Measuring phonological distances in a tonal language: an experimental and computational investigation with Cantonese

1 INTRODUCTION

Speakers have consistent intuitions about which sound sequences are more *wordlike* than others in their native language, often thought of as phonotactic knowledge. Native speakers can not only tell which existing sound sequences are more typical (e.g., *bag* [bæg] is more typical than *squad* [skwad] in English) but also judge which non-word sequences are more wordlike (e.g., *bnick* [bnɪk] is better than *bdick* [bdɪk] in English). Researchers have explored the underlying mechanisms of phonotactic knowledge. Among many, the probabilities of the subparts of a lexical item appearing in a language, or phonotactic probabilities, was suggested to be a main factor (Coleman & Pierrehumbert, 1997; Dankovicova, West, Coleman, & Slater, 1998; Frisch, Large, & Pisoni, 2000; Gathercole & Martin, 1996; Vitevitch et al., 1997). The degree to which a sound sequence overlaps with existing words in lexicon, or neighborhood density, was also found to affect wordlikeness judgments (Martin and Gathercole ms in Gathercole & Martin, 1996; Greenberg and Jenkins, 1964; Bailey and Hahn, 2001). Various models have been proposed based upon phonotactic probabilities or neighborhood density in an attempt to predict speakers' wordlikeness judgements. Some models project a phonotactic grammar from the lexicon, from which the degree of wordlikeness is predicted in forms of probability, likelihood, or scalar (the Phonotactic Probability Calculator (Vitevitch & Luce, 2004), the Phonotactic Learner (Hayes & Wilson, 2008), the featural bigram model (Albright, 2009), the syllabic parser (Coleman & Pierrehumbert, 1997), the bigram model (Jurafsky & Martin, 2014)). Other models directly project lexical effects to the well-formedness score without assuming an intervening grammar (the Generalized Neighbourhood Model of Bailey and Hahn, 2001).

Despite fruitful results, most work on speakers' phonotactic knowledge has focused extensively on phoneme sequences, while neglecting suprasegmental features of language such as tone, stress, or intonation. Our research program aims to fill this gap. Our ultimate goal is to propose a model that can learn regulations on sound sequences incorporating both segmental and suprasegmental features and eventually predict human's wordlikeness judgments. The current paper is a first step in this line of research. For suprasegmental features, this paper takes tone as an example. For a factor affecting wordlikeness judgments, the current study builds a methodological background to model neighborhood density. To model neighborhood density in tonal languages, a fundamental thing to establish is proper measurements of phonological distances between words incorporating segmental as well as tonal information.

Phonological distance measures have been applied in numerous linguistic fields, including phonology and phonetics (e.g. Pierrehumbert, 1993; Frisch, Broe and Pierrehumbert, 1997), psycholinguistics (e.g. Gildea & Jurafsky, 1996; Saiegh-Haddad, 2004) (often in the form of neighbourhood density), dialectology (e.g. Nerbonne & Heeringa, 1997; Heeringa, 2004; Heeringa, Kleiweg, Gooskens and Nerbonne, 2006; Tang, 2009; Tang and van Heuven, 2009, 2011, 2015), historical linguistics (e.g. Oakes, 2000), and some older methods of automatic speech recognition (Fisher & Fiscus, 1993). Because of its wide application, a variety of distance metrics have been proposed, which we introduce in Section 2.

Research on phonological distance, again, has been predominantly on segmental features (e.g., Nerbonne & Heeringa, 1997; Heeringa, 2004) and work using suprasegmental features is rare. Although some studies have utilized tonal distance measures with limited discussions on their quality or nature (e.g. Tang and van Heuven, 2009), to the best of our knowledge, few have taken tonal distance metrics themselves as an object of study, such as comparing the quality of different tonal distance metrics. Relevant studies will be discussed in the Section 2.3. To establish proper measurements of phonological distances in tonal languages, this study takes experimental and modeling approaches. Our case study is Cantonese. Specifically, we focus on three questions about phonological distances in Cantonese. First, we explore relative weightings of tonal and segmental distances. In our Bayesian multilevel model, we investigate the posterior distributions of the means of the tonal and segmental distances' weightings. Based on this, we determine whether, on average, speakers tend to weigh one heavier than another. Second, we investigate which phonological distance measures, among the ones described in the literature, match best with human behavior data we obtained from a distance judgment test. Third, we question whether different parts of a word may also be weighted differently in calculating phonological distances in Cantonese, a possibility first raised by Cheng (1997). The comparison of the predictions from various distance measurements with human subject data will lend some insight into the underlying mechanisms of phonological distance measures in tonal languages.

This paper will first provide an overview of the distance metrics that we will test against our experimental data. We then describe the experiment we have performed to obtain distance judgements and present the results of our statistical modelling to answer the three research questions above. Finally, we will discuss implications of the current finding to the studies of phonotactics incorporating suprasegmental features.

2 DISTANCE METRICS[1]

2.1 SEGMENTAL DISTANCE

Phonemic distance. To determine the phonological distance between sound sequences, we first measure the distance between *phonemes.* The simplest approach is to assume no distance between them when they are identical and full distance otherwise (e.g. Tang & van Heuven, 2015; Heeringa, Kleiwing, Gooskens & Nerbonne, 2006; inter alia). This approach does not take the gradient differences between phonemes into account, e.g., /b/ is equidistant to /p^h/ and /ts^h/.

In phonology, there are two other influential methods of measuring phonological distance of phonemes, using phonological features. One is to compute Hamming distances between binary feature vectors of phonemes (e.g. Pierrehumbert, 1993; Gildea and Jurafsky, 1996). The distance is sometimes normalized by dividing by the total number of features. We will always normalize the distance in this way, so that all distances range from 0 to 1, ensuring comparability between metrics. The formula for determining the normalised phonemic distance is as follows:[2]

$$(1)\ Distance_{Hamming} = \frac{Unshared\ feautres\ between\ phonemes}{Total\ number\ of\ phonological\ features}$$

[1] We will not cover distance measures based on historical sound changes (e.g. Oakes, 2000), methods to combine phonological distances to allow for comparison of languages (Ellison & Kirby, 2006), or distance metrics that rely on lists of correspondences between different dialects (Wieling, Margaretha, & Nerbonne, 2012; Wieling, Nerbonne, Bloem, Gooskens, Heeringa, & Baayen, 2014). As our focus is on phonological rather than phonetic distances, we do not discuss purely phonetic distances such as those based on spectrograms (Gooskens and Heeringa, 2004) and cochleagrams (Heeringa, 2004, pp. 79-120); however, one of the phonological distances we discuss, the one based on multivalued features, does claim to have phonetic basis.

[2] Null features are usually thought to be different from both positive and negative values (Pierrehumbert, 1993). We will adopt this assumption in this study, except when using Broe's information gain weighting (see Supplementary Materials 1).

An alternative is to create natural classes from the features, then construct a class membership-based distance metric (Frisch, Broe and Pierrehumbert, 1997). The formula is shown below:[3]

$$(2)\ Distance_{NC} = \frac{Unshared\ natural\ classes}{Total\ number\ of\ natural\ classes}$$

In the distance measurement in (2), the number of unshared natural classes is divided by the total number of natural classes, i.e. the distance between two phonemes is defined by the proportion of unshared natural classes. The exact feature set of Cantonese on which the Hamming distance calculation is based is presented in Table 1 in Supplementary Materials 8.1 with reference to Hayes (2011).

A third way, influential especially in dialectological studies, is to use multivalued features, either categorical or numeric. One may still use the Hamming metric between multivalued feature vectors. If the values are numeric, one can also use Euclidean and Manhattan distances (Nerbonne & Heeringa, 1997[4]). The formulas for Euclidean and Manhattan distance, once again normalized to fall into [0, 1], are as follows, where $f_i(p_j)$ refers to the *i*-th feature value of the *j*-th phoneme:

$$(3)\ Distance_{Euclidean} = \frac{\sqrt{\sum_i \left(f_i(p_1) - f_i(p_2)\right)^2}}{\max_{j,k}\left[\sqrt{\sum_i \left(f_i(p_j) - f_i(p_k)\right)^2}\right]}$$

$$(4)\ Distance_{Manhattan} = \frac{\sum_i |f_i(p_1) - f_i(p_2)|}{\max_{j,k}\left[\sum_i |f_i(p_j) - f_i(p_k)|\right]}$$

In the Euclidean distance measure in (3), phonological distance is calculated by evaluating the square of the difference between the feature values of the two phonemes under comparison and taking the square root of the sum. Manhattan distance sums up the absolute values of the differences between the corresponding feature values of the phoneme pair. Again, the two distances are normalized so that they fall between 0 and 1, this time by dividing the result by the maximum distance. To establish a system of multivalued features (Kessler, 1995; Kondrak, 2002) in Cantonese, we construct a feature matrix for Cantonese based on Ladefoged's (1975) table, which incorporates articulatory and some acoustic features. The features are shown in Table 2 in Supplementary Materials 1. We calculated Hamming, Euclidean and Manhattan distances between each of these feature vectors. When the Hamming distance was calculated, the numeric values are ignored; thus, the result is same as the distance calculated as if the multivalued features were categorical.

