Assessing feature specification in surface phonological representations through simulation and classification of phonetic data

Abstract

Many previous studies have argued that phonology may leave some phonetic dimensions unspecified in surface representations either because of deletion or lexical underspecification. Lacking a phonologically specified phonetic target, the phonetic signal in these cases can only be structured by phonetic interpolation between segments flanking the targetless element. However, natural variability in the phonetic signal presents a challenge for identifying instances of phonetic interpolation. Our approach to this challenge is to explicitly model phonetic variability providing a computational link between phonological hypotheses and phonetic data. To this end, we set up 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 phonetic interpolation, defined as a smooth interpolation between flanking segments, or, alternatively, that it is structured by a phonological target. The empirical material used to demonstrate the approach comes from Electromagnetic Articulography recordings of high vowel devoicing in Japanese. We use Discrete Cosine Transform 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 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 the targetless trajectory assigns posterior probabilities to the data. Results indicate that /u/ is optionally produced without a vowel height target in Tokyo Japanese and that the frequency of targetlessness varies across phonological environments.

1.0 Introduction

1.1 Phonetic 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 is English /h/, which on spectrograms can look like an interpolation from the preceding segment to the following segment. Another example is nasal airflow data in English (Cohn 1993), in which vowel nasalization before a tautosyllabic nasal consonant involves phonetic interpolation from [-nasal] to [+nasal], with the vowel itself being underspecified for [nasal]. Other research has argued that various types of vocalic transitions between consonants are not phonologically specified but, rather, are best described as periods of open vocal tract with no vocalic target. Cases such as this include the transitional vocoids surfacing onset consonant clusters in Yine/Piro (Hanson, 2010: 28), the vocoid that surfaces between final consonant clusters in Moroccan Arabic (Gafos, 2002), and the production of phonotactically illicit consonant clusters by non-native speakers (Davidson, 2010). See Hall (2006: 390) for a list of 29 languages with phonologically inert "excrescent vowels" and a discussion of their common properties.

Intonation is another area in which the idea of phonetic underspecification has played a central role in theory development. 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 for tone. Pierrehumbert and Beckman (1988) demonstrate that apparent H-tone spreading in Japanese unaccented words, proposed by Haraguchi (1977), is better characterized with phonetic underspecification. The phonetic data shows a roughly linear decline from a H-tone to the next L-tone (see also discussion in section 6.4). Building on these observations, sparse tonal specification has been extended to the intonational analysis of many languages (e.g., Beckman & Pierrehumbert, 1986; Myers,

1998) , and now constitutes a fundamental assumption in the Autosegmental Metrical theory of intonation (Arvaniti & Ladd, 2015; Jun, 2014; c.f., Xu, Lee, Prom-on, & Liu, 2015).

Generalizing across these cases, there is a large body of literature arguing that phonetic behavior is determined by sparse (surface) phonological specification. Determining which phonetic dimensions are under phonological control on the basis of the phonetic signal alone is challenging, as it involves discovering phonological control in the presence of influences on the phonetics from many other factors, speech rate, individual differences, etc. At times, the indeterminacy of phonetic data has given rise to highly disparate characterizations of the same language by different researchers. For example, Tashlhyit Berber has been described variably as a language with many epenthetic vowels (Coleman, 2001) and, alternatively, as having syllabic consonants and no epenthetic vowels (Dell & Elmedlaoui, 1985; Ridouane, 2008). This dichotomy hinges on whether transitions between consonants are treated as being under the phonological control of vowels or not and has been largely resolved through converging evidence from multiple data sources, including appropriate phonetic analyses (Ridouane, 2008). Similar ambiguity is present in other languages. Hall (2006) argues that vocoids between stops and sonorants in Hocank (Winnebago), which are invisible for the purpose of primary stress placement, are not true vowels but merely open transitions between consonants. Other researchers have argued on theoretical grounds that the Hocank vocoids are epenthetic vowels (Davis & Baertsch, 2010), which makes stress placement rules opaque and has consequently spawned a range of theoretical proposals to account for stress-epenthesis interactions, including iterative application of metrical feet (Hale & Eagle, 1980), positional faithfulness (Alderete, 1995), and ordered application of the same epenthesis process in different environments (Strycharczuk, 2009). In the absence of a robust phonetic record,[1] other researchers have reinterpreted the facts, arguing that stress occurs on the epenthetic vowel (Stanton & Zukoff, to appear). Vowels that are invisible to stress in other languages have been shown to differ variably in phonetic quality and duration from vowels that influence stress placement, raising questions about the degree of surface opacity (Hall, 2013). The broader point is that theoretical debates can emerge from ambiguity about surface phonological form, particularly when appropriate analysis of phonetic data are unavailable. This paper develops analytical tools to strengthen interpretation of surface phonological (non-)specification on the basis of phonetic data.

In many of the phonological domains described above, phonetic interpolation has been a key argument for the phonological non-specification of some dimension, whether it be tone, a phonological feature, or a segment. The general logic is as follows. Consider an ABC sequence, where the phonological specification of B is at issue and B is assumed to dictate the phonetic parameter, *p*. Whether observed in the domain of intonation (Pierrehumbert & Beckman, 1988: 37-38), vowels (Browman & Goldstein, 1992) or consonants (Cohn, 1993; Keating, 1988) phonetic interpolation on dimension, *p*, between A and C has been motivated as an argument for the "targetlessness" of B.

Rigorously assessing phonetic interpolation is not always straightforward, owing in part to the natural variability associated with phonetic data. Moreover, listeners show remarkable tolerance for phonetic variation (Shaw et al., to appear). Importantly, the specific patterning of phonetic variability can reveal the phonological form that structures the signal (e.g., 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 phonetic data can be attributed 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 based on some phonetic parameter, *p*, requires distinguishing complete targetlessness from phonetic reduction due to, for example, susceptibility to coarticulation with surrounding segments (c.f., Recasens & Espinosa, 2009) or high predictability in context (e.g., Cohen-Priva, 2017; Shaw & Kawahara, 2017). Although rigorous assessment of phonetic interpolation is a challenging problem, it is one that can greatly enhance our confidence in the identity of surface phonological representations.

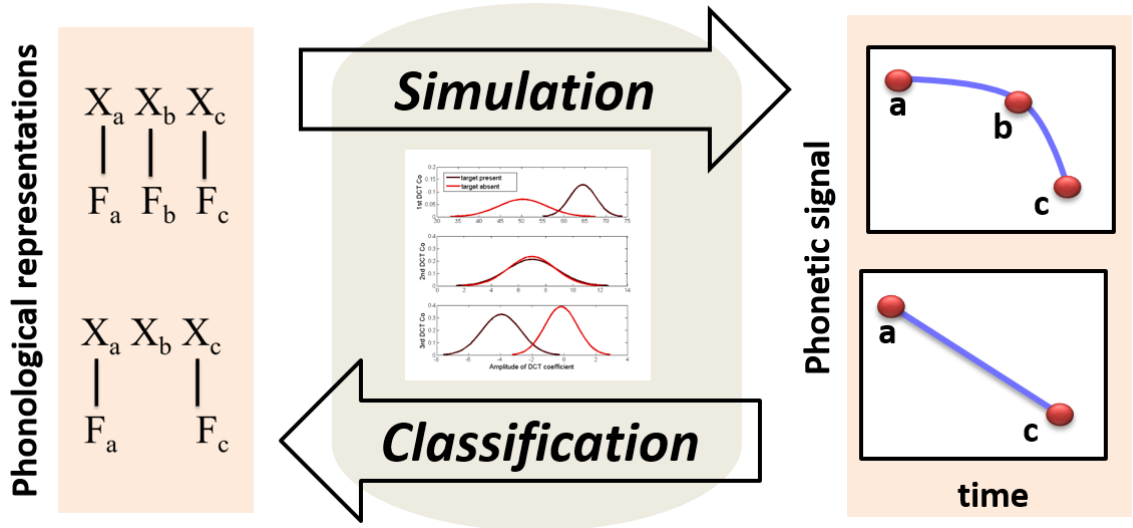
[1] An instrumental phonetic analysis of Miner's original recordings, which serves as the empirical basis of the primary descriptions is now underway (Hall & Sue, 2018)

This paper develops a general methodology for assessing feature specification in surface phonological representations on the basis of the phonetic signal. 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.

A schematic of the approach is presented in Figure 1. The key idea is to link surface phonological form (Figure 1: left) to time-varying phonetic data (Figure 1: right) through the stochastic processes of simulation and classification (Figure 1: middle). Our proposal for the stochastic representational space that supports these processes makes use of parametric (Gaussian) distributions over frequency components of the phonetic signal. We use Discrete Cosine Transform (DCT) to decompose high dimensional phonetic data into a low-dimensional frequency space that can be mapped to phonological form. In this frequency space, we formulate competing phonological hypotheses, including the phonetic interpolation ("targetless") hypothesis (Figure 1: bottom). For the purposes of this paper, we follow others (e.g., Lammert, Goldstein, Ramanarayanan, & Narayanan, 2014) in making the simplifying assumption that interpolation will take the form of a linear transition between flanking segments, though we return to this assumption in the discussion.[2] We estimate distributions over signal components in frequency space and sample from these distributions to convert competing phonological hypotheses into the real world spatial-temporal dimensions of the data. This step, simulation, factors into the analysis the range of natural variability found in the phonetic data, allowing us to generate realistically variable phonetic signals from discrete phonological hypotheses. Finally, we train a Bayesian classifier on the data simulated from competing phonological hypotheses (feature present vs absent) and use it to compute, on a token-by-token basis, the probability of interpolation (target absent) given the phonetic signal. Taken together, this computational toolkit renders stochastic representations that support rigorous assessment of "targetlessness" through simulation and classification of phonetic data.

[2] Although we follow others in making the pragmatic choice to use "linear interpolation" as an estimate of the actuation of targetless elements, the analysis presented below is capable of expressing other shapes of interpolation.

Figure 1: Schematic depiction of the modelling approach. The left box shows phonological representations with (top) and without (bottom) a particular feature, F_b, while the right box shows corresponding phonetic singles with (top) and without (bottom) a phonetic target for F_b. The link between phonological form and the phonetic signal is a stochastic representational space—Gaussian distributions over (DCT) frequency modulation components—used for simulation and classification.



