/ from all six speakers. The data used to illustrate the modelling approach in section 2 comes from S01. Red lines show change in tongue dorsum height over time for /ɸusoku/; blue lines show /ɸuzoku/. The tongue height trajectories begin with the /e/ of the carrier phrase and continue for 350 ms. The dip in the trajectories corresponds to lowering of the TD for the vowel /o/. This is followed by a rise of the TD for /k/ at the ends of the trajectories. Only the portion of the trajectory spanning from /e/ to /o/ was included in subsequent analysis.
Figure 9: Change in tongue dorsum (TD) height (y-axis) over time (x-axis) for /ɸusoku/ (redlines) and /ɸuzoku/ (blue lines).

Following the method introduced in section 2, we fit DCT coefficients (section 2.1) to each TD height trajectory and defined a targetless trajectory (section 2.2). We then compared DCT components by MANOVA. For each speaker, we evaluated the effect of voicing on TD trajectory as well as differences between the actual trajectories and the targetless trajectory. Results are summarized in Table 3. Since the targetless trajectory is stochastically sampled, statistical comparisons vary depending on the particular sample evaluated. To ensure stable and replicable MANOVA results, below we report the average across 10 independent simulations of the targetless trajectory. Since we conducted these analyses for each speaker and each pair of items separately, we adjust the alpha level to correct for multiple comparisons. The Bonferroni corrected alpha is $\alpha = 0.00138 \times (0.05/36)$, where 36 is the total number of comparisons: 6 speakers $\times$ 2 item pairs $\times$ 3 comparisons per item pair.

In Table 3, significant differences are indicated by asterisk. Of the six participants, four produced reliable differences between the vowels in /ɸusoku/ and /ɸuzoku/. For all six participants, the voiced vowel in /ɸuzoku/ was significantly different from the targetless trajectory. Of the four participants who produced /ɸusoku/ and /ɸuzoku/ differently, only one speaker, S04, produced the devoiced vowel in /ɸusoku/ consistent with the targetless trajectory. For the other three, /ɸusoku/ was significantly different from the targetless trajectory. Thus, only one of the six speakers, S04, produced the devoiced vowel with a tongue dorsum height trajectory that could not be distinguished from the targetless trajectory.
Table 3: MANOVA results for /ɸusoku/ and /ɸuzoku/ for each speaker.

<table>
<thead>
<tr>
<th>Speakers</th>
<th>Comparison</th>
<th>df</th>
<th>F</th>
<th>p</th>
<th>lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>ɸusoku - ɸuzoku</td>
<td>21</td>
<td>22.9</td>
<td>0.0001*</td>
<td>0.2798</td>
</tr>
<tr>
<td></td>
<td>ɸuzoku -null</td>
<td>21</td>
<td>30.2</td>
<td>0.0000*</td>
<td>0.1912</td>
</tr>
<tr>
<td></td>
<td>ɸusoku -null</td>
<td>21</td>
<td>18.2</td>
<td>0.0012*</td>
<td>0.3641</td>
</tr>
<tr>
<td>S02</td>
<td>ɸusoku - ɸuzoku</td>
<td>25</td>
<td>20.8</td>
<td>0.0004*</td>
<td>0.3891</td>
</tr>
<tr>
<td></td>
<td>ɸuzoku -null</td>
<td>25</td>
<td>45.6</td>
<td>0.0000*</td>
<td>0.1289</td>
</tr>
<tr>
<td></td>
<td>ɸusoku -null</td>
<td>25</td>
<td>25.1</td>
<td>0.0002*</td>
<td>0.3270</td>
</tr>
<tr>
<td>S03</td>
<td>ɸusoku - ɸuzoku</td>
<td>23</td>
<td>26.8</td>
<td>0.0000*</td>
<td>0.2624</td>
</tr>
<tr>
<td></td>
<td>ɸuzoku -null</td>
<td>23</td>
<td>47.5</td>
<td>0.0000*</td>
<td>0.0948</td>
</tr>
<tr>
<td></td>
<td>ɸusoku -null</td>
<td>23</td>
<td>27.8</td>
<td>0.0002*</td>
<td>0.2602</td>
</tr>
<tr>
<td>S04</td>
<td>ɸusoku - ɸuzoku</td>
<td>23</td>
<td>23.3</td>
<td>0.0001*</td>
<td>0.3114</td>
</tr>
<tr>
<td></td>
<td>ɸuzoku -null</td>
<td>23</td>
<td>28.4</td>
<td>0.0000*</td>
<td>0.2459</td>
</tr>
<tr>
<td></td>
<td>ɸusoku -null</td>
<td>23</td>
<td>5.8</td>
<td>0.3138</td>
<td>0.7559</td>
</tr>
<tr>
<td>S05</td>
<td>ɸusoku - ɸuzoku</td>
<td>27</td>
<td>0.2</td>
<td>0.9953</td>
<td>0.9917</td>
</tr>
<tr>
<td></td>
<td>ɸuzoku -null</td>
<td>27</td>
<td>51.1</td>
<td>0.0000*</td>
<td>0.1204</td>
</tr>
<tr>
<td></td>
<td>ɸusoku -null</td>
<td>27</td>
<td>54.5</td>
<td>0.0000*</td>
<td>0.1047</td>
</tr>
<tr>
<td>S06</td>
<td>ɸusoku - ɸuzoku</td>
<td>29</td>
<td>0.2</td>
<td>0.9971</td>
<td>0.9940</td>
</tr>
<tr>
<td></td>
<td>ɸuzoku -null</td>
<td>29</td>
<td>32.9</td>
<td>0.0000*</td>
<td>0.2854</td>
</tr>
<tr>
<td></td>
<td>ɸusoku -null</td>
<td>29</td>
<td>25.2</td>
<td>0.0002*</td>
<td>0.3832</td>
</tr>
</tbody>
</table>