An underlying assumption behind the distance metrics so far is that features are weighted equally. However, this assumption may be false. There have been several attempts to assign different weights to the features. One approach (Kondrak, 2002) regards the weights as free parameters and finds weights to optimize the distances' performance. In our present study, this could theoretically be achieved by introducing the weights as parameters in our multilevel model. However, we have refrained from using this approach due to its possibility of increasing the complexity of the model, since each the weight of each feature is a new parameter. Moreover, it

[3] This was originally a similarity measure. It was converted into distance measures by subtracting the maximum similarity by the similarity value. This creates a valid measure of distance, since two identical items will have zero distance between them, whereas two completely distinct items will have maximum distance between them.

[4] They also used a distance based on the Pearson correlation between feature vectors, though Heeringa (2004) points out theoretical problems with this approach, and in Heeringa's perception experiment, the Pearson-based method performed worst by far. Therefore, we have not adopted it.

will add complexity to the model-fitting procedure, since indel distance is dependent on the distances between phonemes (see section 2.1.2).[5] Instead, we tried Nerbonne & Heeringa's (1997) information-theoretic approach. Each feature is multiplied by a weight determined by information gain. For Hamming distances between binary features, we also tried a Broe's (1996) modification of the information gain formula, which takes into account the fact that certain feature values may be null. The formula and Broe's modification are presented in Supplementary Materials 8.1.

Distance between phoneme sequences. The segmental distances between words were computed using the Wagner-Fischer algorithm for Levenshtein distances (Jurafsky & Martin, 2014). Following Nerbonne & Heeringa (2001), indel (i.e. insertion and deletion) cost was set at half the average substitution cost. The metrics considered in our study are summarized below. Numbers below match the numbers of corresponding formulas in Section 2.1:

Abbreviation	Segmental representation	Distance metric (between segments)	Distance metric (between strings)
Simple	None	All-or-nothing	Levenshtein
Natural class	Binary	Natural class (2)	Levenshtein
Binary	Binary	Hamming (1)	Levenshtein
Multivalued (E)	Multivalued	Euclidean (3)	Levenshtein
Multivalued (M)	Multivalued	Manhattan (4)	Levenshtein
Multivalued (H)	Multivalued	Hamming (1)	Levenshtein

Table 1: Summary of distance metrics investigated in this paper

2.2 TONAL DISTANCE

Tonal representations. Of the six tonal representations compared in Yang & Castro (2008), we retained the Chao tone letter, autosegmental, onset-contour, onset-contour-offset and contour-offset representations of tone.[6] The Chao tone letters used were Chao's original proposal for Cantonese, except that tone 1 has been fixed at 55 instead of 53 because 53 is mostly absent in Hong Kong Cantonese (Bauer & Benedict, 1997). The autosegmental representations are based on Yip's (1980) framework. She describes the tonal phonology of Chinese varieties using a two-tiered system, including register, which is either upper (+) or lower (-), and Tone, which consists of two binary features (which can be H or L). The onset and contour representations follow standard descriptions such as Bauer & Benedict (1997), while the offset is extrapolated using the onsets and Chao tone letters. The assumed tonal representations of Cantonese are displayed in Table 2. The tone contours are diagrammed according to the Chao tone letters in Figure 1.

Tone	Chao tone letters	Autosegmental		(Onset)-Contour-(Offset)		
		Register	Tone	Onset	Contour	Offset
1	55	+	HH	H	L	H
2	35	+	LH	M	R	H
3	33	+	LL	M	L	M

[5] Kondrak actually manually modifies the weights through trial and error to optimize the distances. However, not only is this approach time-consuming, but it is also impossible to estimate the standard error of 'estimates' computed this way. Therefore, we do not adopt this approach.

[6] We excluded the Target representation (Xu and Wang, 2001). Xu and Wang propose characterising Mandarin tones by the static and dynamic targets H (high), R (rising), L (low) and F (falling), which would be difficult to replicate in Cantonese since there are multiple rising tones, i.e. the second and fifth tones.

4	21	-	LL	L	F	V
5	23	-	LH	L	R	M
6	22	-	HH	L	L	L

Table 2: A table of five different representations of Cantonese tone tested in our study.

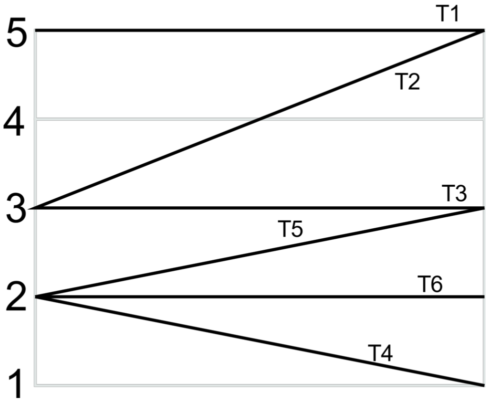

Figure 1: A graphical illustration of the Chao tone letter representations of the six Cantonese tones.

From tonal representations to tonal distance metrics. For the five tonal representations in Table 3, Levenshtein and Hamming distances can be considered identical: we calculated both and found no differences between the two (Table 1-9 in Supplementary Materials 8.1 for the calculated values). Therefore, the Hamming distances discussed below also refers to Levenshtein distances. For Chao tone letters, we also computed Euclidean and Manhattan distances because each tone letter bears its own numeric meaning.

2.3 PREVIOUS STUDIES ON DISTANCE METRIC COMPARISON

Previous studies comparing different distance metrics have largely focused on segmental features of languages. Somers (1998) compared three segmental feature sets defining similarity metrics in the context of aligning phonemes in child language with their adult counterparts: binary articulatory features, Ladefoged-style multivalued features, and a perceptual distance based on frication and pitch. The results show that the perceptual distance performed worst. Heeringa (2004) compares the simple all-or-nothing distance, a binary feature system, two multi-valued feature systems, and a variety of phonetic distance measures. He used them to compare different Norwegian dialects and compares the results to empirically collected perceptions of dialect distance. It was found that the simple system using the all-or nothing distance measure works best, though Kessler (2005) cautioned that this may be because of the binary nature of the task itself. Heeringa also found Euclidean distance to perform better than Manhattan distance. Nerbonne and Heeringa (1997) evaluated the performance of several distance metrics in dialect comparison by comparing the results of the different distances against traditional dialectologists' groupings. They compared the dialect distance results to compare Euclidean, Manhattan and 'Pearson' distance between multivalued features, with or without information gain weighting, and with one-segment or two-segment representations of diphthongs, along with a simple Levenshtein baseline treating distance between phones as all-or-nothing. It was found that the Manhattan distance between multivalued features without information gain weighting and with two-phone representation of diphthongs worked best.

To our knowledge, the only study to compare tonal distance metrics is Yang and Castro (2008), which compared the distances derived from different types of tone representations in Bai and Zhuang. Their results revealed higher Pearson correlation coefficients between the onset-contour, onset-contour-offset and contour-offset representations and mutual intelligibility, compared to Chao tone letter, autosegmental or target representations. Despite the novelty of exploring tonal distance metrics, their approach has a disadvantage of not considering the potential confounding effect of segmental distance. Their experiment involved measuring the intelligibility of texts spoken in different dialects to speakers of other dialects; thus, it is possible that tonal and segmental distance are correlated in their texts. Yet they only assessed the simple Pearson correlation coefficient between the various tonal distance metrics and mutual intelligibility, without partialing out the effect of segmental distance. Therefore, the small differences in Pearson correlation may not be necessarily due to the quality of the tonal distance metrics alone. Tang and van Heuven (2011) also looked at the association between three tonal distance metrics and mutual intelligibility, including Levenshtein distances between Chao tone letters[7] and onset-contour representations as well as a 'perceptually weighted' distance. Though they did not directly compare the metrics, their point estimates of Pearson correlation coefficients seem to suggest that onset-contour representation outperforms the other two measures as well.

Yang and Castro also compare relative importance of tone and distance by fitting multiple regression models with segmental and tonal distances as independent variables and conclude that tones may be more important than segments in Bai. Unfortunately, they only provide *t*-statistics and *p*-values, which give us information about the strength of evidence for tonal and segmental effects on intelligibility, rather than the strength of the effects themselves, which are better represented by point and interval estimates of the regression coefficients. Standard errors were not reported, so we were unable to recover the coefficients in their model or calculate confidence intervals for them. Therefore, it is unclear how great the difference between tone and segmental distance really are from their reported figures. Also, it is often found in perception studies that different people's weightings of different cues may differ wildly (Yu and Zellou, 2018), which Yang and Castro's modelling method (fixed-effects linear regression that does not contain by-subject effects) does not allow for.

2.4 INTERIM CONCLUSION

We have introduced several ways of evaluating segmental distance which assume different ways of calculating distances between phonemes, including the proportion of unshared natural classes, the Hamming distance between their binary and multivalued feature vectors, and the Euclidean and Manhattan distance between multivalued numeric feature vectors. We have also looked at five different types of tonal representations in Cantonese, including the Chao tone letter, autosegmental, onset-contour, onset-contour-offset and contour-offset representations, and explained how we derived distances between tones using them. Our literature review reveals that no systematic investigation has been conducted on phonological distance measures incorporating segments and tones. Against this background, we will now move on to how to estimate the phonological distances of sound sequences in tonal languages with our current study of Cantonese. In order to evaluate the performance of the distance measures, it is necessary to have native speakers' phonological distance judgment data in hand. In our distance judgment test, we presented pairs of two items varying degree of segmental and tonal distances and asked native speakers to judge the similarities between the two items.