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 key debate regarding this phenomenon is whether the surface phonological representation contains a vowel or not. A classic description of the facts is that "high vowels are devoiced between two voiceless consonants and after a voiceless consonant before a pause". As we review below, one proposal is that the vowel is not only devoiced but entirely absent from the surface representation, due to deletion. When vowels are devoiced, it is difficult to ascertain from the acoustics whether they are also deleted. For this reason, we look to the articulatory signal to adjudicate between competing proposals, taking the presence/absence of a lingual articulatory target for the vowel as an indicator of surface phonological specification. We consider 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

Several previous studies relate to one or more of these hypotheses. Kawakami (1971: 24-26) argues that vowels delete in some phonological environments (=H3) and devoice in others. Sometimes, the only traces of vowels found in the (acoustic) phonetic signal are vowel-conditioned allophony on surrounding consonants, which has led some researchers to conclude that the vowels are entirely deleted (Beckman, 1982; Beckman & Shoji, 1984). If deletion is phonological, then the vowel should not exhibit a lingual gesture, predicting H3. Similarly, Kondo (2001) argues that high vowel devoicing is actually deletion based on a phonological consideration. Devoicing in consecutive syllables is often

prohibited, and Kondo (2001) argues that this prohibition stems from a constraint against complex onsets or codas. Even if vowel devoicing is due to phonological deletion, some studies show that its application is optional or variable (Fujimoto, 2015; Nielsen, 2015), suggesting H4.

On the other hand, Tsuchida (1997) and Kawahara (2015) argue that bimoraic foot-based truncation (Poser, 1990) counts a voiceless vowel as one mora (e.g. [su̥to] from [su̥toraiki] 'strike', *[stora]).[3] If [u] was completely deleted losing its mora, the bimoraic truncation should result in *[stora], but in actuality devoiced vowels always count toward the bimoraic requirement. This sort of proposal implies that the lingual gesture of devoiced high vowels should be phonologically present, and predicts either H1 or H2. In particular, H1 is predicted by a "gestural overlap theory" of high vowel devoicing (Beckman, 1996; Jun & Beckman, 1993; Jun, Beckman, & Lee, 1998). In this theory, high vowel devoicing occurs when laryngeal abduction gestures of surrounding consonants heavily 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).

We use the general methodology described in section 1.1 to distinguish between the hypotheses in (1) on the basis of phonetic data. 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 more general: how do we assess phonetic interpolation in order to confirm or reject phonological specification? Some potential broader applications of our proposed toolkit are discussed in section 6.4.

The remainder of the paper is organized as follows. Section 2.0 describes the experimental methods involved in collecting articulatory data. Section 3.0 and Section 4.0 motivate the computational approach, specifically the tools used for simulation (Section 3.0) and classification (Section 4.0). Section 5.0 provides an analysis of the data addressing the hypotheses in (1). Section 6.0 provides some discussion of the results as well as alternative approaches to assessing surface phonological specification on the basis of phonetic data.

2.0 Electromagnetic Articulography Experiment

The phonetic data used to illustrate our computational approach were drawn from a larger experiment 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 full report of the experiment can be found in Shaw and Kawahara (2018b). This paper focuses on illustrating the computational tools.

2.1 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 and had spent no more than 3 months outside of the Tokyo region.

2.2 Materials

The stimuli in the experiment consisted of words presented in the carrier phrase: *ookee* _____ *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 suggest the presence of a vowel target for [u]. To illustrate the computational approach, we focus on

[3] The voiceless vowel is underlined.

the two dyads (=four words) in Table 1. 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.

Devoiced vowels	Voiced vowel
ʃu̥taisei 'willingness'	ʃudaika 'theme song'
ɸu̥soku 'shortage'	ɸuzoku 'enclosed'

Table 1: (A subset of) stimulus items in the experiment used in Shaw and Kawahara (2018). These two dyads are used in this paper to illustrate our computational method.

2.3 Equipment

We used an NDI Wave EMA system sampling at 100 Hz to capture articulatory movement in 3D. The spatial accuracy of this system in generally within 0.5 mm. 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 height of the tongue dorsum (TD) sensor is the focus of our analysis. The TD sensor was the most posterior of the three sensors on the tongue, attached as far back as was comfortable for the participant (~5-6 cm behind the tip). Acoustic data were recorded simultaneously at 22 KHz with a Schoeps MK 41S supercardioid microphone.

2.4 Post-processing

We 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.

2.5 Trajectories for analysis

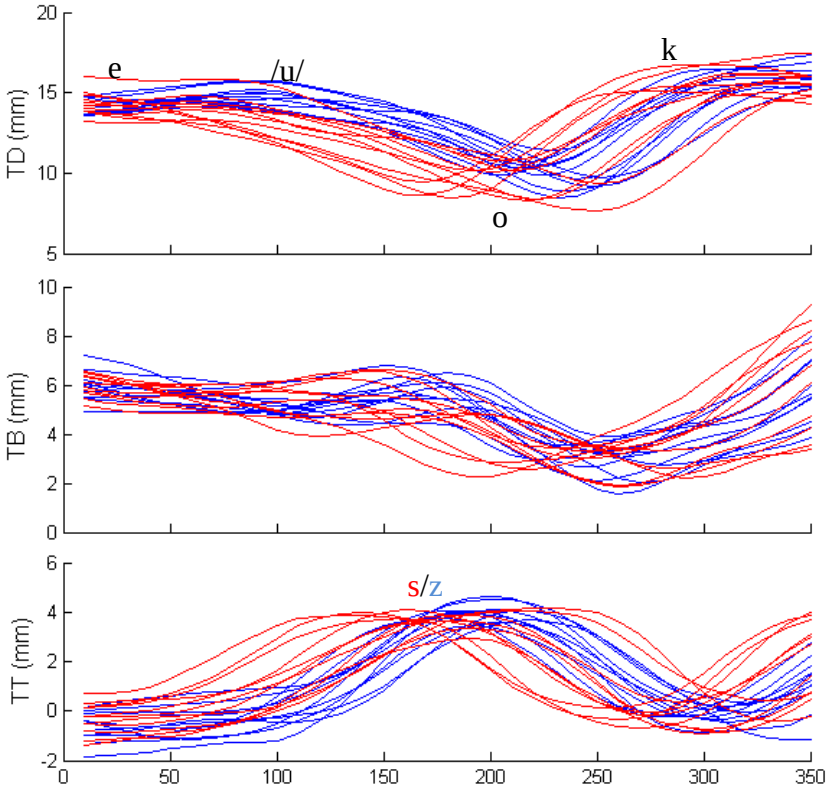
We first visualized the data using the Matlab-based software Mview (Tiede, 2005), which displays the EMA movement trajectories along with the waveform and spectrogram from the audio signal. In Mview, we verified that /u/ was devoiced in devoicing environments by visual inspection of the spectrogram. At this stage of analysis, we also identified articulatory landmarks associated with V_1 and V_3, the vowels preceding (V_1) and following (V_3) the target /u/. The point of minimal velocity in the TD signal corresponding to the location of V_1 and V_3 in the spectrogram, a reliable method of parsing vowel targets in articulatory data (for discussion, see, e.g., Blackwood Ximenes, Shaw, & Carignan, 2017), was used to left-delimit (V_1) and right-delimit (V_3) the interval, V_1CuCV_3, containing /u/. This interval was the subject of all subsequent analyses. To illustrate the raw data, at times we also provide plots of longer intervals so that movements of following consonants can also be visualized.

2.6 Preview of raw data

A full analysis of the data is provided in section 5.0. Here, we preview the raw data in order to make concrete the general difficulty involved in assessing phonological specification from continuous phonetic data. Figure 2 provides representative data from one speaker producing 11 repetitions of the minimal pair /ɸusoku/~/ɸuzoku/. The movement trajectories span from the /e/ in /oke̥e̥/ to the /k/ in /ɸuso̥ku/ and in /ɸuzo̥ku/. In line with descriptions of high vowel devoicing in contemporary Tokyo Japanese, this speaker produced voiced /u/ in /ɸuzoku/ and always devoiced /u/ in /ɸusoku/(Shaw & Kawahara, 2018b). 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 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 subset of the data. 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). The *y*-axis shows the height of the sensors; the x-axis shows time (in ms). The trajectories span a 350 ms window starting from the /e/ of the carrier phrase and extending to the TD rise for the /k/ near the end of the words. Subsequent analysis focuses just on the interval from /e/ to /o/.



3.0 Simulation

This section introduces the computational tools that support simulation of competing phonological hypotheses in the dimensions of the data. We estimate distributions over a small number of parameters 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 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 V_1 to V_3 (Browman & Goldstein, 1992; Choi, 1995; Lammert et al., 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 Figure 2. Since we are interested in the presence of a vowel, we focus on TD movement, which is the primary articulator for (non-front) vowels (see, e.g., Wood, 1979). 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 TD 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 modelling addresses whether the observed trajectory from /e/ to /o/ is 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).

3.1 Discrete Cosine Transform

Due to the 100 Hz sampling rate used in the EMA recording, there is one data point for every 10 ms, e.g., the 350 ms TD trajectories in Figure 2 consist of 35 data points per trajectory. The data points in a trajectory 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. At a deeper level, the statistical dependencies between data points across the entire trajectory are due (at least in part) to phonologically controlled movement. 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, is that it represents the data with a small set of parameters, a general property of DCT (Jain, 1989: 151). In addition, as we will see below, each of the DCT coefficients may have 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 changes in vowel formants over time, can be represented quite well with a small number of cosine components (Elvin, Williams, & Escudero, 2016; Watson & Harrington, 1999).