Figure 10 shows the trajectory of tongue dorsum height for another pair of words, /ʃutaisei/ and /ʃudaika/. The red lines show the word containing the devoiced vowel, /ʃutaisei/; the blue lines show the comparison word, /ʃudaika/, which contains a voiced /u/. For all six speakers, the TD trajectory is somewhat lower for the devoiced vowel (red lines) than for the voiced vowel (blue lines). Moreover, for several speakers the red lines take an almost linear trajectory between flanking vowels /e/ and /a/. To assess the statistical significance of these trends we again followed the methods outlined in section 2. We fitted DCT components to each trajectory, simulated a targetless trajectory and compared these via MANOVA. The results are summarized in Table 4.
As shown in Table 4, all six speakers produced /ʃutaisei/ and /ʃudaika/ with significantly different tongue height trajectories. Moreover, of the six speakers, only one, S05, produced /ʃutaisei/ so that it was significantly different from the targetless trajectory. For completeness, we also note that one speaker, S02, who did not produce a difference between /ʃutaisei/ and the targetless trajectory also did not produce a difference between /ʃudaika/ and the targetless trajectory that was significant after Bonferroni correction ($p = 0.009$ where $\alpha = 0.001$).
To summarize, most speakers showed significant differences in tongue height trajectories between voiced and voiceless vowels. This result allows us to rule out the possibility that devoiced vowels are produced with the same lingual articulatory gestures as voiced vowels, H1 in (1). With respect to targetlessness, the statistical evaluation indicates that /u/ may sometimes be targetless.6 One speaker, S04, produced /usoku/ and five speakers produced /utaisei/ without a clear height target. Thus, a conclusion based upon this analysis is that devoiced vowels are often reduced and sometimes even produced without a target. Moreover, we could divide speakers, based on this analysis into three groups. For /usoku/, there are speakers who produce /u/ without a height target (S04), those who reduce /u/ (S01, S02, S03) and those who produce full vowels (S05, S06). For /utaisei/, there are two groups: those who reduce (S05) and those who produce a targetless /u/ (S01, S02, S03, S04, S06). We caution, however, that analysis by MANOVA treats as a homogenous group all tokens of /usoku/ and /utaisei/ for a given speaker. If there is within-speaker optionality, then this assumption is not justified. We therefore conclude with an approach that allows us to evaluate targetlessness on a token by token basis. This approach will evaluate H4 in (1), optional targetlessness, and offer an additional angle on the other hypotheses, H1-H3.