[7] In their languages, Chao tone letters do not always contain the same number of pitches, so Hamming distance cannot be calculated.

METHODOLOGY

2.5 JUDGEMENT TASK

Design. We created a question set of 72 monosyllabic and 72 disyllabic sound sequences. The stimuli list is provided in Supplementary Materials 8.2. To ensure that the items are well spread out across different segmental and tonal distances and that the two are not correlated among our experimental items, we chose the distance metrics for segmental and tonal distance during experimental design as follows. To evaluate segmental distance, natural class distances were computed with deletions and insertions set at half of the average substitution cost; we chose natural class distance following Bailey and Hahn (2001). Hamming distances between onset-contour tonal representations were used to evaluate tonal distances following Yang and Castro (2008) and Tang and van Heuven (2011).

A simulation by picking monosyllables from the Hong Kong Cantonese Corpus (Luke and Wong, 2015) at random showed that segmental distances rarely went above 2.5. Hence, we divided segmental distances into four types within the interval of [0, 2.5]: high (>1.67), mid (0.83 to 0.1.67), low (below 0.84 but nonzero) and zero, where each region occupies one third of the interval. We also divided tonal distance into three types: high (1), low (0.5) and zero (0). Each segmental distance and each tonal distance appeared the same number of times in the experimental items. Moreover, each segmental distance-tonal distance pair also appeared the same number of times. We ensured that every possible segment in every position appeared at least once. The design for disyllables was similar, except the first item of each pair was used six times each. Again, we divided segmental distance into four levels: high (>3.33), mid (1.67 to 3.33), low (<1.67) and zero (0). The situation is simply double that of the situation for monosyllables, with each region occupying one-third of the interval [0, 5]. Tonal distance was classified as high (>1), low (<1) or zero (0) as in monosyllables. We ensured that each segmental distance level, tonal distance level, and segmental distance level-tonal distance level pair appeared the same number of times. In both monosyllabic and disyllabic pairs, the first item of each pair is an existing word in Cantonese, whereas the second syllable is an existing word or a non-word for a general interest[8]. When creating the non-words, we excluded absolutely illegal segments in onset, nucleus and coda positions; for example, no fricatives were in coda position, which Cantonese phonotactics disallow. However, we did not consider any other constraints as we view them as constraints to be discovered later through the phonotactic models based on results of the current study.

A native speaker of Hong Kong Cantonese recorded the test items. All the recordings were recorded in a sound-attenuated booth n the authors' institute. The recordings were scaled to 70 dB using the Scale intensity feature in Praat. They were then converted to MP3 format in Audacity, allowing the files to be embedded in HTML5 <audio> tags.

Procedure. The experiment was implemented on the survey website Qualtrics (Qualtrics, 2018). Each experimental item was placed on a separate page. On each page, the participants heard the two audio recordings and judged their similarity using a slider. As we believed that it would be easier to understand similarity than distance, the participants were asked to rate

[8] The present study is a part of an ongoing study to build ongoing models of Cantonese phonotactics. The results of this paper will be used to build a Generalised Neighbourhood Model (GNM) of Cantonese phonotactics (Bailey and Hahn, 2001), which will be tested against phonotactic judgements and compared with other models of Cantonese phonotactics. Therefore, in our experiment, we show participants two recordings in each trial, including one existent word and one word that may or may not be existent, and ask them to judge the distance between the two. This simulates the type of computation that will be performed by the GNM model to determine the neighbourhood density of unseen words.

similarity between the two items instead. The similarities were rated from 0 to 100, where 0 means the two syllables were completely different and 100 means they were identical. The similarities were then converted into distances by subtracting the similarity from 100.

Before the judgment test, a screening task was added in forms of AXB tests to ensure that participants could perceptually distinguish between [n] and [l] initials, which are merging in some Cantonese speakers (Bauer and Benedict, 1997, pp. 24-25), and that they could distinguish between tones 2 and 5, 3 and 6 and 4 and 6, which are merging in some Cantonese speakers (Mok, Zuo and Wong, 2013). If participants submitted an incorrect answer to any question, the questionnaire would terminate and the experiment would not be performed.

Participants. In total, data were collected from 61 anonymous participants after circulating the survey on social media platforms. Twenty-eight participants completed all 144 questions while others submitted the incomplete forms. The data from all of the participants were used to fit the model regardless of they actually completed the experiment. We did not discard any data as the model is able to handle variable sample sizes. Participants who have answered few questions will simply have their estimates shrunk to the subject-level mean, whereas participants who have answered all of the questions will have coefficient estimates influenced largely by their own judgements (Gelman and Hill, 2007).

2.6 DESCRIPTIVE DATA AND SET-UP OF MODELING

We first explored the descriptive patterns in the data through scatterplots of segmental and tonal distances against distance judgements as in Figure 1 in Supplementary Materials 8.3. This examination was meant to arrive at descriptive analyses of our data and to inform our modelling decisions. As shown in the scatterplots, the strength of the relation between distance judgements and the theoretical distances varies greatly among participants. Based on this observation, we chose a multilevel model that allows with an item-level random intercept as well as subject-level random slopes for tonal and segmental distance.

As we know of no statistical package that fits models with our exact specification using a frequentist approach, we opted for Bayesian multilevel modelling (Gelman & Hill, 2007; Nicenboim & Vasishth, 2016). The use of multilevel modelling, similar to frequentist mixed-effects models, allows the *partial* pooling of data from different items and participants. This approach helps us avoid the pitfalls of complete pooling (as Yang and Castro have done), which ignores variability in the data, and no pooling, which ignores information in the data and results in high-variance estimates (Gelman & Hill, 2007; Barth and Kapatsinki, 2018). The exact model is as follows:

$$
\begin{gathered}
(5)\ Y_{ij} \sim N\left(\mu + \alpha_i + \beta_j + \gamma_j t_i + \delta_j s_j, \sigma^2\right), i = 1, \ldots, 72, j = 1, \ldots, n \\
\alpha_i \sim N(0, \sigma_\alpha^2) \\
\begin{bmatrix} \beta_j \\ \gamma_j \\ \delta_j \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu_\gamma \\ \mu_\delta \end{bmatrix}, \begin{bmatrix} \sigma_\beta^2 & \rho_{\beta\gamma}\sigma_\beta\sigma_\gamma & \rho_{\beta\delta}\sigma_\beta\sigma_\delta \\ \rho_{\beta\gamma}\sigma_\beta\sigma_\gamma & \sigma_\gamma^2 & \rho_{\gamma\delta}\sigma_\gamma\sigma_\delta \\ \rho_{\beta\delta}\sigma_\beta\sigma_\delta & \rho_{\gamma\delta}\sigma_\gamma\sigma_\delta & \rho_{\beta\delta}\sigma_\beta\sigma_\delta \end{bmatrix} \right) \\
Y_{ij}^* = \begin{cases} Y_{ij}\ if\ Y_{ij} \le 4 \\ 4\ \ if\ Y_{ij} > 4 \end{cases}
\end{gathered}
$$

where Y_{ij}^* is the *j*th participant's response to the *i*th item, μ is the overall (population-level) intercept, α_i and β_j are respectively item-level and subject-level intercepts centred at zero, μ_γ and μ_δ are the mean coefficients of segmental and tonal distance, and γ_j and δ_j are their subject-level counterparts. ρ_{AB} indicate the population correlation between A and B, and σ_A indicates the standard deviation of A. For disyllables, 4 is replaced with 8 as the maximal distance. The models were fit using the R package brms (Bürkner, 2017; Bürkner, in press), which fits models with an lme4-like syntax using the Stan language (Carpenter et al., 2017). Since we have little evidence for relevant priors on the topic, we relied on the non-informative and weakly informative default priors provided by the package. In the full model, the distance judgements are treated as a right-censored variable (Gelman, Carlin, Stern and Rubin, 2014, pp.225-226). This assumes that there is some underlying distance which may exceed 4 but the data is truncated if the number goes beyond it, the setting of which can be justifiable from the raw data - see the scatterplots of the raw data in Figure 1 in Supplementary Materials 8.3.

We fit the full model to different tonal and segmental distances, comparing the predictive power to find the optimal distance metrics. We also run a model that separates onset, nucleus and coda distance to see if the syllabic components will differ in weighting. Apart from the models we use to compare different tonal and segmental distances, all of the models use natural class distance and Hamming distances between onset-contour tonal representations, since these were used to design the experiment.

3 RESULTS FOR MONOSYLLABLES

Before fitting the models, the judged distances have been scaled to lie between 0 and 4 for ease of interpretation; theoretically, the maximum tonal distance ranges from 0-1 and maximum segmental distance ranges from 0-3, so they can sum up to 4. Comparing the Watanabe-Akaike Information Criterion (WAIC) values of the full model with various reduced models shows that the optimal model is the full model, i.e. containing the item-level random intercept, all subject-level random effects, as well as the censoring assumption. Full justification of the model specification, as well as details of the model comparison procedure, are provided in Supplementary Materials 8.3.