A mathematical expression of DCT transform is provided in (2), with an example from the data in Figure 2. 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 $w(k) = \sqrt{\frac{2}{L}}$ 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 $\frac{\sum_{n=1}^{L} x(n)}{\sqrt{L}}$, the sum of all data points in the trajectory divided by the square root of the number of data points. 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

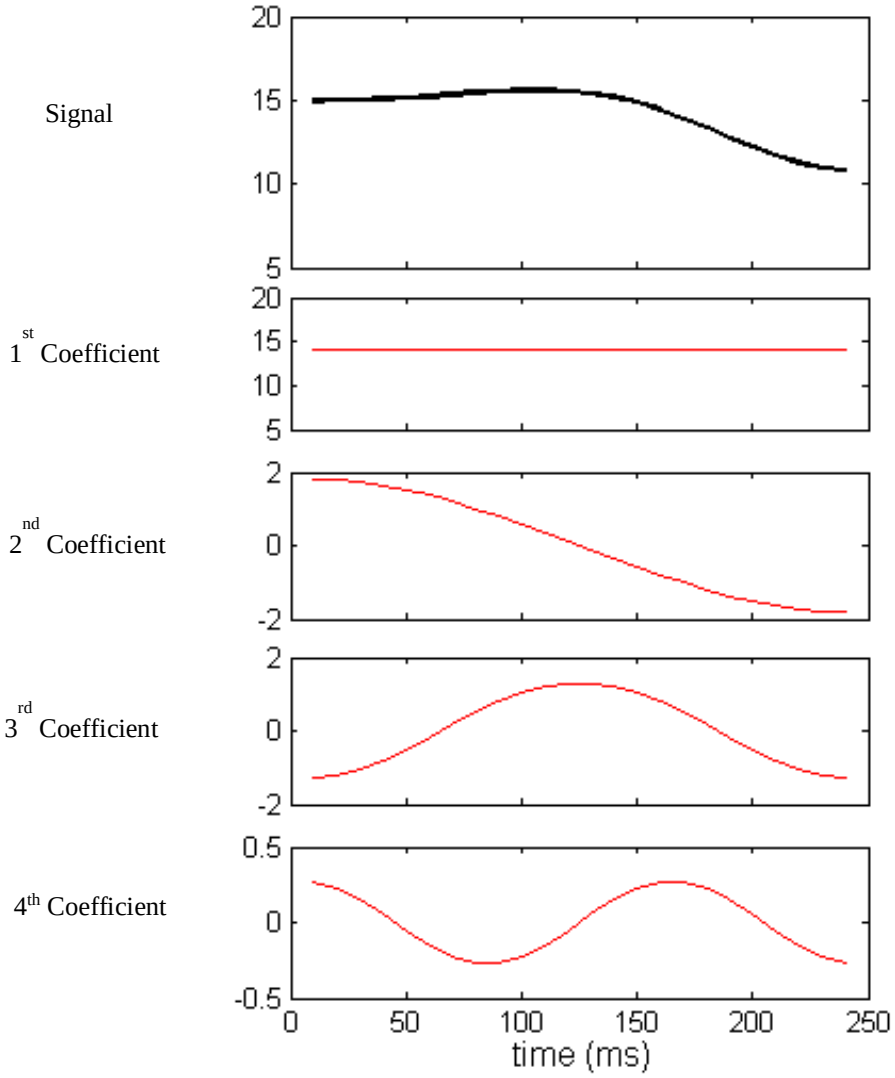
$$y(k) = w(k) \sum_{n-1}^{L} \cos \frac{\pi(2n-1)(k-1)}{2L} \quad k = 1, 2, ...L$$

where

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

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 occlusal plane). 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 much smaller, specifying modulations that are less than ±0.5 mm.

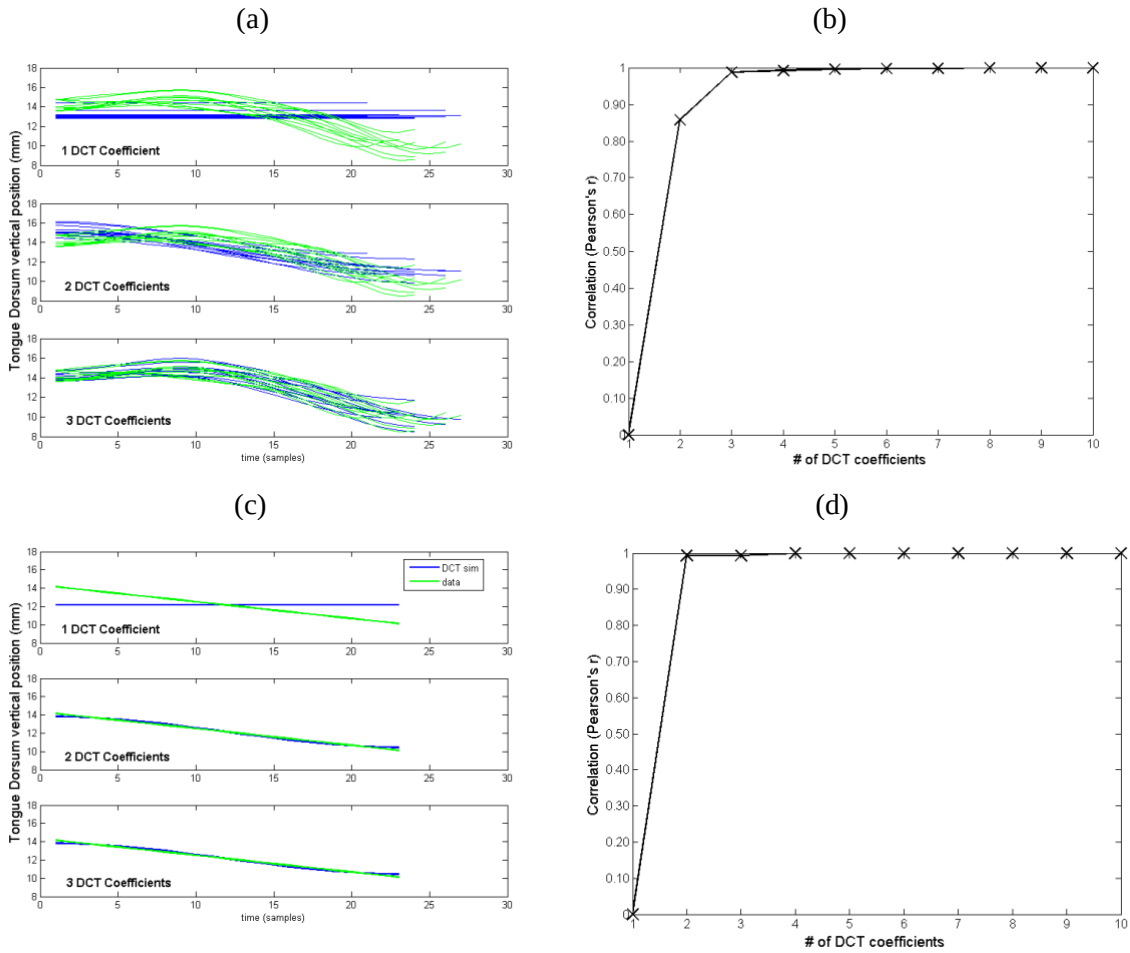
Figure 3: An illustration of DCT components for a TD trajectory spanning the VCuCV portion of /e#ɸusoku/. The top panel shows the signal. The bottom four panels show individual DCT components contributing to the signal.



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. Figure 4(b) shows how the correlation between the raw data and the simulated data increases with the number of DCT coefficients. The Pearson correlation *r* is nearly zero with just one DCT coefficient. With two DCT coefficients, the correlation rises to .858. Increasing the number of parameters to three increases *r* up to .989. Subsequent increases in the number of DCT coefficients exact only more marginal improvement—the correlation with four DCT coefficients is .992; the correlation with six is .998. As illustrated here, one general advantage of the DCT analysis is that for each number of DCT coefficients, we can generate the predicted trajectories, compare them with actual trajectories, and examine the goodness-of-fit. Figure 4(a) illustrates the goodness of fit token by token. The same set of eleven /ɸuzoku/ tokens from 2 is re-displayed in green. The blue lines show trajectories that were simulated from DCT coefficients. With

three coefficients, the blue and the green trajectories overlap almost completely, illustrating nearly lossless compression of the trajectories.

Figure 4: panels for (a) show the raw data (green lines) and simulated trajectories (blue lines) for 11 tokens of the VCuCV portion of /e#ɸuzoku/. The trajectories span the VCuCV interval under analysis. Simulated trajectories (blue lines) were based on one (top), two (middle) and three (bottom) DCT coefficients. (b) shows the Pearson correlation between raw data and simulated data as a function of the number of DCT coefficients employed in data representation. (c) shows a linear trajectory (green line) for the same data along with DCT components fit to the linear trajectory. (d) shows the correlation between the linear trajectory and trajectories simulated based on different numbers of DCT coefficients. For the linear trajectory, high precision is obtained with just two DCT coefficients.



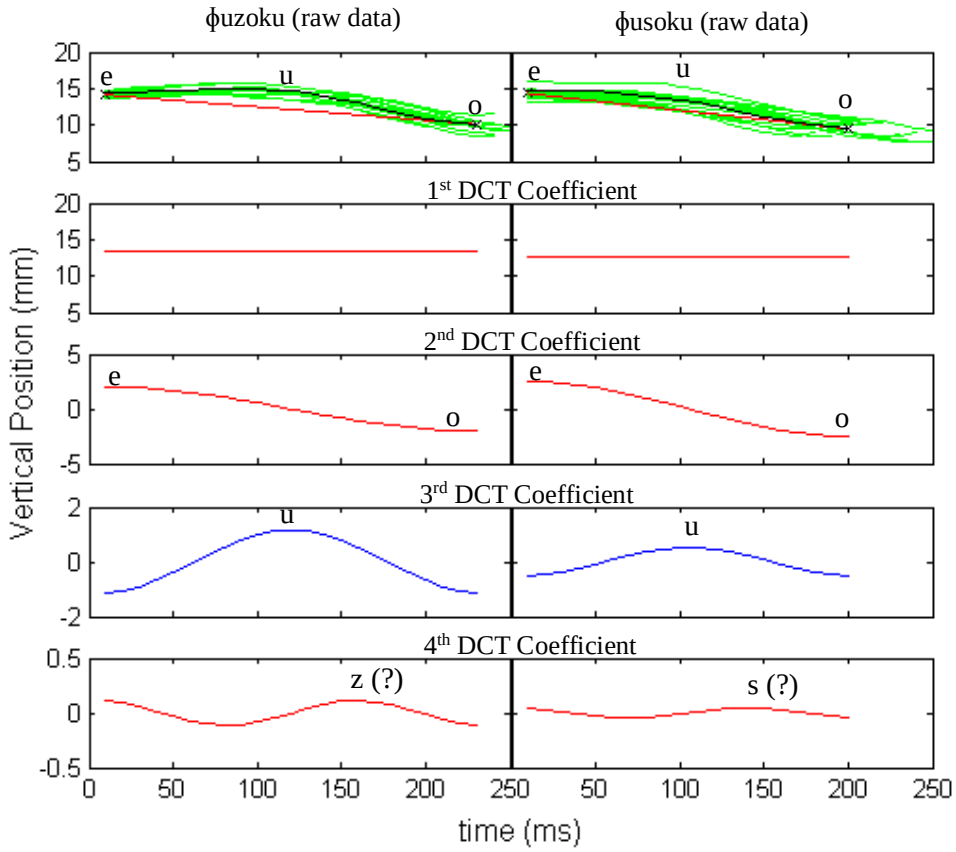
By using DCT, we can reduce the dimensionality of the data without loss of precision. Figure 4 indicates that three DCT coefficients are sufficient to retain a detailed phonetic representation of the TD signal, for the Japanese case under discussion. These DCT coefficients have plausible linguistic interpretations, on which we now elaborate.