5.2 Bayesian classification

We submitted each token of /usoku/ shown in Figure 9 and each token of /utaisei/ in Figure 10 to a Bayesian classifier as described in section 2.3. Recall that, as illustrated in Figure 8, all four hypotheses in (1) can be expressed as patterns of posterior probabilities, the output of the Bayesian classifier. For easy comparison to Figure 8, we have summarized the posterior probabilities as histograms.

Figure 11 provides a histogram of posterior probabilities for /usoku/. The left panel aggregates across speakers. The right panel provides a breakdown by speaker. The pattern is remarkably clear. There are two modes in the probabilities. One of them is around 0.05 probability of targetlessness; the other is around .95 probability of targetlessness. In fact, there are very few tokens at all that have intermediate probabilities, i.e., a token which we could call phonetically reduced. Across the six

---

6 As an anonymous reviewer pointed out, this reasoning runs the risk of concluding the lack of difference based on null results in statistical hypothesis testing. The Bayesian classification developed in the next section overcomes this problem. See Gallistel (2009) for relevant discussion on “proving the null” with Bayesian analyses.
speakers, rather, it seems like /u/ in /ɸusoku/ is optionally targetless. The breakdown of individual speakers in the right panel helps us to make sense of the MANOVA results. Recall that speakers S01, S02, S03, showed significant differences between /ɸusoku/ and /ɸuzoku/ as well as between /ɸusoku/ and the deletion trajectory. From the right panel of Figure 10, it is clear why. These speakers optionally produce the vowel without a height target. Speaker S04 does so most often. The main difference between speakers S01 through S04, therefore, is not a difference between phonetic reduction and phonological deletion but rather in the frequency with which the vowel is targetless. The other speakers, S05 and S06, produced no tokens classed as targetless. S05 was particularly remarkable in that all 11 tokens had a targetless probability of less than .05, and 10 of them had a probability less than .01.

Figure 12 provides histograms of posterior probability for /ʃutaisei/. The left panel shows the aggregate across speakers and the right panel shows the breakdown by speakers. Again, the pattern in the posterior probabilities is bimodal, with one peak (the larger peak) at a high probability of targetlessness and the other at a very low probability of targetlessness. Just like /ɸusoku/, in /ʃutaisei/ the vowel /u/ is produced with an optional vowel target. Noticeably absent are tokens that are intermediate between the full vowel and the linear interpolation trajectory. Amongst the 77 tokens from 6 speakers there are only a handful of tokens in the ambiguous, .3 to .7 probability range. Consistent with the MANOVA results, the individual speaker results indicate that five speakers tend to produce the /u/ in /ʃutaisei/ without a vowel height target while one speaker, S05, reliably produced the word with a vowel height target.

Figure 11: Posterior deletion probabilities for 77 tokens of /ɸusoku/ from 6 speakers (the TD trajectories shown in Figure 9). The left panel aggregates across speakers; the right panel shows probabilities by speaker.
The Bayesian classification provides converging evidence for some of the conclusions based on MANOVAs but also provides additional insight. Both analyses indicate that targetlessness is more common in /ʃutaisei/ than in /ɸusoku/. This is true for the aggregate data as well as for each speaker individually. Table 5 highlights this point, providing average targetless probabilities by speaker for each word. Although targetlessness varies across speakers, all of them show a higher probability of targetlessness in /ʃutaisei/ than in /ɸusoku/. This goes as well for S05, who has a low probability of targetlessness in both words. Although the MANOVA analysis also indicated that targetlessness was more common in /ʃutaisei/ than in /ɸusoku/, Bayesian classification reveals that this holds across all speakers individually as well. Thus, while the overall probability of targetlessness seems to be a matter of personal preference, relative probabilities of targetlessness are shared across speakers.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Targetless probability /ʃutaisei/</th>
<th>Targetless probability /ɸusoku/</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>0.9385</td>
<td>0.4590</td>
</tr>
<tr>
<td>S02</td>
<td>0.6664</td>
<td>0.4646</td>
</tr>
<tr>
<td>S03</td>
<td>0.8234</td>
<td>0.7499</td>
</tr>
<tr>
<td>S04</td>
<td>0.9338</td>
<td>0.8269</td>
</tr>
<tr>
<td>S05</td>
<td>0.0591</td>
<td>0.0064</td>
</tr>
<tr>
<td>S06</td>
<td>0.8476</td>
<td>0.0968</td>
</tr>
</tbody>
</table>