3.1 TONAL AND SEGMENTAL WEIGHTING

Recall that our hypothesis is to test the weights of segmental distance, tonal distance, and their relation. To test this hypothesis more directly, we now turn to the investigation of the relevant model parameters. The population-level estimates[9] of the coefficients of segment was higher than that of the tonal distance (μ_γ and μ_δ); the former is estimated at 1.50 (SE: 0.14, 95% CI: (1.23, 1.77)), which is around twice of the latter, estimated at 0.77 (SE: 0.22, 95% CI: (0.34, 1.19)) respectively. This indicates that the segmental distance plays a more crucial role in predicting the distance judgement in this study than the tonal distance. A 95% credible interval of the difference between the two ($\mu_\gamma - \mu_\delta$), as calculated by the brms package using posterior draws, is (0.25, 1.19) (point estimate: 0.72; SE: 0.24; evidence ratio that $\mu_\gamma - \mu_\delta > 0$: 570.43), indicating very strong evidence that segments are, on average, weighted heavier.

3.2 COMPARISON OF DISTANCE AND SEGMENT MEASURES

The full model was fit to the segmental and tonal representations this study considers, and the WAIC value was computed for each of these models. We find that the best WAIC value was achieved with the onset-contour offset representation with Hamming distances between multivalued features (4682.1), followed by the contour-offset representation with the same segmental distance metric (4683.5) as in Table 3.

[9] Apart from the population-level conclusions, we also find that there is slightly more variation in segmental weighting than tonal weighting, and that we lack strong evidence for correlation between segmental and tonal distance. More details are given in Supplementary Materials 3.

On the tonal side, onset-contour, onset-contour-offset and contour-offset representations consistently outperform the other tonal distance measures, regardless of the segmental distance. One possible reason is the inclusion of contour representation, which is absent in the other two tonal representations we investigated. This suggests that participants may focus on whether the syllable is rising, falling or level when producing distance judgements. On the segmental side, the natural class distance did not perform better than the simple baseline assuming all-or-nothing costs, and the Hamming distance between binary features fared much worse than the rest. However, the multivalued feature performed better, especially when Hamming distances were computed:

	Chao (H)	Chao (M)	Chao (E)	Autosegmental	O-C	O-C-O	C-O
Simple	4764.8	4788.1	4781.7	4780.3	4711.5	4711.3	4709.6
Natural class	4763.5	4786.2	4779.5	4780.4	4727.1	4706.8	4709.4
Binary (H)	4794.2	4817.5	4810.3	4810.6	4762.5	4744.2	4747.2
Multivalued (E)	4752.7	4774.9	4769.8	4770.1	4714.4	4693.3	4696.8
Multivalued (M)	4755.5	4778.8	4774.2	4770.5	4717.8	4697.4	4700.8
Multivalued (H)	4737.1	4759.4	4752.2	4753.7	4702.2	4682.1	4683.5

Table 3: WAIC values of the monosyllable model using different segmental and tonal distances without information gain weighting. (H): Hamming, (E): Euclidian, (M): Manhattan distances.

After applying information gain weighting to both the tonal and segmental distances, the results did not improve much, and in the case of the natural class-based distance, the WAIC values even increased as in Table 14 in Supplementary Materials 8.3. This is consistent with Nerbonne and Heeringa's results.

For comparison, we also fitted a model that, instead of segmental and tonal distances on conceptual grounds, directly calculates acoustic distance from the audio recordings. We calculated them by obtaining cochleagrams of each of the recordings using Praat with the default parameters, then calculating the Euclidean distances between the cochleagrams. The problem of different numbers of samples was resolved similarly to the method described in Heeringa (2004):[10] if one recording had *n* samples and the other had *m*, we calculated the distance using a number of samples equal to the least common multiplier (LCM) of the two. For example, if one recording has six samples and the other has four, then we use each sample from the first recording twice and each sample from the second recording three times, so there are twelve samples from both recordings. This purely acoustic distance performed far worse than any of the phonological ones, at WAIC value 5070.2.

3.3 Relative weighting of syllable components

To investigate relative weightings of subparts of a syllable, we fitted a version of the model that separates segmental distance into onset, nucleus and coda distance. Again, when fitting this model, instead of choosing the best tonal and segmental representations, we used the natural class distance for segmental distance and onset-contour representation for tonal distance. This makes the model more comparable between monosyllables and disyllables.

As seen in Figure 2, the coefficients of onsets, nuclei and tone were estimated at 1.80 (SE: 0.27; 95% CI: (1.30, 2.36)), 2.12 (SE: 0.29, 95% CI: (1.30, 2.36)), 0.68 (SE: 0.68, 95% CI: (0.20, 1.16), and 0.84 (SE: 0.0.84, 95% CI: (0.20, 1.16)) respectively. The difference between onsets

[10] Conceptually, our acoustic distance is very different from Heeringa's. Heeringa was computing acoustic distances between phones: he averaged the distance over different recordings of the same sound. By contrast, we computed acoustic distances between the recordings used in the stimuli themselves.

and nuclei and between codas and tones are respectively estimated at -0.32 (SE: 0.4, 90% CI: (-1.09, 0.45) and -0.15 (SE: 0.34, 90% CI: (-0.84, 0.49)), so we cannot conclude that they are different. However, we have strong evidence that nuclei are weighted heavier than codas, with an estimated difference of 1.44 (SE: 0.4, 95% CI: (0.67, 2.25)).

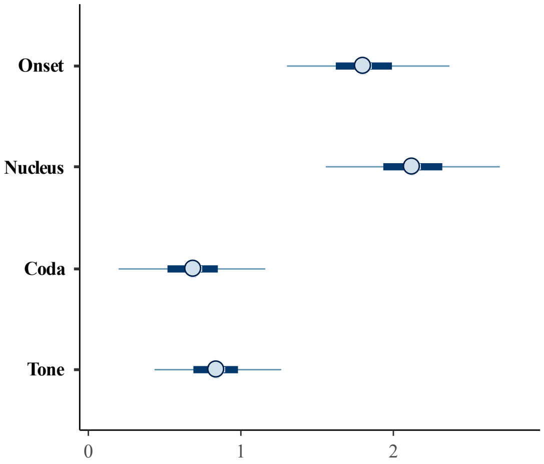

Figure 2: Estimates of the weightings of onset, nucleus, coda and tone along with 95% and 50% credible intervals.

4 Results for disyllables

Scatterplots of the data as in Figure 4 in Supplementary Materials 8.4 show that in the disyllabic case, there seems to be a sharper discrepancy between the fully categorical judges and gradient judges. Thus, we attempted to model the situation by assuming that the tone and segmental distances come from a Gaussian mixture model with different means, but were not able to generate a model without divergent transitions in the MCMC chains. Therefore, we have retained the same full model as in the previous section, without adding additional complexity. Again, the full model was found to have the best WAIC compared to reduced models.

4.1 Tonal and segmental weighting

No strong evidence was found that the population-level coefficients of segmental and tonal distance (μ_γ and μ_δ) are different; the former is estimated at 1.67 (SE: 0.16, 95% CI: (1.37, 2.00)), while the latter is estimated at 1.34 (SE: 0.26, 95% CI: (0.81, 1.85)) respectively. Thus, unlike the case of monosyllables, there is no evidence that segments are, on average, weighted heavier than tones. A 95% credible interval of the difference between the two ($\mu_\gamma - \mu_\delta$), as calculated by the brms package using posterior draws, is (-0.23, 0.91) (point estimate: 0.34; SE: 0.28), indicating that there is no strong evidence to suggest a difference in weighting.

4.2 Comparison of distance and segment measures

To compare different distance measurements, we applied the full model to other measures of tonal and segmental distance, as we have done with monosyllables. The results are in Table 4. This time, of the tonal distances, the autosegmental representation stands out as the worse. Among the rest, the contour-offset representation seems to be the best, but there is not much difference between its WAIC and the model using Hamming distances between Chao tone letters. As for the segmental distances, multivalued features perform the best, especially with Hamming distance, whereas the phonology-based distances perform even worse than the baseline.

	Chao (H)	Chao (M)	Chao (E)	Autoseg-mental	O-C	O-C-O	C-O
Simple	7168.7	7172.0	7185.4	7237.2	7177.2	7176.4	7168.5
Natural class	7185.7	7194.0	7201.0	7247.7	7194.5	7189.9	7179.2
Binary (H)	7191.2	7204.5	7213.0	7249.9	7200.7	7193.2	7188.0
Multivalued (E)	7161.6	7172.6	7181.1	7226.6	7175.1	7164.9	7158.8
Multivalued (M)	7162.0	7175.1	7181.1	7226.6	7177.9	7168.5	7158.5
Multivalued (H)	7163.5	7173.4	7181.3	7227.5	7178.5	7165.7	7153.0

Table 4: WAIC values of the monosyllable model using different segmental and tonal distances without information gain weighting.

The addition of information gain weighting greatly inflated the WAIC of most models, implying that information gain weighting did not improve the models. The details are provided in Supplementary Materials 4. Note that the (onset)-contour-(offset) representation failed to outperform the Chao tone letters in the simulations of disyllables. We think that this is because the (onset)-contour-(offset) representation overlooked the change in pitch level between the two syllables. Therefore, we created several extensions of the two representations. In the first type (O-C-O+ type 1 in Table 5), we used the offset of the first syllable and the onset of second syllable to determine the inter-syllable pitch-level change, then added this to the onset-contour-offset representation. In the second type (O-C-O+ type 2 in Table 5), we took the 'average' pitch of the onset and offset of the two syllables, with very low denoted by '1' and high denoted by '4', then determined whether the average pitch was rising, falling or level. Then we added this to the onset-contour-offset representation. Finally, we determined the pitch level change between the two offsets and added the result to the contour-offset representation (C-O+ type 3 in Table 5).