Figure 5 re-displays the /ɸusoku/~/ɸuzoku/ data along with mean DCT components fit to the data. Here, four DCT coefficients are shown. 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). This average trajectory was computed by averaging DCT coefficients. The average trajectory (black line) is closer to the linear trajectory for /ɸusoku/ (right) than for /ɸusoku/ (left). More importantly, each of the DCT coefficients has a plausible 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 generally captures a fall in TD height from /e/ to /o/. This component is thus likely to represent the vowel-to-vowel transition, which is similar for both words. Vowel to vowel intervals have long provided building blocks for speech production 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 roughly consistent with coarticulatory effects of coronal consonants /s/ and /z/ on TD height but is so small that it is under the average measurement error of the NDI system (Berry, 2011). We therefore proceed in modelling the data with three DCT coefficients.

In determining the number of DCT components used for analysis, we emphasize that the choice of using three 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) three 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. Criteria (2) is particularly important for constructing a stochastic representational space enabling simulation and classification of phonetic data in terms of phonological hypotheses, as it serves to evaluate whether the statistical dependencies picked up by DCT are indeed those that are a consequence of phonological control of articulation.

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. Only the target VCuCV interval is shown.



To summarize, the first computational step is to express TD trajectories over /VCuCV/ sequences in frequency space, as the sum of three DCT components. 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).

3.2 Stochastic sampling

The next computational tool 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). Building on the preceding section, 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 associated with the target segment.[4] Parsing gestural landmarks often relies

[4] We do parse specific articulatory landmarks for non-target segments, V_1 and V_3, the vowels flanking the target segment, which are used to delimit the start and end of the TD trajectory across V_1CuCV_3.

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 such data, selecting a single point in time that corresponds to the vowel is largely arbitrary. 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 (see Figure 5). If there is no independent TD height target for /u/, i.e., H3 in (1), then the tongue dorsum position should follow a smooth 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 three DCT coefficients to the straight line from V_1 to V_3. Figure 4(c,d) shows that the linear trajectory can also be captured with high precision with a small number of DCT coefficients. 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. Formalized as distributions over three DCT coefficients, corresponding to the middle three 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.[5]

This computational method expresses the targetless hypothesis in the phonetic dimensions of the data, specifically the TD height over time, as phonological control of the vocal tract passes from one vowel to the next. Table 2 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 the standard deviation in parenthesis. The bottom two rows provide the parameters for the targetless hypothesis for the same data. The mean parameters come from a three-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 nearly zero, for the targetless hypothesis, indicating no rise from the trajectory defined by Co2 (see also Figure 5). 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.

[5] An anonymous reviewer asked what the results would look like if we assumed that the targetless trajectory was more variable than the vowel. We have provided simulation results addressing this question in the supplementary materials.

	Distributions over DCT coefficients		
	Co1	Co2	Co3
/ɸusoku/ (raw)	56.5 (6.91)	7.09 (3.93)	-1.80 (0.93)
/ɸuzoku/ (raw)	64.6 (5.03)	7.06 (2.42)	-3.41 (2.15)
/ɸusoku/ (simulated)	53.2 (6.91)	5.18 (3.93)	-0.15 (0.93)
/ɸuzoku/ (simulated)	59.9 (5.03)	5.92 (2.42)	-0.07 (2.15)

Table 2 Mean and standard deviation (in parentheses) of DCT coefficients.

Having defined distributions over DCT coefficients, we can sample from the DCT coefficients to simulate trajectories corresponding to the targetless 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 $k = 3$ DCT coefficients.

(3) Numerical expression of inverse Discrete Cosine Transform

$$y(k) \sim N(\mu(k), \sigma(k))$$

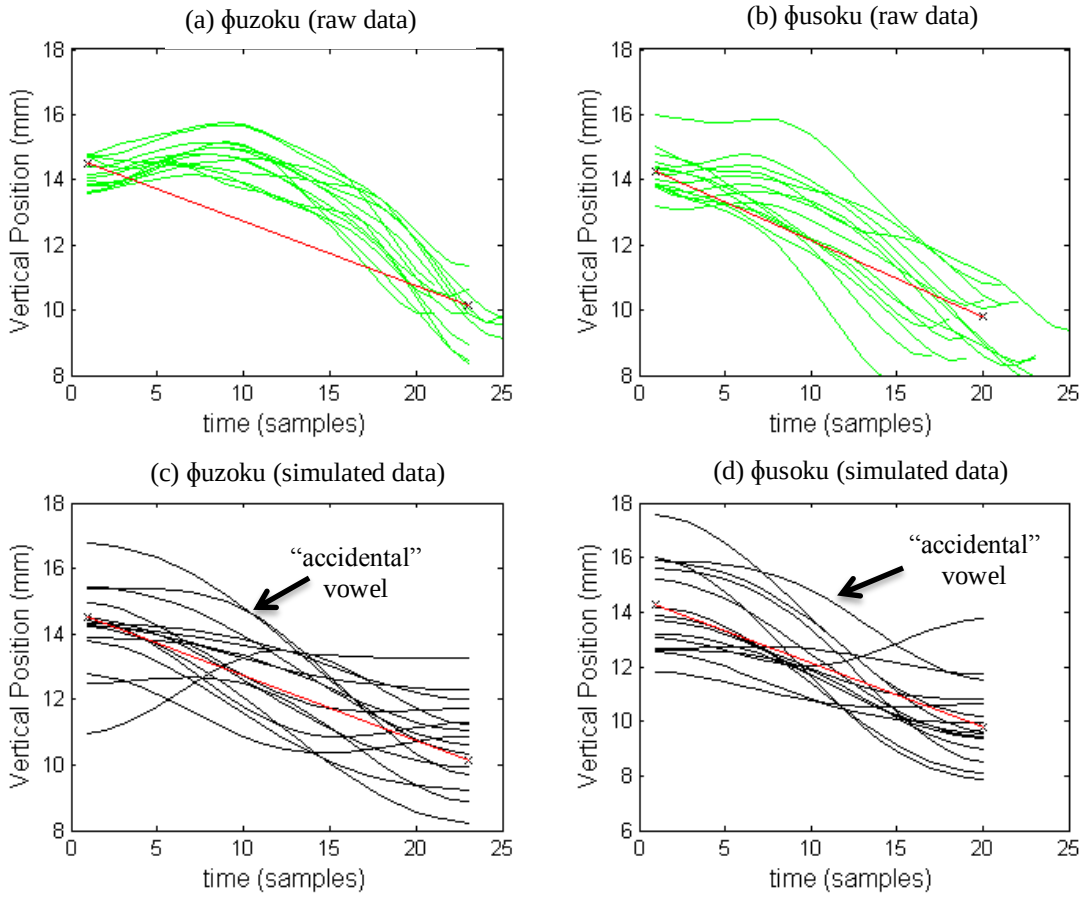
$$x(n) = \sum_{n-1}^{L} w(k) y(k) \cos \frac{\pi(2n-1)(k-1)}{2L} \quad n = 1, 2, \ldots L$$

where

$$\begin{cases} \frac{1}{\sqrt{L}} & k = 1 \\ \sqrt{\frac{2}{L}} & 2 \leq k \leq L \end{cases}$$

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 /ɸuzoku/ are shown on the left; the trajectories from /ɸusoku/ are on the right. The 'x''s denote average vowel targets for 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 /ɸusoku/ tends to be closer to the line than does TD height in /ɸuzoku/, 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 2. 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, because the distributions over DCT coefficients define the same range of variability as is present in the TD trajectories, which are also not perfectly linear.

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)). All figures show only the target /V_1CuCV_2/ interval. The individual trajectories differ in length because they are delimited by the targets of V_1 and V_2.



Focusing on Figure 6(c,d), 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 sampled from the linear interpolation trajectory, 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 in the simulations 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.

We can now statistically adjudicate between three of our four hypotheses in (1). In asking whether the vowel is targetless, 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 noisy targetless trajectory in 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

differs 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 /ɸusoku/ to the simulated targetless trajectory to test whether the data significantly differs 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). 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 smooth interpolation.