Table 5: average probability of targetlessness by speaker and by word

Another new insight gained from Bayesian classification is the status of phonetic reduction, i.e., H2 in (1). When comparing DCT components via MANOVA, we found that four out of six speakers showed significant differences between voiced (/ɸuzoku/) and voiceless (/ɸusoku/) contexts. Of these four speakers, three of them also showed a significant difference between /ɸusoku/ and the full vowel deletion trajectory. Since the productions of /ɸusoku/ are different, as a group, from /ɸuzoku/ and also, as a group, from linear interpolation, we might be tempted to conclude that the vowels are reduced (but not targetless). The Bayesian classification reveals that this conclusion is unwarranted. Rather, a group of /ɸusoku/ tokens from a single speaker may be different from both the full vowel trajectory in /ɸuzoku/ and the targetless trajectory because it contains a mix of full vowel and targetless tokens. The Bayesian classification revealed that this is indeed the case for /ɸusoku/. Production of a lingual target in devoiced vowels in Tokyo Japanese is optional but phonetic reduction is rare. Tokens are either
produced with a full vowel, similar to the voiced context, or with no lingual vowel target at all, as in the linear interpolation assumed for underspecified segments.

6.0 Discussion

6.1 Summary

We have illustrated three computational tools which, put together, allow rigorous assessment of the linear interpolation hypothesis. The general strategy is to express competing phonological hypotheses (including the linear interpolation hypothesis) in the dimensions of the data, including realistic levels of variability, and then to classify the phonetic data according to phonological structure. DCT compresses the phonetic data into a small number of phonologically relevant parameters that preserve phonetic detail. We defined linear interpolation in this DCT frequency space. Stochastic sampling from distributions over DCT coefficients enabled simulation of competing phonological hypotheses (target present vs. target absent) with the level of phonetic variability observed in the data, a step we refer to as Micro-prosodic sampling. Finally, we trained a Bayesian classifier to assign probabilities of targetlessness to new data. We have illustrated the method with tongue dorsum movements produced by speakers of Tokyo Japanese as a case study. Based on existing literature, we motivated four possible hypotheses about lingual articulatory targets and demonstrated step by step how our computational approach can adjudicate between them.

6.2 What we have learned about high vowel devoicing

Results for Tokyo Japanese indicate that the lingual articulatory gesture of devoiced vowels is rarely reduced, despite the fact that, given the devoicing, it can have only negligible auditory consequences. There are, however, two distinct phonetic outcomes for devoiced vowels. They can be produced with or without a vowel height target. This result supports H4 in (1), the hypothesis that devoiced vowels are optionally targetless.

Another interesting aspect of the results is that the probability of vowel targetlessness varied systematically across words. For all speakers, the probability of producing a vowel without a height target was higher for /ʃutaisei/ than for /ʃusoku/. Considering C₁ (/ʃ vs. /θ/) only, the result is somewhat surprising, but it could be due to resulting consonant cluster phonotactics. Deletion of /a/ in /ʃutaisei/ would give rise to a fricative-stop cluster, [ʃt], which may be a better surface form than the fricative-fricative cluster [ʃs] resulting from /a/ deletion in /ʃusoku/. If we assume that a syllable boundary remains between these surface consonants, a preference for fricative-stop clusters over fricative-fricative clusters is consistent with syllable contact laws (Murray & Vennemann, 1983; Vennemann, 1988). Since there is a greater decrease in sonority between the offset of one syllable and the onset of the next, [ʃt] is unmarked relative to [ʃ.ʃ] (Gouskova, 2004). It is not clear exactly what other facts of Japanese, if any, motivate this fine-grained grammatical preference, although similar types of patterns have been observed in the production and perception of unfamiliar consonant clusters (Berent, Lennertz, Smolensky, & Vaknin-Nusbaum, 2009; Berent, Steriade, Lennertz, & Vaknin, 2007; Davidson, Shaw, & Adams, 2007; Davidson & Shaw, 2012; Shaw & Davidson, 2011). See Authors (2016) for the discussion of other words, which conform to the above generalizations based on syllable contact laws.