As shown in Table 5, the type 1 did not result in much improvement while the type 2 resulted in much lower WAICs than the original onset-contour-offset representation. The type 3 resulted in much lower WAICs than the original contour-offset representation: when paired with the natural class distance, the WAIC value was only 7162.8, compared to the original 7179.2. Based on this observation, we conclude that for disyllabic items, the best distance metric to predict distance judgements involved the natural class distance between the segment strings and the Hamming distance between the modified contour-offset representation of the tones reflecting the change in pitch level between the two syllables.

	O-C-O	O-C-O+ (type 1)	O-C-O+ (type 2)	C-O	C-O+ (type 3)
Simple	7176.4	7177.8	7164.6	7168.5	7152.8
Natural class	7189.9	7188.4	7176.7	7179.2	7162.8
Binary (H)	7193.2	7180.2	7189.4	7188.0	7169.7
Multivalued (E)	7164.9	7153.8	7161.3	7158.8	7142.3
Multivalued (M)	7168.5	7153.0	7163.7	7158.5	7143.8
Multivalued (H)	7165.7	7153.0	7163.6	7153.0	7138.6

Table 5: WAIC values of the monosyllable model using different segmental and tonal distances without information gain weighting, using newly developed tonal representations.

Again, the purely acoustic distance measure fared far worse, with a WAIC of 7510.5.

4.3 Relative weighting of syllable components

To explore the relative weights of syllable components among disyllables, we fitted a version of the model that separates segmental distance into onset, nucleus and coda distance. Note that we assumed equal weighting of two syllables within an item; in other words, onset, nucleus, and coda in both syllables were treated equally in our modeling. When fitting this model, we used the natural class distance for segmental distance and onset-contour representation for tonal distance. This makes the model more comparable between monosyllables and disyllables.

The coefficient of onsets, nuclei, and coda was estimated at 2.53 (SE: 0.43; 95% CI: (1.68, 3.36)), 1.38 (SE: 0.41, 95% CI: (0.51, 2.18)), and 0.68 (SE: 0.38, 95% CI: (0.18, 1.70) respectively, and that of tones was estimated at 1.29 (SE: 0.25, 95% CI: (0.78, 1.78)). Based on posterior draws, the difference between onset and nucleus, nucleus and coda, and coda and tone weighting is estimated at 1.15 (SE: 0.68, 95% CI: (-0.2, 2.46)), 0.43 (SE: 0.62, 95% CI: (-0.81, 1.68)), and -0.35 (SE: 0.43, 95% CI: (-1.19, 0.48)) respectively. Clearly, we do not have strong evidence that the nucleus, coda and tone differ in weighting, though we do have weak evidence that onsets are weighted heavier than nuclei, since a 90% credible interval is (0.02, 2.26).

The relative weightings of onset, nucleus, coda and tone are provided below:

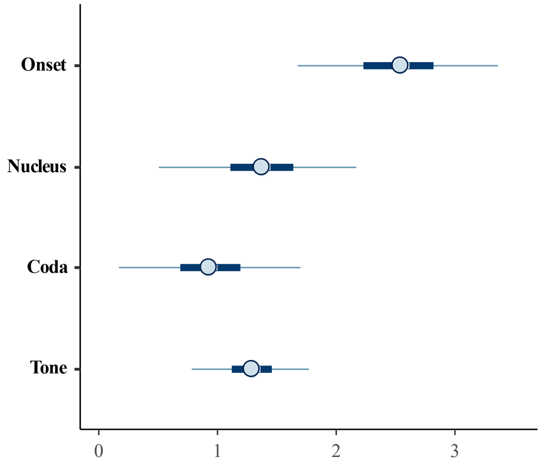

Figure 3: Estimates of the weightings of onset, nucleus, coda and tone along with 95% and 50% credible intervals.

5 Discussion and conclusion

5.1 Tonal and segmental weighting

Our study provides evidence that segments are weighted heavier than tones in Cantonese monosyllables in measuring phonological distances. It may appear to contrast with Yang and Castro's (2008) findings that segments and tones were equally important in Zhuang whereas segments were *less* important than tone in Bai. This difference can be interpreted in several ways: there may exist typological difference in the relative weighting of tones and segments, the difference could be due to differences in the task performed (direct distance judgements vs. mutual intelligibility), or the coefficients in Yang and Castro's model (which they do not report) may not directly support their conclusion. [11] Note though that the results of disyllables did not

[11] Another potentially related paper is Cham (2003), which looks at the perception of Thai tones and segments by Cantonese-speaking children and adults by administering phonological awareness tests where participants were instructed to select the odd one out among three syllables. Cham claims that 'children and

support those of monosyllables in our study; segments were not weighted heavier than tones in disyllables. We want to point out that we do not have good evidence to the contrary either, since their 50% credible intervals do not overlap. A less clear pattern among disyllables can be attributed to the fact that the disyllabic test items are less representative of the lexicon than monosyllables. Recall that our test included the same number of monosyllables (*n*=72) and disyllables (*n*=72) to avoid any bias toward one or the other structure. Due to this setting, fewer number of logically possible combinations of disyllables were tested, which in turn could have resulted in wider variabilities in judgments.

An overarching assumption of our study is that tone is considered separate from segments, hence tonal and segmental distances are computed independently as inputs to the final distance. It is possible to assume that the tone is tied to the nucleus instead. However, even if we consider nucleus-tone combinations, the effect of nucleus and tone would still be additive, as far as we determine the distance between nucleus-tone combinations using Levenshtein distance as usual. Therefore, the result would be similar to the current model except the nucleus is forced to be weighted the same as each element of the tone. For example, the distance between aaHL and aMF would still be the distance between [aa] and [a], between H and M and between L and F summed up. For this reason, in order to build a model where tone is tied to segments and which is significantly different from our current assumption, a different class of distances than Levenshtein needs to be identified.

5.2 METRIC COMPARISONS

For segmental distances, we have demonstrated that multivalued features are a better representation of phonemes for predicting distance judgements than binary distinctive features, whether we use the natural class distance or the Hamming distance with binary features. At the same time, it was found that a purely acoustic/auditory measure of distance works far worse than any of the other features mentioned. This result can be interpreted in two ways. First, it is possible to consider that articulation is most relevant to distance judgements. This is because most of the multivalued features are articulation-based; the binary distinctive features were designed with reference to articulation, but abstracted away from it; and the cochleagram had no articulatory component at all. This interpretation aligns with conclusions drawn by previous studies like Somers (1998) and Heeringa (2004), as well as results in phonetics and phonology more generally, e.g. the view that speech perception involves processes also used in production (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967). Second, it is also possible to propose that a balance between phonetics and phonology, which is what the multivalued features provide, may be the best. Unlike the binary features, the multivalued features distinguish between allophones and allows for gradient features, but at the same time does not take into account minor, non-systematic phonetic detail as the cochleagram does.

We have also shown that Hamming distance between multivalued features without information gain weighting works better than any other type of distance metric between multivalued features. This is partially consistent with Nerbonne & Herringa's (1997) results,

adults generally performed better in phone awareness tasks than tone awareness tasks'. If this were true, it would imply that tones are perhaps perceptually less salient than phones for Cantonese speakers, and thus the higher weighting of phones for monosyllables would not be surprising. (Cham did not look at disyllables.) However, in fact, her data is more ambiguous for adults. She only mentions that the performance in the Phone 1 task significantly better than all the Tone tasks. However, her graph seems to show that performance in the Tone 2 task was better than the corresponding Phone 2 task for adults, while performance in the Tone 3 task was worse than the corresponding Phone 3 task. Therefore, her results seem ambiguous as to the relative salience of tones and phones, and hence does not lend additional support to the idea that tones are perceptually less salient than phones.

which show distances between multivalued features without information gain weighting works best for determining dialect distance among different metrics using multivalued features, though the best distance measures in their study were Manhattan distances. This difference can be attributed to the fact that the actual magnitude of the differences between numerical values should be more crucial to Nerbonne and Heeringa's case, which Manhattan distances take into account. We want to note that the lack of effectiveness of information gain weighting does not necessarily imply that the features are equally weighted, since information gain is only one possible type of weighting scheme and there may well be other theoretical or empirical schemes that can improve the predictive power of phonological distances. We leave this for furute research.

For tonal distances, we showed that representations with a contour component worked best for both monosyllables and disyllables. In particular, the contour-offset representation, augmented with a component representing the change in pitch between the two syllables, works best on disyllables by far. This implies that tone contours are very important for distance judgements in Cantonese, consistent with the results by Yang and Castro (2008), who find that the best distance metric are those derived from one of the onset-contour, onset-contour-offset and contour-offset representations of tone, and also those of Tang and van Heuven (2011), whose estimates suggest that onset-contour outperforms Chao tone letters. All in all, it is possible that tone contours are important for determining phonological distance in languages with contour tones contrasting with level tones.

5.3 RELATIVE WEIGHTING OF SYLLABLE COMPONENTS

We further split segments into onsets, nuclei, codas and tones to investigate relative weightings of syllable components in phonological distance judgments. For monosyllables, we have shown that onsets and nuclei are weighted heavier than codas and tones. For disyllables, onsets are weighted heavier than nuclei, codas and tones. We do not have strong evidence of the relative weighting between elements within the two groups in either case.