4.0 Classification

Many 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 targetlessness 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). To distinguish between these possibilities, we make use of the distributions built for simulation to classifier phonetic data in terms of phonological structure. For this purpose, we use 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 (with three 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 formula is provided in (4). The output of the formula, i.e. $p(T|Co_1, \dots, Co_n)$, is the posterior probability of targetlessness, which designates the probability of a targetless articulation, given the 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 $\prod_{i=1}^{n} p(Co_i|T)$. 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: $\prod_{i=1}^{n} p(Co_i)$.[6]

(4) Formula for Naïve Bayesian Classifier

$$\prod_{i=1}^{n} p(T|Co_1, \dots, Co_n) = \frac{p(T) \times \prod_{i=1}^{n} p(Co_i|T)}{\prod_{i=1}^{n} p(Co_i)}$$

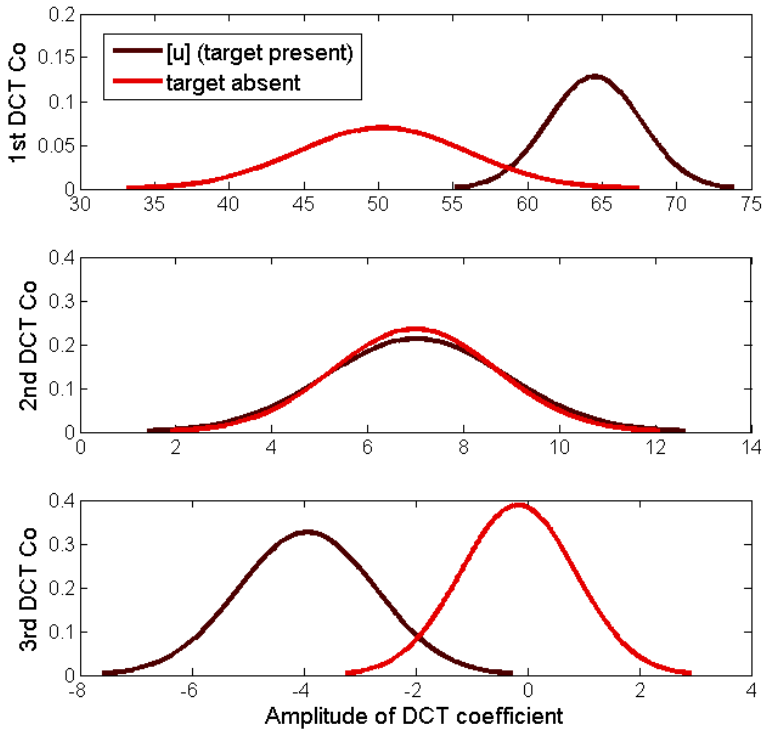
where Co_i is i-th DCT coefficient (and $n = 3$ for the case at hand)

[6] The denominator guarantees that the posterior 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.

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 categories *target present*, H1 in (1), and *target absent*, H3 in (1) hypotheses. Thus, $p(T)$ is set to .5.[7] 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 three 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(c,d)). 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 $\prod_{i=1}^{n} p(Co_i|T)$) 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 baseline (target present) hypothesis, based on [ɸuzoku] tokens; the red lines show the "target absent" hypothesis, based on noisy simulation of linear interpolation. As can be seen from Figure 7, the PDF of the 2nd DCT coefficient is heavily overlapped. The main differences between the presence and absence of a vowel are found in the PDF of the 1st and the 3rd DCT coefficients. This is expected since the first DCT coefficient is related to the average TD height across the trajectory and the 3rd 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.

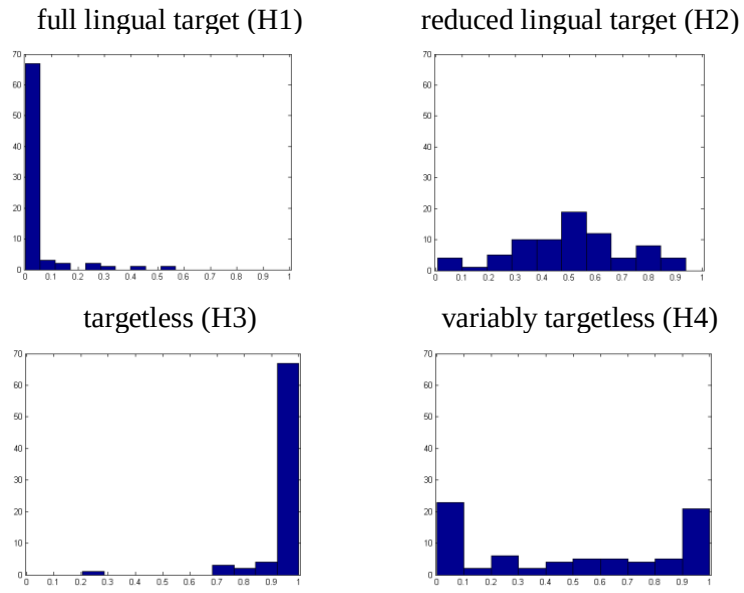
Figure 7: The probability distribution functions for DCT coefficients given the targetless hypothesis and the alternative target present hypothesis based on the data in Figure 2.



[7] 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.

Four possible patterns, displayed as histograms over posterior probabilities of targetlessness, are illustrated in Figure 8. These hypothetical results correspond to the four hypotheses in (1). The histogram in the top left panel was obtained by submitting /ɸuzoku/ 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 targetlessness pattern. We created this pattern by sampling at random from distributions characterizing the target present data and the target absent data.

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 supplementary material A for different instantiations of H3 which assume different degrees of variability.



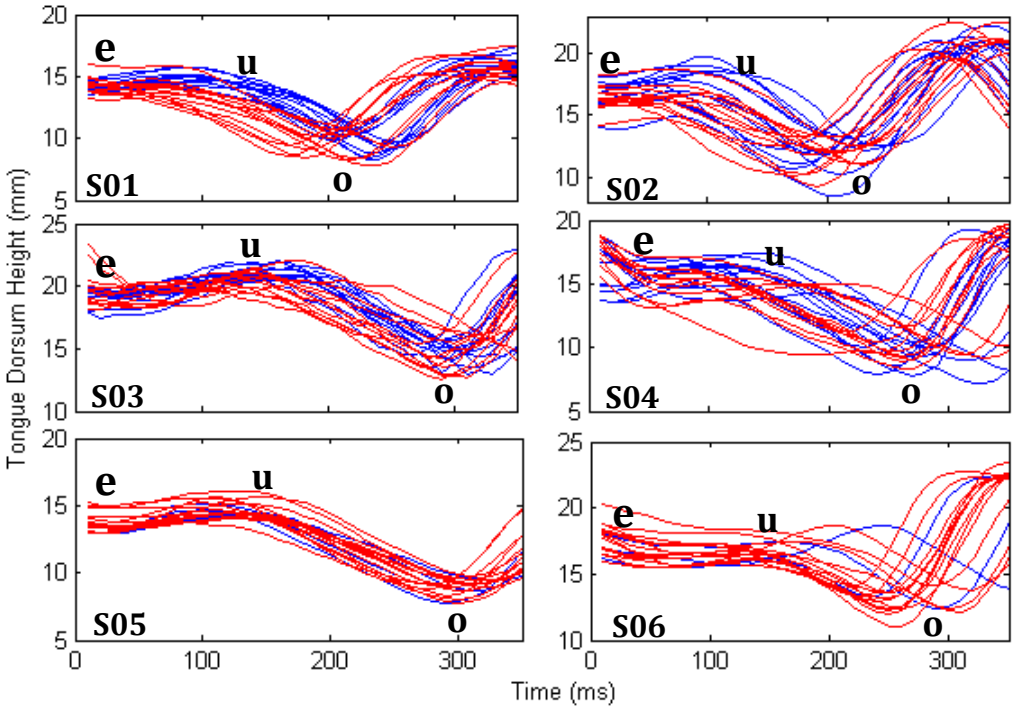
5.0 Results

5.1 Simulation results

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 3 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. There is variation across speakers in the degree to which red and blue lines overlap. They are very closely overlapped for S05 and S06 but less so for other speakers. As described in the methods, 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/ (red lines) and /ɸuzoku/ (blue lines). The figure shows a fixed window beginning with the /e/ of the carrier phrase and extending for 350 ms, which is longer than the analysis window (note, for most tokens a rise in TD height for /k/ can be observed following the TD minima for /o/).



Following the method introduced in section 3, we fit DCT coefficients to each TD height trajectory and defined a targetless trajectory. 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, 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 $\alpha = 0.00138$ (.05/36), where 36 is the total number of comparisons: 6 speakers × 2 item pairs × 3 comparisons per item pair.

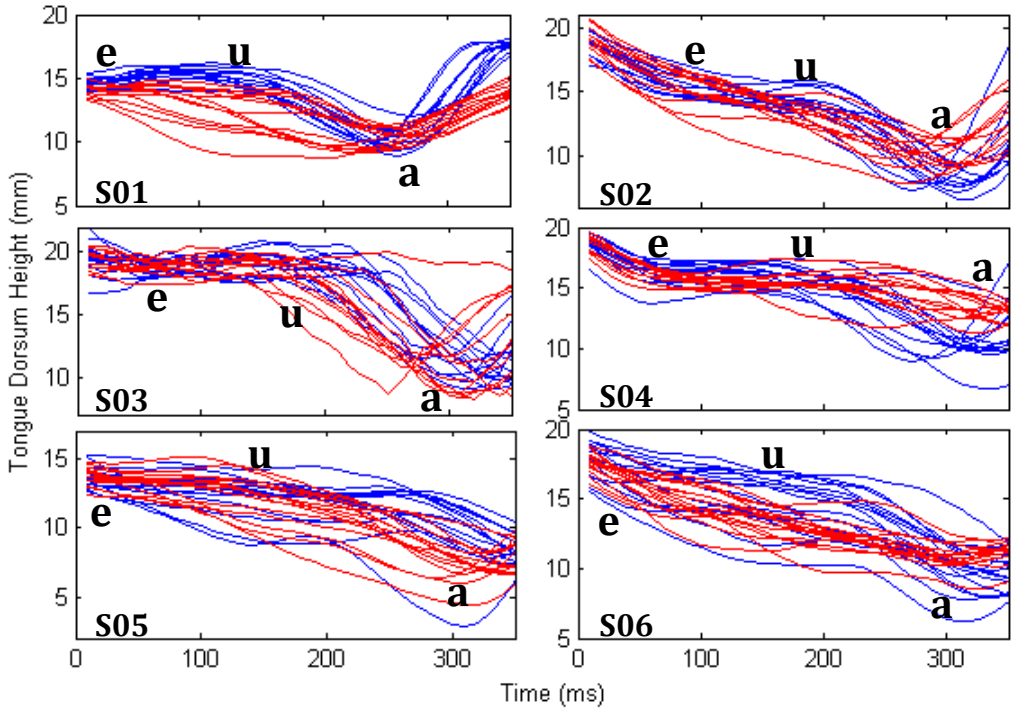
In Table 3, significant differences are indicated by asterisk. Of the six participants, four produced reliable differences between the vowels in /ɸusoku/ and /ɸuzoku/. For all six participants, the voiced vowel in /ɸuzoku/ significantly differed from the targetless trajectory. Of the four speakers who 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 speaker, S04, produced the devoiced vowel with a TD height trajectory that could not be distinguished from the targetless trajectory.

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

Table 3: MANOVA results for /ɸusoku/ and /ɸuzoku/ for each speaker.