Consistency across speakers in the relative targetlessness of /ʃutaisei/ and /ʃusoku/ resembles other well-studied cases of phonological variation, such as t/d deletion, in which grammatical influences remain constant even as overall deletion rates vary across speakers (Coetzee & Kawahara, 2013; Guy, 1997). Beyond what is reported in Authors (2016), which tested five dyads including the two that are reported here, we have run additional experiments to further evaluate the effect of surface consonant cluster as well as other linguistic factors on targetless probabilities, although the data have not yet been analyzed. Thus, although the final word on grammatical factors conditioning targetlessness must wait for additional data, it is clear that the computational approach we have developed can bring order to phonetic data, revealing phonological factors conditioning continuous movement dynamics with clarity and quantitative rigor.
6.3 Comparison with other approaches

Our approach differs from other quantitative attempts to assess phonological hypotheses, including targetlessness, on the basis of phonetic data. To highlight the uniqueness points, we briefly summarize past approaches, which can be divided into four categories: (i) heuristic use of phonetics (ii) statistical comparison of two samples of phonetic data (iii) predicting one part of the phonetic signal from another (iv) hypothesis testing by simulation.

The first approach, heuristic use of phonetics, involves drawing some conclusion about phonological form on the bases of visual inspection of the phonetic signal. Phonetic heuristics have played an important role in foundational work in laboratory phonology, including in the context of arguing for phonetic underspecification (Cohn, 1993; Keating, 1988). Phonetic heuristics can be extremely useful in augmenting transcriptions or auditory impressions of phonological form, particularly from field workers who are non-native speakers of the target language. However, phonetic heuristic may also break down. They are sometimes too sensitive and sometimes not sensitive enough. Consider, for example, a common phonetic heuristic for a vowel between stop consonants: “a period of voicing...with formant structure containing a visible second formant that ended with abrupt lowering of intensity at the onset of the second stop (Davidson, 2010).” Application of this heuristic to Tashlihiyt Berber, for example, greatly overestimates the frequency of vowels in the language (Ridouane, 2008). A Berber word like /tbdg/ “it is wet” contains no vowels in the phonological representation but is normally pronounced with three periods of voicing that would meet the above-stated phonetic heuristic (Fougeron & Ridouane, 2011). In this case, the phonetic heuristic is too sensitive. On the other hand, English words such as support, which contain two phonological vowels, are sometimes produced with just one phonetic segment meeting the above heuristic for a vowel (Davidson, 2006b). In this case, the phonetic heuristic is not sensitive enough. Heuristics breakdown because they do not capture the full range of phonetic signals consistent with phonological form.

An alternative to visual inspection of the phonetic signal is to statistically compare one or more phonetic dimensions in two groups of words hypothesized to differ in phonological structure. A wide range of statistical tools have been deployed to this end. For example, Smoothing Spline ANOVAs can be used to compare populations of splines (Gu, 2013), such as tongue shapes or the tongue height trajectories analyzed here, and have been applied to a range of phonetic signals (Davidson, 2006a). However, a significant difference between two populations of signals does not necessarily indicate the nature of the phonological difference. As our case study demonstrates, the same word can be produced with different phonological specifications. Populations of signals can therefore be different not because they actuate different phonological structures but because they actuate different mixtures of phonological structures (also Shaw & Davidson, 2011). Alternatively, populations of signals can differ due to phonetic factors. For example, Shaw et al. (2016) demonstrate that tongue height in Mandarin Chinese vowel production varies across tones. Despite the common claim that tones and vowels are phonologically independent (Yip, 2002), there are dependencies between laryngeal and supralaryngeal articulation that result in small but statistically significant sub-categorical differences in lingual articulation for the same vowel produced with different tones. Thus, statistical differences between surface measurements offer no guarantee of a categorical phonological difference between samples. As with heuristic use of phonetics, statistical comparison of continuous dimensions can be over-sensitive, picking up differences that do not correspond to phonological structure.