Since phonological distances are relevant to distinguishing between lexical items, we may expect that the less predictable an element is, whether it is an onset, a nucleus, a coda, or a tone, the more important it is for distinguishing between words. In turn, the unpredictable elements may play a more significant role in evaluating phonological distances. We may estimate an element's lack of predictability using entropy. Because such simple point estimates of entropy do not give us information about variability in the estimates, we computed confidence intervals for the differences between the entropies to ensure that the differences are not simply due to sampling error. Since no standard formula is available for confidence intervals of differences between marginal entropy measures, we derived our own using the asymptotic properties of the probability estimates along with the delta method; details are given in Supplementary Materials 8.5. The 95% confidence intervals for the differences are given in Table 6; we applied a Bonferroni correction with $g = 6$, so the monosyllable and disyllable estimates each have at least 95% confidence as a whole. All of the confidence intervals are quite far from including 0. Thus, we have very strong evidence that the entropies of the four elements are ranked onset > nucleus > coda > tone for both monosyllables and disyllables.

Type	Difference	Point estimate	Confidence interval
Mono-syllable	Onset – Nucleus	1.0224141	(1.0145622, 1.030266)
	Onset – Coda	1.2220209	(1.2132533, 1.2307884)
	Onset – Tone	1.3616846	(1.3560588, 1.3673103)
	Nucleus – Coda	0.1996068	(0.1902098, 0.2090038)
	Nucleus – Tone	0.3392704	(0.3332731, 0.3452678)
	Coda – Tone	0.1396637	(0.1323494, 0.146978)
Di-	Onset – Nucleus	1.4175104	(1.39113759, 1.4438831)

syllable	Onset – Coda	1.5250996	(1.49897581, 1.5512234)
	Onset – Tone	2.0967519	(2.0722213, 2.1212826)
	Nucleus – Coda	0.1075892	(0.08266781, 0.1325107)
	Nucleus- Tone	0.6792416	(0.65845484, 0.7000283)
	Coda – Tone	0.5716523	(0.54861183, 0.5946928)

Table 6: Point and interval estimates of the differences between the entropies of various syllable components.

Another information-theoretic measure of importance is functional load. The functional load of a component c is computed by comparing the entropy $H(L)$ of the entire language to the entropy of a fictional language state $H(L'_c\ )$ where the all contrasts in that component are neutralised (Hockett, 1966; Carter, 1987; Surendran & Levow, 2004; Oh, Coupé. Marsico and Pellegrino, 2015):

(6) $FL_c(L) = \frac{H(L)-H(L'_c)}{H(L)}$

We computed functional loads for onsets, nuclei, codas and tones and calculated confidence intervals for the differences between the functional loads, as in Table 7 (note that all but the difference between nucleus and coda in disyllables do not cover zero). As shown, in terms of functional load, tones are in fact slightly more important than nuclei and tones which are relatively close to each other. This is in contrast with simple entropy calculations.

Type	Difference	Point estimate	Confidence interval
Mono-syllable	Onset – Nucleus	0.07282048	(0.07175836, 0.07388259)
	Onset - Coda	0.11316083	(0.11210798, 0.11421367)
	Onset – Tone	0.06108325	(0.06000738, 0.06215911)
	Nucleus – Coda	0.04034035	(0.0394496, 0.0412311)
	Nucleus- Tone	-0.01173723	(-0.01268993, -0.01078453)
	Coda - Tone	-0.05207758	(-0.05305742, -0.05109774)
Di-syllable	Onset – Nucleus	0.010521195	(0.00847304, 0.01256935)
	Onset - Coda	0.012062699	(0.01004672, 0.01407868)
	Onset – Tone	0.008590597	(0.00655996, 0.01062123)
	Nucleus – Coda	0.001541504	(-0.00010232, 0.00318533)
	Nucleus- Tone	-0.001930599	(-0.00370368, -0.00015752)
	Coda - Tone	-0.003472103	(-0.00525529, -0.00168892)

Table 7: Point and interval estimates of the differences between the functional loads of various syllable components.

From the examinations of entropies, we would expect the weight hierarchy of onsets > nuclei> codas > tones . From the examination of functional loads, we would expect the weight hierarchy of onsets > tones > nucle, codas. Considering our modeling results, it suggest that information-theoretic predictability and functional load do have a partial power to account for the weightings of syllabic components in phonological distance measures but cannot predict its full range. One possible intuitive explanation for the erratic behaviour of nuclei in the monosyllable case is that nuclei are the ‘central’ part of the word when it is monosyllabic, but its role is weakened in a disyllabic word due to an additional transitional property incurred between syllables. Further investigations are needed to determine the exact reasons behind.

Given our study is on a single tonal language, we cannot come to a general conclusion about the relative roles of segments and tones in phonological distance measures overall. We cannot guarantee that the distance metrics that work best for Cantonese are applicable to other tonal languages either. Also, it is not possible to draw a generalization about the role of syllabic components in measuring phonological distances that are universally pertinent to other tonal languages. Moreover, our study cannot ensure that the current findings are applicable when other suprasegmental features such as stress or intonation are considered in measuring phonological distances. Despite its limited scope to a single tonal language situation, we do believe that our study provides methodological insights to the work on phonological distance measures especially when suprasegmental features are concerned. We also hope that this study can open doors to wider explorations of phonotactic models incorporating suprasegmental features, which have been relatively missed in the current literature.

6 REFERENCES

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1), 9-41.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods?. *Journal of Memory and Language*, *44*(4), 568-591.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology* (Vol. 102). Berlin: Walter de

Gruyter.

Barth, D., & Kapatsinski, V. (2018). Evaluating logistic mixed-effects models of corpus data. In D. Speelman, K. Heylen & D. Geeraerts (Eds.), *Mixed Effects Regression Models in Linguistics*, 99-116. Cham: Springer International Publishing.

Beijering, K., Gooskens, C., & Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. *Linguistics in the Netherlands*, 15, 13-24.

Broe, M. (1996). A generalized information-theoretic measure for systems of phonological classification and recognition. In *Computational Phonology in Speech Technology: Proceedings of the Second Meeting of the ACL Special Interest Group in Computational Phonology*, pp 17-24. Santa Cruz. Association for Computational Linguistics.

Bürkner P. C. (2017). brms: An R Package for Bayesian Multilevel Models using Stan. *Journal of Statistical Software*. 80(1), 1-28. doi:10.18637/jss.v080.i01

Bürkner P. C. (in press). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M Brubaker, M.;Guo, J.,Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Carter, D. M. (1987). An information-theoretic analysis of phonetic dictionary access. *Computer Speech & Language*, 2(1), 1–11.

Cham, H. Y. (2003). A cross-linguistic study of the development of the perception of lexical tones and phones. (Unpublished BSc thesis.) University of Hong Kong, Hong Kong.

Cheng, C. C. (1997). Measuring relationship among dialects: DOC and related resources. International Journal of Computational Linguistics & Chinese Language Processing, Volume 2, Number 1, February 1997*: Special Issue on Computational Resources for Research in Chinese Linguistics*, *2*(1), 41-72.

Clumeck, H., Barton, D., Macken, M. A., & Huntington, D. A. (1981). The aspiration contrast in Cantonese word-initial stops: data from children and adults [Guangdonghua Saiyin Shengmu Songqi Duili: Ertong ji Chengren de Ziliao]. *Journal of Chinese Linguistics*, 9(2), 210-225.

Coleman, J., and Janet P. (1997). Stochastic phonological grammars and acceptability. In *Third Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop*, ed. by John Coleman, 49-56. East Stroudsburg, PA: Association for Computational Linguistics.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory.* Hoboken, N.J.: Wiley-Interscience.

Dankovicova, J., West, P., Coleman, J., & Slater, A. (1998). *Phonotactic grammaticality is gradient*. Poster presented at the 6th International Conference on Laboratory Phonology, University of York, 2–4 July 1998

Ellison, T. M., & Kirby, S. (2006). Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 273-280).

Frisch, S., Broe, M., & Pierrehumbert, J. (1997). Similarity and phonotactics in Arabic. Rutgers Optimality Archive [Online], ROA-223-1097. Available at http://www.webslingerz.com/cgi-bin/oa_list.cgi.

Frisch, S., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on processing non-words. *Journal of Memory and Language*, 42(4), 481–496.

Fisher, W. M., & Fiscus, J. G. (1993, April). Better alignment procedures for speech recognition evaluation. In *icassp* (pp. 59-62). IEEE.

Gathercole, S. E., & Martin, A. J. (1996). Interactive processes in phonological memory. In M.A. Conway (Ed.), *Cognitive models of memory*. Hove, UK: Psychology Press/MIT Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*. New York, NY: Chapman and Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models* (Vol. 1). New York, NY: Cambridge University Press.

Gildea, D., & Jurafsky, D. (1996). Learning bias and phonological-rule induction. *Computational Linguistics*, *22*(4), 497-530.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, 20(2), 157–177

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3), 379–440.

Hayes, B. (2011). *Introductory phonology* (Vol. 32). Oxford : John Wiley & Sons.

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance* (Unpublished doctoral dissertation.) University Library Groningen, Groningen, Netherlands.

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In *Proceedings of the workshop on linguistic distances* (pp. 51-62). Association for Computational Linguistics.