Figure 10 shows the trajectory of TD 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 fitted DCT components to each trajectory, simulated a targetless trajectory and compared these via MANOVA. The results appear in Table 4.

Figure 10: Change in TD height (y-axis) over time (x-axis) for /ʃutaisei/ (red lines) and /ʃudaika/ (blue lines). The figure shows a fixed window of 350 ms, beginning with the /e/ of the carrier phrase.



As shown in Table 4, all six speakers produced /ʃutaisei/ and /ʃudaika/ with significantly different TD trajectories. Moreover, of the six speakers, only one, S05, produced /ʃutaisei/ differently 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 Boneferroni correction (p = 0.009 where α = 0.001).

Speaker	Comparison	*df*	*F*	*p*	lambda
S01	ʃu̲taisei ~ ʃu̲daika	21	22.0	0.0002*	0.2949
	ʃu̲daika ~null	21	47.9	0.0000*	0.0703
	ʃu̲taisei ~null	21	3.5	0.5050	0.8250
S02	ʃu̲taisei ~ ʃu̲daika	25	20.3	0.0004*	0.3980
	ʃudaika ~null	25	16.9	0.0097	0.4739
	ʃu̲taisei ~null	25	8.2	0.1826	0.6994
S03	ʃu̲taisei ~ ʃu̲daika	23	17.9	0.0013*	0.4078
	ʃu̲daika ~null	23	42.1	0.0000*	0.1232
	ʃu̲taisei ~null	23	16.9	0.0054	0.4357
S04	ʃu̲taisei ~ ʃu̲daika	23	52.3	0.0000*	0.0732
	ʃu̲daika ~null	23	27.2	0.0000*	0.2598
	ʃu̲taisei ~null	23	11.1	0.0378	0.5770
S05	ʃu̲taisei ~ ʃu̲daika	27	22.8	0.0001*	0.3861
	ʃu̲daika ~null	27	21.0	0.0012*	0.4218
	ʃu̲taisei ~null	27	41.5	0.0000*	0.1796
S06	ʃu̲taisei ~ ʃu̲daika	29	49.3	0.0000*	0.1504
	ʃu̲daika ~null	29	17.8	0.0013*	0.5041
	ʃu̲taisei ~null	29	12.0	0.0176	0.6312

Table 4: MANOVA results for /ʃutaisei/ and /ʃudaika/ for each speaker.

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.[8] One speaker, S04, produced /ɸusoku/ and five speakers produced /ʃutaisei/ 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 /ɸusoku/, 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 /ʃutaisei/, 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 /ɸusoku/ and /ʃutaisei/ for a given speaker. If there is within-speaker optionality, then this assumption is not justified. We next turn to phonological classification of the data 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 Classification results

We submitted each token of /ɸusoku/ shown in Figure 9 and each token of /ʃutaisei/ in Figure 10 to a Bayesian classifier as described in section 4. 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 /ɸusoku/. The left panel aggregates across speakers. The right panel provides a breakdown by speaker. The pattern clearly shows that 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 speakers, rather, it

[8] This reasoning may run the risk of concluding the lack of difference based on null results in statistical hypothesis testing. The Bayesian classification analysis reported below overcomes this problem.

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 targetless trajectory. From the right panel of Figure 11 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 targetlessness but rather in the frequency with which the vowel is targetless. The other speakers, S05 and S06, produced no tokens that were classified as targetless. S05 had just one reduced token, with a targetless probability of 0.60; S06's most reduced token had a targetless probability of just 0.30.

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 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. When we bore down to subject level data (right panel), we see just one speaker, S02, prone to gradient reduction. Outside of S02, the other speakers produce only a small number (three) of tokens in the ambiguous, .3 to .7 probability range. Consistent with the MANOVA results, the individual speaker results indicate that five speakers (including S02) 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 probabilities of targetlessness 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.

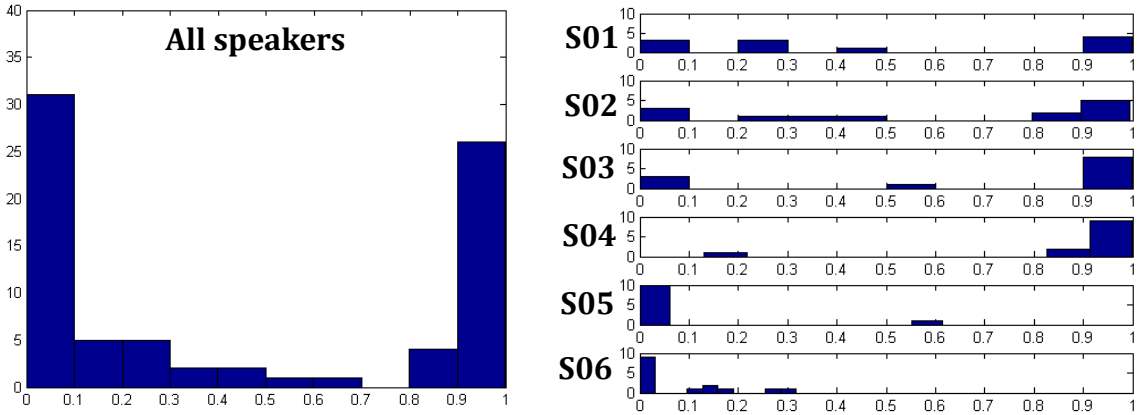
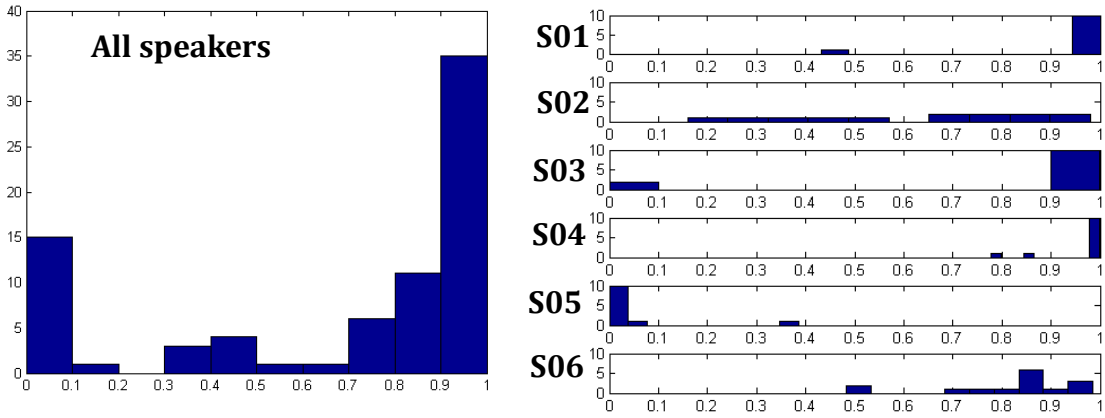


Figure 12: Posterior probabilities of targetlessness for 77 tokens of /ʃutaisei/ from 6 speakers (the TD trajectories shown in Figure 10).



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/. Table 5 shows that this holds true for each speaker: 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.

	Targetless probability	
Speaker	/ʃutaisei/	/ɸusoku/
S01	0.9195	0.5756
S02	0.6646	0.5312
S03	0.7213	0.7185
S04	0.9869	0.8680
S05	0.0273	0.0152
S06	0.6595	0.1281

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 targetless 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 tokens that lack a vowel in the surface representation.

6.0 Discussion

6.1 Summary

We have illustrated a computational approach to the assessment of surface phonological form based on phonetic data. The general strategy was to develop a stochastic representational space that links discrete phonological form to continuous phonetic data through simulation and classification. Our specific proposal is to express phonological hypotheses in terms of distributions over harmonic (frequency) components, extracted using DCT. We showed that DCT compresses the phonetic data into a small number of phonologically relevant parameters that preserve phonetic detail. As a proxy for phonetic interpolation, we defined a linear trajectory between flanking vowels 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. Finally, we used the distributions to assign probabilities of targetlessness to unseen data, according Bayes' rule. We have illustrated the method with TD movements produced by Tokyo Japanese speakers 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 this pair of words. For all speakers, the probability of producing a vowel without a height target was higher for /ʃutaisei/ than for /ɸusoku/. This difference could be due to resulting consonant cluster phonotactics. Deletion of /u/ in /ʃutaisei/ would give rise to a fricative-stop cluster, [ʃt], which may be a better surface form than the fricative-fricative cluster [ɸs] resulting from /u/ deletion in /ɸusoku/. If we assume that a syllable boundary remains between these surface consonants (for evidence that it does, see Shaw & Kawahara, 2018a), a preference for fricative-stop clusters over fricative-fricative clusters follows from syllable contact laws (e.g., Vennemann, 1988) . Since there is a greater decrease in sonority between the offset of one syllable and the onset of the next, [ʃ.t] is less marked relative to [ɸ.s] (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 (e.g., Berent, Lennertz, Smolensky, & Vaknin-Nusbaum, 2009; Berent, Steriade, Lennertz, & Vaknin, 2007; Davidson & Shaw, 2012).

Consistency across speakers in the relative targetlessness of /ʃutaisei/ and /ɸusoku/ 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) . Some additional discussion of possible factors influencing deletion is taken up in Shaw and Kawahara (2018b), where the analysis developed here is extended to more words and presented alongside converging phonetic evidence for variable deletion of lingual articulatory targets.

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 useful in augmenting auditory impressions of phonological form, particularly from researchers who are non-native speakers of the target language. However, phonetic heuristics 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 Tashlhiyt Berber, for example, greatly overestimates the frequency of vowels in the language (Ridouane, 2008). A Berber word like /t-bdg/ "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 (Davidson, 2006a; Lee, Byrd, & Krivokapić, 2006; Wieling et al., 2016). For example, SSANOVAs can be used to compare populations of splines (Gu, 2013), such as tongue shapes, TD trajectories, or even more complex derived variables (e.g., change in tongue curvature over time Ying et al., 2017) and have been applied to various phonetic signals (Davidson, 2006a). Similarly, Functional Data Analysis (FDA) fits a series of splines to time aligned signals, and has been shown to differentiate temporal differences associated with prosodic context (Lee et al., 2006). Another approach is Generalized Additive Models (GAMs) which have been developed to support random effects, e.g., of talker. GAMs have recently been applied to EMA data, detecting dialect variation on the basis of movement trajectories from large samples of speakers (Wieling et al., 2016). These techniques can all be used to assess significant differences between populations of trajectories, such as those produced in different prosodic contexts or by speakers of different regional accents. 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 (e.g., 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 (or to the phonological difference of theoretical interest).