A third approach is to use one part of the phonetic signal to predict another. Specifically in the area of targetlessness, Pierrehumbert and Beckman (1988: 37-38) rely on this approach to argue for sparse tonal specifications in Japanese unaccented words. They argue that in unaccented words, only the first two syllables are specified as LH, and there is a general decline toward the L tone in the next Accentual Phrase. They show that the longer the duration between H and L, the shallower the slope. Their illustrative figure is reproduced here as Figure 13. In this case, the duration between H and L tones is used to predict the slope of the f0 fall. The relationship between these phonetic variables constitutes an argument for the tonal targetlessness of intervening syllables.
This specific correlation, since it requires manipulating the duration of the hypothesized “targetless” material, does not easily generalize to other cases of targetless, e.g., nasal specification of English vowels, etc, but conceptually similar approaches have been applied to other arguments for targetlessness. Browman and Goldstein (1992) used a multiple regression framework to assess whether English schwa contains an articulatory target. They reasoned that schwa could be claimed to be targetless in sequences such as /pV1papV2p/ if the spatial position of the articulators could be reliably predicted by the flanking vowels in a two parameter (one coefficient for each flanking vowel) linear regression model (also Lammert et al., 2014). They argued that schwa in such words has a target of its own, since regression models with an intercept term, representing the mean height of the signal, tended to outperform models informed only by flanking vowel positions.

In modelling contextual effects on English schwa, Browman and Goldstein (1992) also deploy what we would describe as a fourth type of approach. They simulated phonetic data from various phonological hypotheses, including targetlessness, and compare the simulated data to the experimental data. They find a qualitative match between simulated data and experimental data when English schwa is specified in the model with a neutral vowel target and overlapped in time with the following vowel, a result that converges nicely with the regression analysis described above, and makes different predictions than the targetless specification (particularly in high vowel environments). Browman and Goldstein (1992) explore several possible phonological configurations by specifying gestural scores by
hand and examining the phonetic consequences. More recent simulations derive gestural scores from coupled oscillators (Goldstein, Nam, Saltzman, & Chitoran, 2009; Nam, 2007; Saltzman, Nam, Krivokapic, & Goldstein, 2008) or posit coordination topologies isomorphic with syllable structure while fitting lower level parameters to the data (Gafos, Charlow, Shaw, & Hoole, 2014; Shaw & Gafos, 2010, 2015). The computational toolkit we have introduced belongs to this fourth type of approach, which assesses competing phonological hypotheses computationally, by simulating those hypotheses in the physical dimensions of phonetic data.

In comparison with other models instantiating the fourth class of approaches described above, our toolkit is, in some ways, more bottom-up, requiring fewer theoretical commitments and also fewer researcher degrees of freedom. First, the parameters in the model, i.e., the values of the DCT coefficients, are determined by the data, according to the algorithm in (2). Second, our approach does not privilege particular points in time, i.e., “magic moments” (Mücke et al., 2014), as having greater phonological relevance than others. In many of the studies described above, specific moments in time are selected for analysis. For example, Browman and Goldstein (1992) and Shaw et al. (2016) select, by automatic algorithm, a specific point in time to represent the spatial position of a vowel. Regardless of the algorithm, whether based on displacement of articulators, formant values, peak velocity, the temporal midpoint of voicing, etc., “target” selection introduces a researcher degree of freedom into the analysis. Our toolkit alleviates the necessity of picking points in time. This aspect of our approach is particularly useful for addressing the presence/absence of a target, as it is problematic to choose a point in time corresponding to a target that might not be there. Thus, our approach makes the presence/absence of targets a largely empirical question which can be addressed with phonetic data. The one assumption that we have adopted is that targetlessness corresponds to linear interpolation in the phonetics. Beyond this, since the parameters capturing phonetic signal modulation are fit to the data quantitatively, the bottom-up approach remains compatible with most higher level theories of phonological representation, including dynamically defined gestural units, as in Articulatory Phonology (e.g., Browman & Goldstein, 1990).