Hockett, C. F. (1966). *The quantification of functional load: A linguistic problem. Report Number RM-5168-PR*. Santa Monica: Rand Corp.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*. London: Pearson.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics* (pp. 60-66). Morgan Kaufmann Publishers Inc.

Kessler, B. (2005). Phonetic comparison algorithms 1. *Transactions of the Philological Society*, *103*(2), 243-260.

Kondrak, G. (2002). *Algorithms for language reconstruction*. (Unpublished doctoral dissertation.) University of Toronto, Toronto, Canada.

Kondrak, G. (2002, August). Determining recurrent sound correspondences by inducing translation models. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, 37(3), 273-291.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-361.

Mok, P. P., Zuo, D., & Wong, P. W. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language variation and change*, 25(3), 341-370.

Nerbonne, J., & Heeringa, W. (1997). Measuring dialect distance phonetically. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, 10(11), 591-613.

Oakes, M. P. (2000). Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, *7*(3), 233-243.

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics*, 53, 153-176.

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. In *Proceedings of NELS* (Vol. 23, pp. 367-381).

Qualtrics. (2010–2011). Qualtrics [Computer software]. Provo, UT: Author.

Rao, C. R. (1973). *Linear statistical inference and its applications* (Vol. 2). Wiley New York.

Saiegh-Haddad, E. (2004). The impact of phonemic and lexical distance on the phonological analysis of words and pseudowords in a diglossic context. *Applied Psycholinguistics*, 25(4), 495-512.

Somers, H. L. (1998). Similarity metrics for aligning children's articulation data. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2* (pp. 1227-1232). Association for Computational Linguistics.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, 12(3), 175–200. doi:10.20982/tqmp.12.3.p175

Surendran, D., & Levow, G. A. (2004). The functional load of tone in Mandarin is as high as that of vowels. In *Speech Prosody 2004, International Conference*.

Tang, C., & van Heuven, V. J. J. P. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, 119, 24.

Tang, C. (2009). *Mutual intelligibility of Chinese dialects: an experimental approach*. (Unpublished doctoral dissertation.) LOT, Utrecht.

Tang, C., & van Heuven, V. J. J. P. (2011). Tone as a predictor of mutual intelligibility between Chinese dialects. *Online Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII 2011)*. International Phonetic Association.

Tang, C., & Van Heuven, V. J. (2015). Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics*, 53(2), 285-312.

Tang, S.-W., Kwok, F., Lee, T. H.-T., Lun, C., Luke, K. K., Tung, P., & Cheung, K. H. (2002). *Guide to LSHK Cantonese romanization of Chinese characters*. Hong Kong: Linguistic Society of Hong Kong.

Tse, H. (2005). *The Phonetics of VOT and Tone Interaction in Cantonese.* (Unpublished doctoral dissertation.) University of Chicago, Chicago, IL.

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2), 307-314.

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., & Baayen, R. H. (2014). A cognitively grounded measure of pronunciation distance. *PloS ONE*, 9(1), e75734.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, 40, 47–62.

Vitevitch, M. S., & Luce, P.A. (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science*, 9, 325–329.

Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech communication*, *33*(4), 319-337.

Yang, C., & Castro, A. (2008). Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing*, *2*(1-2), 205-219.

Yip, M. J. (1980). *The tonal phonology of Chinese* (Unpublished doctoral dissertation.) Massachusetts Institute of Technology, MA.

Yu, A. C., & Zellou, G. (2018). Individual Differences in Language Processing: Phonology. *Annual Review of Linguistics*, *4*(1).

Zee, E. (1999). Chinese (Hong Kong Cantonese). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

7 SUPPLEMENTARY MATERIALS

7.1 SUPPLEMENTARY MATERIALS TO SECTION 2

Segmental representations

The distinctive binary values for Cantonese phonemes are presented in Table 1. The exact number of phonemes may be debatable, and we assume that each symbol in Jyutping, a standard phonological transcription system for Cantonese (Tang, Kwok, Lee, Lun, Luke, Tung and Cheung, 2002), is a phoneme.

	cons	son	syll	lab	cor	dor	round	nas	lat	tens	voic	stri	cont	high	spr_gl	low	front
b	+	-	-	+	-	-	-	-	-	0	-	0	-	0	-	0	0
p	+	-	-	+	-	-	-	-	-	0	-	0	-	0	+	0	0
m	+	+	0	+	-	-	-	+	-	0	+	0	-	0	0	0	0
f	+	-	-	+	-	-	-	-	-	0	-	0	+	0	0	0	0
d	+	-	-	-	+	-	-	-	-	0	-	-	-	0	-	0	0
t	+	-	-	-	+	-	-	-	-	0	-	-	-	0	+	0	0
n	+	+	-	-	+	-	-	+	-	0	+	-	-	0	0	0	0
l	+	+	-	-	+	-	-	-	+	0	+	-	+	0	0	0	0
z	+	-	-	-	+	-	-	-	-	0	-	+	-	0	-	0	0
c	+	-	-	-	+	-	-	-	-	0	-	+	-	0	+	0	0
s	+	-	-	-	+	-	-	-	-	0	-	+	+	0	0	0	0
g	+	-	-	-	-	+	-	-	-	0	-	0	-	0	-	0	0
k	+	-	-	-	-	+	-	-	-	0	-	0	-	0	+	0	0
ng	+	+	0	-	-	+	-	+	-	0	+	0	-	0	0	0	0
h	+	-	-	-	-	-	-	-	-	0	+	0	-	0	+	0	0
gw	+	-	-	+	-	+	+	-	-	0	-	0	-	0	-	0	0
kw	+	-	-	+	-	+	+	-	-	0	-	0	-	0	+	0	0
w	-	+	-	+	-	+	+	-	-	+	+	0	+	+	0	-	-
j	-	+	-	0	-	-	0	-	-	+	+	0	+	+	0	-	+
aa	-	+	+	-	-	-	-	-	-	+	+	0	+	-	0	+	-
a	-	+	+	-	-	-	-	-	-	-	+	0	+	-	0	+	-
e	-	+	+	-	-	-	-	-	-	0	+	0	+	-	0	-	+
oe	-	+	+	+	-	-	+	-	-	+	+	0	+	-	0	-	+
eo	-	+	+	+	-	-	+	-	-	-	+	0	+	-	0	-	+
o	-	+	+	+	-	-	+	-	-	0	+	0	+	-	0	-	-
i	-	+	+	-	-	-	-	-	-	0	+	0	+	+	0	-	+
u	-	+	+	+	-	+	+	-	-	0	+	0	+	+	0	-	-
yu	-	+	+	+	-	-	+	-	-	+	+	0	+	+	0	-	+

Table 1: Distinctive binary values for Cantonese phonemes

The multivalued phonological features for Cantonese segments are presented in Table 2. The exact values themselves involved educated guesswork, as with the original Ladefoged table. Since phonetic features are involved, we had to determine the values of certain allophones separately. For Cantonese sounds which have close English parallels, such as [s] and [l], we largely used Ladefoged's values. The values for phonemes without English equivalents were determined based on Ladefoged's definitions of the features, extrapolation from other sounds, and previous work on Cantonese phonetics (Clumeck, Barton, Macken, & Huntington, 1981; Tse, 2005; Bauer & Benedict, 1997; Zee, 1999). For example, for voicing, we follow Ladegofed in assigning 80 to all voiced consonants and vowels, setting all others to 0. Different from Ladefoged, we assigned zero to [h], given no voicing involved in the phoneme. Moreover, we added a small positive value to the unaspirated consonants, which are not present in English except in limited contexts.

seg	glot	voi	asp	place	lab	stop	nas	lat	son	sib	height	back	round	wide
b	50	20	8	100	100	100	0	0	2.5	0	100	49	0	50
p	60	0	63	100	100	100	0	0	0	0	100	49	0	50
m	50	80	0	100	100	100	100	0	75	0	95	49	0	50
f	50	0	51	95	90	90	0	0	5	10	100	49	0	50
d	50	20	9	85	5	100	0	0	2.5	0	100	49	0	50
t	60	0	70	85	5	100	0	0	0	20	100	49	0	50
n	50	80	0	85	5	100	100	0	75	0	95	49	0	50
l	50	80	0	85	5	70	0	100	80	0	90	49	0	50
z	50	20	58	85	5	95	0	0	8.75	50	100	49	0	50
c	50	0	100	85	5	95	0	0	7.5	60	100	49	0	50
s	50	0	51	85	40	90	0	0	15	100	100	49	0	50
g	50	20	22	60	5	100	0	0	2.5	0	100	49	0	30
k	60	0	77	60	5	100	0	0	0	0	100	49	0	30
ng	50	0	0	60	5	100	100	0	75	0	100	49	0	30
h	50	0	44	60	5	0	5	0	5	10	100	49	90	50
gw	50	50	46	60	80	90	0	0	36.25	0	100	49	90	40
kw	50	50	80	60	80	90	0	0	35	0	100	49	90	40
w	50	80	0	60	80	80	0	0	70	0	90	49	0	40
j	50	80	0	70	5	80	0	0	70	0	90	49	0	95
aa	50	80	0	44	5	0	0	0	95	0	15	40	0	20
a	50	80	0	29	5	0	0	0	95	0	25	65	0	30
e	50	80	0	59	5	5	0	0	95	0	50	15	0	30
oe	50	80	0	50	30	5	0	0	95	0	53	30	40	30
eo	50	80	0	32	30	5	0	0	95	0	53	60	60	30
o	50	80	0	9	60	0	0	0	95	0	50	97	50	20
i	50	80	0	62	25	75	0	0	80	5	85	10	0	70
u	50	80	0	11	80	60	0	0	85	0	85	95	90	40
yu	50	80	0	59	80	70	0	0	80	5	83	15	0	60
[ɪ]	50	80	0	59	20	5	0	0	95	0	60	15	0	50
[ʊ]	50	80	0	14	55	20	0	0	90	0	60	90	0	30
[e]	50	80	0	62	5	5	0	0	95	0	55	10	0	50
[o]	50	80	0	11	50	5	0	0	95	0	58	95	70	40