A third approach is to use one part of the phonetic signal to predict another. For example, 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. Because following syllables are unspecified, there is a general decline in f0 toward the L tone in the next Accentual Phrase. They show that the longer the duration between H and L, the shallower the slope of the f0. 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, # of intervening syllables and the slope of the f0 fall, constitutes an argument for the tonal targetlessness of intervening syllables.

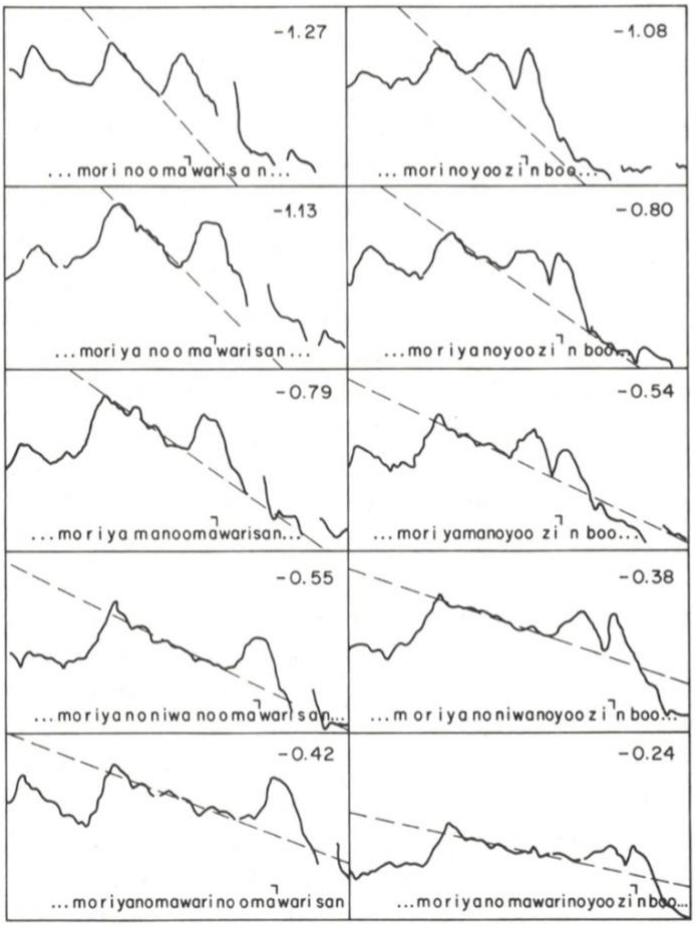


Figure 13: An illustrative figure from Pierrehumbert and Beckman (1988), which was used to argue for tonal underspecification. There is a correlation between the distance of the H tone and the L tone on the one hand, and the slopes between the two tones.

This specific correlation requires manipulating the duration of the hypothesized "targetless" material, which may not always be possible[9], 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 /pV$_1$pəpV$_2$p/ 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. This same approach has been generalized to assess vowel specification on the basis of formant trajectories (Choi, 1995). Choi (1995) demonstrates that the F2 of Marshallese vowels can be predicted by flanking consonants and, therefore, argues that they are unspecified for backness.

[9] Manipulation of speech rate (Solé, 1992; Strycharczuk, Van'T Veer, Bruil, & Linke, 2014) has also been used to probe phonological specification, the key assumption being that only phonologically specified features (not mechanical consequences of coarticulation) maintain proportional influence over the phonetic signal across speech rates.

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 compared the simulated data to the experimental data. They found 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 (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. This can of course be combined with other methods described above. For example, in their investigation of the effects of prosodic structure on articulation, Parrell, Lee, & Byrd (2013) proceed by first simulating trajectories from the TaDA articulatory synthesizer under different prosodic conditions. They tested their FDA-based measure of prosodically-dictated temporal modulation on the simulated data to verify that it picked up the *a priori* known prosodic differences before extending the measure to investigate prosodic effects in naturally produced speech. Like our approach, this method relies on phonologically-informed simulation to guide statistical analysis of the data in terms of phonological structure.

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 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), Shaw et al. (2016), and Blackwood Ximenes et al. (2017) 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, min/max velocity, the temporal midpoint of voicing, etc., "target" selection introduces a researcher degree of freedom. Our toolkit alleviates the necessity of picking points in time associated with the target phonological structure. This aspect of the 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. One assumption that we have adopted here 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. Perhaps most importantly, the number of parameters in our representation of the signal is small, and each has a phonological interpretation. This property contrasts with GAMs, FDA and other powerful algorithms capable of fitting non-linear data and it, in particular, facilitates a phonological interpretation of the phonetic signal.

6.4 Broader applications

Although the computational toolkit we have assembled to assess 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 approach does not dictate such categorical outcomes (see Figure 8 and supplementary material A). With respect to deletion of high vowels 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.Although there is a long line of research on formal architectures that can model variable phonological processes (for overview, see Coetzee & Pater, 2011), the development of formal tools for assessing whether the data require a phonological solution lags behind. We are optimistic about the prospects of applying our computational toolkit to a wider variety of phenomena and curious about the extent to which other cases of "phonetic reduction" are actually manifestations of optional phonological processes.[10] We strongly hope that the proposed toolkit will be used broadly in reassessing alleged cases of reduction to test whether they should be modelled as reduction or as optional processes of phonological deletion/targetlessness.

As discussed in the introduction, the toolkit is designed to address the general issue of phonetic underspecification, whether the source be phonological deletion or lexical under-/non-specification. 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 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 informative. For example, the tradeoff between signal length and f0 slope identified by Pierrehumbert and Beckman (1988) and shown in Figure 12 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 three 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 Supplementary Material A). The broader point is that our approach is flexible. Although it is clear that the representational space dealing with simulation and classification is stochastic in nature, our 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

[10] For another case of categorical but optional phonology, see Strycharczuk *et al.* (2014) for results on voicing in Quito Spanish fricatives.

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 the other, as in, for example, non-native speech production (Davidson, 2010; Wilson & Davidson, 2013). 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 a set of computational tools to assess the presence vs. absence of phonological specification in phonetic data. From end to end, the set of tools, consisting of Discrete Cosine Transform, stochastic sampling, and Bayesian classification, does so without requiring explicit labelling of the target structure. 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 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.

Acknowledgements

We like to thank audiences at ICU, RIKEN, Yale, PAIK, JK 2016, and the Seoul International Conference on phonology. Comments from the associate editor and four anonymous reviewers were very helpful in improving the argumentation of this paper. This research was funded by JSPS grant #15F15715.

References

Alderete, J. (1995). Winnebago accent and Dorsey's law. In J. Beckman, L. Walsh Dickey, & S. Urbanczyk (Eds.), *Papers in Optimality Theory* (pp. 21-52). Amherst, Mass.: GLSA Publications.

Anttila, A. (1997). Deriving variation from grammar. In F. Hinskens, R. van Hout, & W. L. Wetzels (Eds.), *Variation, Change, and Phonological Theory* (pp. 35-68). Amsterdam: John Benjamins.

Archangeli, D. (1988). Aspects of underspecification theory. *Phonology, 5*, 183-208.

Arvaniti, A., & Ladd, D. R. (2015). Underspecification in intonation revisited: a reply to Xu, Lee, Prom-on and Liu. *Phonology, 32*(03), 537-541.

Bayles, A., Kaplan, A., & Kaplan, A. (2016). Inter-and intra-speaker variation in French schwa. *Glossa: a journal of general linguistics, 1*(1).

Beckman, M. (1982). Segment duration and the 'mora'in Japanese. *Phonetica, 39*(2-3), 113-135.

Beckman, M. (1996). When is a syllable not a syllable? In T. Otake & A. Cutler (Eds.), *Phonological Structure and Language Processing* (pp. 95-124). New York: Mouton de Gruyter.

Beckman, M., & Shoji, A. (1984). Spectral and Perceptual Evidence for CV Coarticulation in Devoiced/si/and/syu/in Japanese. *Phonetica, 41*(2), 61-71.

Berent, I., Lennertz, T., Smolensky, P., & Vaknin-Nusbaum, V. (2009). Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology, 26*, 75-108.

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition, 104*, 591-630.

Berry, J. J. (2011). Accuracy of the NDI wave speech research system. *Journal of Speech, Language, and Hearing Research, 54*(5), 1295-1301.

Blackwood Ximenes, A., Shaw, J., & Carignan, C. (2017). A comparison of acoustic and articulatory methods for analyzing vowel variation across American and Australian dialects of English. *The Journal of Acoustical Society of America, 142*(2), 363-377.

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry, 32*, 45-86.

Browman, & Goldstein, L. (1992). 'Targetless' schwa: An articulatory analysis. In G. Docherty & R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 26-56). Cambridge: Cambridge University Press.

Carré, R., & Chennoukh, S. (1995). Vowel-consonant-vowel modeling bysuperposition of consonant closure on vowel-to-vowel gestures. *Journal of Phonetics, 23*(1), 231-241.

Choi, J. D. (1995). An acoustic-phonetic underspecificationaccount of Marshallese vowel allophony. *Journal of Phonetics, 23*(3), 323-347.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Coetzee, A., & Pater, J. (2011). The place of variation in phonological theory. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The Handbook of Phonological Theory*. Oxford: Blackwell.

Coetzee, A. W., & Kawahara, S. (2013). Frequency biases in phonological variation. *Natural Language & Linguistic Theory, 31*(1), 47-89.

Cohen-Priva, U. (2017). Informativity and the actuation of lenition. *Language, 93*(3), 569-597.

Cohn, A. C. (1993). Nasalisation in English: phonology or phonetics. *Phonology, 10*(01), 43-81.

Coleman, J. S. (2001). The phonetics and phonology of Tashlhiyt Berber syllabic consonants. *Transactions of the Philological Society, 99*, 29-64.

Davidson, L. (2006a). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variancea). *The Journal of the acoustical society of America, 120*(1), 407-415.