6.4 Broader application of the toolkit

Although the computational toolkit we have assembled to assess linear interpolation takes continuous phonetic data as input, the results for devoiced vowels in Japanese are remarkably categorical. Most tokens are either produced without a vowel target or with a full vowel target. The toolkit itself does not dictate such categorical outcomes (see Figure 8 and the Appendix). With respect to high vowel devoicing in Japanese, the categorical nature of the variation, as revealed by application of our approach, and its interaction with other grammatical factors suggests a distinctly phonological character to the phenomenon. We are optimistic about the prospects of applying our computational toolkit to a wider variety of phenomenon and curious about the extent to which other cases of “phonetic reduction” are actually manifestations of optional phonological processes. We strongly hope that the proposed toolkit will be used broadly in reassessing alleged cases of reduction to test whether they should be modelled reduction or as optional processes of phonological deletion/targetlessness.

As discussed in the introduction, our toolkit is designed to address the general issue of phonetic underspecification, including, e.g., formant transition patterns of English /h/ (Keating 1988), and nasalization of vowels before a tautosyllabic nasal consonant in English (Cohn 1993). Our toolkit can be used to re-examine claims that have been made about phonetic underspecification. One domain within which the current toolkit may be particularly applicable is intonation. As mentioned in the introduction, the issue of underspecification (targetlessness) is particularly important in the domain of intonation, because the dominant analytical framework of intonation, the Autosegmental/Metrical model of intonation, generally assumes sparse tonal specification (see, e.g., Xu et al. (2015) vs. Arvaniti and Ladd (2015) for a recent exchange of opinions on this matter). Since intonation comes with much

---

7 “Researcher degrees of freedom” refers to flexibility in methods of data collection, reporting and analysis that affect statistical assessment of a hypothesis (see, e.g., Simmons et al. 2011).
natural variability including individual variation, just like the articulatory data reported here, application of these tools to the tonal underspecification hypothesis may prove to be highly informative.\(^8\) For example, the tradeoff between signal length and f0 slope identified by Pierrehumbert and Beckman (1988) and shown in Figure 13 is a natural consequence of DCT, since the amplitude of DCT components are inversely related to the length of the signal (see (2) where \(y(k)\) is amplitude and \(L\) is signal length). Moreover, there are some “bumps” on the “linear” f0 trajectories, which could be due either to non-linguistic perturbations of the signal or phonological specifications, precisely the type of distinction that can be addressed in our framework.

In closing this section on the broader applicability of our approach, we would like to summarize aspects of our analysis that we expect will vary depending on the specific dataset being analyzed. We chose four DCT coefficients to model tongue dorsum trajectories over VCuCV sequences, but the number of DCT coefficients deployed in a given analysis will depend on the complexity of the data. For example, DCT fits to formant transitions in diphthongs have typically used just two DCT components (Elvin, Williams, & Escudero, 2016); longer sequences influenced by multiple overlapping gestures will likely require more. In the general case, we advise selecting DCT components based on two criteria: the precision with which they fit the data and the clarity of their linguistic interpretation. The maximum hypothesized number of phonologically dictated modulations in the signal under analysis may serve as an appropriate guideline. Second, we reported simulations of the targetless (linear interpolation) trajectory based on variance around DCT coefficients equivalent to the level of variability observed in voiced vowels. It is conceivable that a devoiced (or reduced) vowel could have greater variability than a full vowel, and we have explored this possibility as well (see Appendix). The broader point is that our approach if flexible. Micro-prosodic sampling as a methodological approach does not dictate the level of variability used in the simulations and it may at times be advisable to consider scaling these parameters. For example, injecting only the level of variability found in full vowels into our linear interpolation simulations still generated the occasional “accident” vowel from a targetless trajectory, but by gradually increasing variability, it would be possible to identify how variability influences the probability of accidental vowels. Finally, in the Bayesian classification stage of our analysis, we did not make use of the prior, but this option is available, and may be useful in cases in which there are independent reasons to suspect that one form or another has greater likelihood than other, as in, for example, non-native speech production (Davidson, 2010). For the case of Japanese high vowel devoicing, we did not have such evidence, so we simply posited that they are equally likely. In short, the computational tools that we have introduced here have the flexibility to be deployed in wide range of cases in which the phonetic specification of a target is at issue.