Davidson, L. (2006b). Schwa elision in fast speech: Segmental deletion or gestural overlap? *Phonetica, 63*(2-3), 79-112.

Davidson, L. (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics, 38*(2), 272-288.

Davidson, L., & Shaw, J. A. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics, 40*(2), 234-248.

Davis, S., & Baertsch, K. (2010). On the relationship between codas and onset clusters *Handbook of the syllable* (pp. 71-98): Brill.
Dell, F., & Elmedlaoui, M. (1985). Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics, 7*, 105-130.
Elvin, J., Williams, D., & Escudero, P. (2016). Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *The Journal of the acoustical society of America, 140*(1), 576-581.
Fougeron, C., & Ridouane, R. (2011). Schwa elements in Tashlhiyt word-initial clusters. *Journal of Laboratory Phonology*(2), 1-26.
Fujimoto, M. (2015). Chapter 4: Vowel devoicing. In H. Kubozono (Ed.), *The handbook of Japanese phonetics and phonology*. Berlin: Mouton de Gruyter.
Gafos, Charlow, S., Shaw, J. A., & Hoole, P. (2014). Stochastic time analysis of syllable-referential intervals and simplex onsets. *Journal of Phonetics, 44*, 152-166.
Gafos, A. (2002). A grammar of gestural coordination. *Natural Language and Linguistic Theory, 20*, 269-337.
Gouskova, M. (2004). Relational Markedness in OT: The Case of Syllable Contact. *Phonology, 21*(2), 201-250.
Gu, C. (2013). *Smoothing spline ANOVA models* (Vol. 297): Springer Science & Business Media.
Guy, G. (1997). Competence, performance, and the generative grammar of variation. In F. Hinskens, R. van Hout, & W. L. Wetzels (Eds.), *Variation, Change, and Phonological Theory* (pp. 125-143). Amsterdam: John Benjamins.
Hale, K., & Eagle, J. W. (1980). A preliminary metrical account of Winnebago accent. *International Journal of American Linguistics, 46*(2), 117-132.
Hall, N. (2006). Cross-linguistic patterns of vowel intrusion. *Phonology, 23*, 387-429.
Hall, N. (2013). Acoustic differences between lexical and epenthetic vowels in Lebanese Arabic. *Journal of Phonetics, 41*(2), 133-143.
Hall, N., & Sue, E. (2018). *Hocank (Winnebago) vowel epenthesis: A phonological re-examination in light of phonetic data*. Paper presented at the OCP15, University College London.
Hanson, R. (2010). *A grammar of Yine (Piro)*. (Ph.D.), La Trobe University, Bundoora, Victoria, Australia.
Haraguchi, S. (1977). *The Tone Pattern of Japanese: An Autosegmental Theory of Tonology*. Tokyo: Kaitakusha.
Jain, A. K. (1989). *Fundamentals of digital image processing*: Prentice-Hall, Inc.
Jun, S.-A. (2014). *Prosodic typology II: the phonology of intonation and phrasing* (Vol. 2): Oxford University Press.
Jun, S.-A., & Beckman, M. (1993). *A gestural-overlap analysis of vowel devoicing in Japanese and Korean.* Paper presented at the 67th annual meeting of the Linguistic Society of America, Los Angeles.
Jun, S.-A., Beckman, M. E., & Lee, H.-J. (1998). Fiberscopic evidence for the influence on vowel devoicing of the glottal configurations for Korean obstruents. *UCLA Working Papers in Phonetics*, 43-68.
Kawahara, S. (2015). A catalogue of phonological opacity in Japanese: Version 1.2. *慶応義塾大学言語文化研究所紀要*(46), 145-174.
Kawakami, S. (1977). Outline of Japanese Phonetics [written in Japanese as" Nihongo Onsei Gaisetsu"]: Tokyo: Oofuu-sha.
Keating, P. (1988). Underspecification in phonetics. *Phonology, 5*, 275-292.
Kondo, M. (2001). Vowel Devoicing and Syllable Structure in Japanese. In M. Nakayama & C. J. Quinn (Eds.), *Japanse/Korean Linguistics* (Vol. 9). Stanford: CSIL.
Ladd, D. R., & Jun, S.-A. (2008). Prosodic typology: the phonology of intonation and phrasing: JSTOR.
Lammert, A., Goldstein, L., Ramanarayanan, V., & Narayanan, S. (2014). Gestural control in the English past-tense suffix: an articulatory study using real-time MRI. *Phonetica, 71*(4), 229-248.

Lee, S., Byrd, D., & Krivokapić, J. (2006). Functional data analysis of prosodic effects on articulatory timing. *The Journal of the acoustical society of America, 119*(3), 1666-1671.
Mooshammer, C., Hoole, P., & Kühnert, B. (1995). On loops. *Journal of Phonetics, 23*(1), 3-21.
Mrayati, M., Carré, R., & Guérin, B. (1988). Distinctive regions and modes: A new theory of speech production. *Speech Communication, 7*, 257-286.
Myers, S. (1998). Surface underspecification of tone in Chichewa. *Phonology, 15*, 367-391.
Nielsen, K. Y. (2015). Continuous versus categorical aspects of Japanese consecutive devoicing. *Journal of Phonetics, 52*, 70-88.
Ohman, S. (1966). Coarticulation in VCV utterances. *Journal of Acoustical Society of America, 39*, 151-168.
Parrell, B., Lee, S., & Byrd, D. (2013). Evaluation of prosodic juncture strength using functional data analysis. *Journal of Phonetics, 41*(6), 442-452.
Pierrehumbert, J. (1980). *The Phonetics and Phonology of English Intonation.* (Ph. D. Dissertation), MIT, Cambridge, Mass.
Poser, W. J. (1990). Evidence for foot structure in Japanese. *Language, 66*, 78-105.
Recasens, D., & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *The Journal of the acoustical society of America, 125*(4), 2288-2298.
Ridouane, R. (2008). Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber. *Phonology, 25*(02), 321-359. doi:doi:10.1017/S0952675708001498
Shaw, J. A., Best, C. T., Docherty, G., Evans, B. G., Foulkes, P., & Hay, J. (to appear). Resilience of English vowel perception across regional accent variation. *Laboratory Phonology*, 1- 65.
Shaw, J. A., Chen, W.-r., Proctor, M. I., & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research, 59*(6), S1566-S1574.
Shaw, J. A., & Davidson, L. (2011). Perceptual similarity in input–output mappings: A computational/experimental study of non-native speech production. *Lingua, 121*(8), 1344-1358.
Shaw, J. A., & Gafos, A. I. (2010). *Quantitative evaluation of competing syllable parses.* Paper presented at the 11th Meeting of the Association for Computational Linguistics Special Interest Group on Computational Morphology and Phonology, Uppsala, Sweden.
Shaw, J. A., & Gafos, A. I. (2015). Stochastic Time Models of Syllable Structure. *PLoS One, 10*(5), e0124714.
Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2009). Syllabification in Moroccan Arabic: evidence from patterns of temporal stability in articulation. *Phonology, 26*, 187-215.
Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2011). Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. *Phonology, 28*(3), 455-490.
Shaw, J. A., & Kawahara, S. (2017). Effects of Surprisal and Entropy on Vowel Duration in Japanese. *Language and Speech*, 1-30.
Shaw, J. A., & Kawahara, S. (2018a). *Consequences of High Vowel Deletion for Syllabification in Japanese*. Paper presented at the Annual Meetings on Phonology (AMP 2017), New York University.
Shaw, J. A., & Kawahara, S. (2018b). The lingual articulation of devoiced /u/ in Tokyo Japanese. *Journal of Phonetics, 66*, 100-119.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 0956797611417632.
Smith, C. L. (1995). Prosodic patterns in the coordination of consonant and vowel gestures. In B. Connell & A. Arvaniti (Eds.), *Papers in laboratory phonology IV: phonology and phonetic evidence* (pp. 205-222). Cambridge: Cambridge University Press.

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: a framework for integrating the continuous and the discrete in cognition. *Cognitive Science, 38*(6), 1102-1138.

Solé, M.-J. (1992). Phonetic and Phonological Processes: The Case of Nasalization. *Language and Speech, 35*(1-2), 29-43.

Stanton, J., & Zukoff, S. (to appear). Prosodic identity in copy epenthesis: Evidence for a correspondence-based approach. *Natural Language and Linguistic Theory*.

Strycharczuk, P. (2009). *The interaction of Dorsey's Law and stress: A non-foot based approach.* Paper presented at the CUNY Conference on the Foot. City University of New York, January.

Strycharczuk, P., Van'T Veer, M., Bruil, M., & Linke, K. (2014). Phonetic evidence on phonology–morphosyntax interactions: Sibilant voicing in Quito Spanish. *Journal of Linguistics, 50*(2), 403-452.

Tiede, M. (2005). MVIEW: software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories.

Tsuchida, A. (2001). Japanese vowel devoicing: Cases of consecutive devoicing environments. *Journal of East Asian Linguistics, 10*(3), 225-245.

Vennemann, T. (1988). *Preference laws for syllable structure and the explanation of sound change : with special reference to German, Germanic, Italian, and Latin*. Berlin ; New York: Mouton de Gruyter.

Warner, N., & Arai, T. (2001). Japanese mora-timing: A review. *Phonetica, 58*(1-2), 1-25.

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the acoustical society of America, 106*, 458.

Whang, J. (2014). Effects of predictability on vowel reduction. *Journal of Acoustical Society of America, 135*(4), 2293.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., . . . Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics, 59*, 122-143.

Wilson, C., & Davidson, L. (2013). *Bayesian analysis of non-native cluster production.* Paper presented at the Proceedings of NELS.

Wood, S. (1979). A radiographic analysis of constriction locations for vowels. *Journal of Phonetics, 7*(1), 25-43.

Xu, Y., Lee, A., Prom-on, S., & Liu, F. (2015). Explaining the PENTA model: a reply to Arvaniti and Ladd. *Phonology, 32*(03), 505-535. doi:doi:10.1017/S0952675715000299

Ying, J., Carignan, C., Shaw, J., Proctor, M., Derrick, D., & Best, C. (2017). *Temporal dynamics of lateral channel formation in /l/: 3D EMA data from Australian English*. Paper presented at the Interspeech 2017, Stokholm, Sweden.

Yip, M. (2002). *Tone*: Cambridge University Press.