**7.0 Conclusion**

We have developed an approach to assessing the presence vs. absence of phonological specification in phonetic data bringing together a set of computational tools to assess the presence vs. absence of a phonological element on the basis of phonetic data. From end to end, the set of tools, consisting of Discrete Cosine Transform, Micro-prosodic sampling, and Bayesian Classification, does so without requiring explicit labelling of particular phonetic landmarks. In this sense, the toolkit can be productively deployed as a phonological feature detector. We demonstrated the approach with analysis of EMA recordings of voiced and devoiced vowels in Tokyo Japanese, contributing to a debate about whether devoiced vowels are specified for lingual articulatory targets. Analyzed within the

---

\(^8\) Relatedly, we note that the Fujisaki model of intonation (Fujisaki, 1983) is somewhat akin to our DCT analysis in that it has two components: the “phrasal control mechanism” determines the overall intonational contour (as our 2nd DCT coefficient) and “the accent control mechanism” realizes local f0 movement due to pitch accent (as our 3rd and 4th coefficients). There is independent observation that tones may differ in terms of how large the domain of their influence is; for example, a tone that is attached to a phrasal node may influence a sequence of syllables, whereas a tone that is associated with a particular syllable may have a more localized effect (Pierrehumbert & Beckman, 1988: 72-75). The former type of tones would be expressed in terms of lower DCT coefficients. In this sense, we already have a fairly good understanding of the linguistic interpretations of resulting DCT coefficients.
computational framework described here, these data elucidated some previously unknown aspects of the pattern, including its highly categorical nature and phonological conditions under which devoiced vowels also lack lingual articulatory targets.

Largely data-driven, adaptable to a range of phonetic signals, and compatible with a broad spectrum of representational frameworks in phonology, we anticipate that the computational toolkit can be widely deployed to link hypotheses about the specification (or non-specification) of phonological elements, including features, gestures, and tones, to phonetic data.

Appendix: a variant of H3, reduced vowels with high levels of variability

In this appendix, we considered a version of H3 in which the vowel target is not only reduced but has greater variability (than a regular vowel). We thank an anonymous reviewer for encouraging us to consider this possibility. What would be the predictions of our Bayesian classifier for a situation in which the devoiced vowel was both reduced and highly variable? We investigated this possibility through simulation, gradually increasing variability of the reduced vowel and observing how the change in variability influences the distribution of posterior probabilities from the classifier. The results are shown in Figure 14. The first panel of Figure 14 shows the result reported in Figure 8 (H3). There is a mono-modal distribution centered on a .5 probability of targlessness. Each subsequent panel shows a distribution over posterior probabilities resulting from increase variance in the reduced vowel. The second panel (from the left) shows results when the reduced vowel is twice as variable as a full vowel, the third panel triples the variance, the fourth quadruples the variance, and so on. We observe that as variance increases the posterior probabilities at the extreme ends of the scale, 0 and 1, also increase. This is because high variance makes extreme values more likely. At very high levels of variability, a tri-modal distribution emerges, with a peak in the distribution around 0, .5, and 1.

<table>
<thead>
<tr>
<th>Level of variability of the reduced vowel is</th>
<th>reduced vowel is two times more variable than full vowel</th>
<th>reduced vowel is three times more variable than full vowel</th>
<th>reduced vowel is four times more variable than full vowel</th>
<th>reduced vowel is five times more variable than full vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD(<em>{\text{reduced}}) = SD(</em>{\text{full}})</td>
<td>SD(<em>{\text{reduced}}) = 2*SD(</em>{\text{full}})</td>
<td>SD(<em>{\text{reduced}}) = 3*SD(</em>{\text{full}})</td>
<td>SD(<em>{\text{reduced}}) = 4*SD(</em>{\text{full}})</td>
<td>SD(<em>{\text{reduced}}) = 5*SD(</em>{\text{full}})</td>
</tr>
</tbody>
</table>

Figure 14: Increasing the inherent variability of reduced vowels for H3.

Acknowledgements (your name here)
References