[p̚]	60	0	0	100	100	100	0	0	0	0	100	49	0	50
[t̚]	60	0	0	85	5	100	0	0	0	20	100	49	0	50
[k̚]	60	0	0	60	5	100	0	0	0	0	100	49	0	30

Table 2: Multivalued values for Cantonese phones

Information gain weighting

Instead of directly taking Nerbonne and Heeringa's notation, we use modified notation that better resembles standard information-theoretic notation. Let a sound $\mathbf{S}$ be a random vector of features with components $\mathbf{f_1}, \mathbf{f_2}, \dots, \mathbf{f_I}$, where each component represents a feature. Each possible value of $\mathbf{S}$, denoted s_i, is thus a phoneme. Suppose there are J phonemes in the language. Then the entropy of $\mathbf{S}$ is as follows:

(1) $H(\mathbf{S}) = -\sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j)\right)$

where $P(\mathbf{S} = s_j)$ is the probability of the j-th possible value of $\mathbf{S}$, and is estimated by the frequency of the phoneme corresponding to that feature vector value in the corpus divided by the total number of segments in the corpus. The conditional entropy of $\mathbf{S}$ on each feature is calculated thus:

(2) $H(\mathbf{S}|\mathbf{f_i}) = \sum_{v \in V} H(\mathbf{S}\,|\mathbf{f_i} = v) P(\mathbf{f_i} = v)$

where V is the set of possible values of $\mathbf{f_i}$ and the value $H(\mathbf{S}|\mathbf{f_i} = v)$ is defined as follows:

(3) $H(\mathbf{S}|\mathbf{f_i} = v) = -\sum_{j=1,\, \mathbf{f_i}=v \text{ when } \mathbf{S}=s_j}^{J} P(\mathbf{S} = s_j|\mathbf{f_i} = v) \log\left(P(\mathbf{S} = s_j|\mathbf{f_i} = v)\right)$

The information gain is then obtained as follows:

(4) $IG(\mathbf{f_i}) = H(\mathbf{S}) - H(\mathbf{S}|\mathbf{f_i})$

Thus, according to Nerbonne & Heeringa (1997), the information gain from a feature is calculated by taking the difference between the entropy of a segment and the conditional entropy of the segment on the feature. Put in a more intuitive way, it calculates the difference between the amount of uncertainty in the identity of segment and the average amount of uncertainty left after we know the value of a feature.

For each of the distance measures we mentioned above, we created an altered version with information gain weighting. Each natural class was considered a 'feature' in the natural class distance measure. We used the Hong Kong Cantonese Corpus (Luke & Wong, 2015) to do the entropy estimations.

A disadvantage of the above formula is that features with null values are considered to have a special value, rather than *lacking* a value for that feature. Broe (1996) extends this notion of information gain to *incomplete* random variables, where certain components (i.e. features) may have null values. Broe uses a different formula for information gain, known as interdependence or mutual information. The formula is shown below, simplified according to our present setting:[12]

(5) $IG(\mathbf{f_i}) = \sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(\frac{1}{P(f_i = v_j)}\right)$,

where v_j is the value of $\mathbf{f_i}$ corresponding to s_j. The two measures of information gain (8, 9) are equivalent. as shown by the argument in Cover and Thomas (2006, pp. 20-21). A version of the proof, adopted to our current setting, is as follows:

(6) $IG(\mathbf{f_i}) = -\sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j)\right) - \sum_{v \in V} H(\mathbf{S}\,|\mathbf{f_i} = v) P(\mathbf{f_i} = v)$

[12] In the actual formula for mutual information, the fraction inside the logarithm should be $\frac{P(\mathbf{S}=s_j \cap f_i = v_i)}{P(\mathbf{S}=s_j)P(f_i=v_i)}$, which can be simplified to the current form since $P(\mathbf{S} = s_j \cap f_i = v_i) = P(\mathbf{S} = s_j)$.

$$
\begin{aligned}
&= -\sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j)\right) \\
&\quad + \sum_{v \in V} \sum_{j=1,\, \mathbf{f_i}=v \text{ when } \mathbf{S}=s_j}^{J} P(\mathbf{S} = s_j | \mathbf{f_i} = v) \log\left(P(\mathbf{S} = s_j | \mathbf{f_i} = v)\right) P(\mathbf{f_i} = v) \\
&= -\sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j)\right) \\
&\quad + \sum_{v \in V} \sum_{j=1,\, \mathbf{f_i}=v \text{ when } \mathbf{S}=s_j}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j | \mathbf{f_i} = v)\right) \\
&= \sum_{v \in V} \sum_{j=1,\, \mathbf{f_i}=v \text{ when } \mathbf{S}=s_j}^{J} P(\mathbf{S} = s_j) \left[\log\left(\frac{P(\mathbf{S} = s_j \ \cap f_i = v_i\ )}{P(f_i = v_i)}\right) - \log\left(P(\mathbf{S} = s_j)\right)\right] \\
&= \sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(\frac{1}{P(f_i = v_j)}\right)
\end{aligned}
$$

The second last line is equal to the last line because the outer sum in in the second last line sums up all possible values v of $\mathbf{f_i}$ while the inner sum sums up all possible values of the expression for phonemes with the value v, meaning that we are summing up the values of expression for all phonemes.

Therefore, we can expand Nerbonne & Heeringa's information gain weighting to account for the fact that certain features may have null values. Let $V' = V \setminus \{0\}$, i.e. the set of values excluding the null value. Then, from Broe's formula (22), we may derive the following formula for computing the information gain of features (taking null values into consideration):

$$
(7)\ \frac{\sum_{j=1, v_j \in V'}^{J} P(\mathbf{S}=s_j) \log\left(\frac{1}{P(f_i=v_j)}\right)}{\sum_{j=1, v_j \in V'}^{J} P(\mathbf{S}=s_j)}
$$

Thus only the phonemes for which the feature is non-null are considered, and the result is normalized by dividing it by the probability of that feature being defined.

When using Broe's formula, we modified the Hamming distance slightly. To account for the fact that null values are ignored rather than being considered as a possible value, we set the distance between +/- and 0 at 0.5 instead of 1, by analogy with the Levenshtein weights, which have 0.5 for insertion and deletion. For example, if phoneme A has a feature vector (+, +, 0) and phoneme B has a feature vector (+, -, -), the distance is 0 + 1 + 0.5 = 1.5 instead of the usual Hamming distance of 0 + 1 + 1 = 2.

Tonal distances

	1	2	3	4	5	6
1	0	0.285714	0.571429	1	0.714286	0.857143
2	0.285714	0	0.285714	0.714286	0.428571	0.571429
3	0.571429	0.285714	0	0.428571	0.142857	0.285714
4	1	0.714286	0.428571	0	0.285714	0.142857
5	0.714286	0.428571	0.142857	0.285714	0	0.142857
6	0.857143	0.571429	0.285714	0.142857	0.142857	0

Table 3: Hamming distances between Chao tone letter representations

	1	2	3	4	5	6
1	0	0.125	0.25	1	0.5	0.75
2	0.125	0	0.125	0.875	0.375	0.625
3	0.25	0.125	0	0.75	0.25	0.5
4	1	0.875	0.75	0	0.5	0.25
5	0.5	0.375	0.25	0.5	0	0.25
6	0.75	0.625	0.5	0.25	0.25	0

Table 4: Manhattan distances between Chao tone letter representations

	1	2	3	4	5	6
1	0	0.171185	0.342371	1	0.553968	0.765564
2	0.171185	0	0.171185	0.959589	0.513556	0.725153
3	0.342371	0.171185	0	0.826556	0.342371	0.684742
4	1	0.959589	0.826556	0	0.484185	0.342371
5	0.553968	0.513556	0.342371	0.484185	0	0.342371
6	0.765564	0.725153	0.684742	0.342371	0.342371	0

Table 5: Euclidean distances between Chao tone letter representations

	1	2	3	4	5	6
1	0	0.333333	0.666667	1	0.666667	0.333333
2	0.333333	0	0.333333	0.666667	0.333333	0.666667
3	0.666667	0.333333	0	0.333333	0.666667	1
4	1	0.666667	0.333333	0	0.333333	0.666667
5	0.666667	0.333333	0.666667	0.333333	0	0.333333
6	0.333333	0.666667	1	0.666667	0.333333	0

Table 6: Hamming distances between autosegmental representations:

	1	2	3	4	5	6
1	0	1	0.5	1	1	0.5
2	1	0	0.5	1	0.5	1
3	0.5	0.5	0	1	1	0.5
4	1	1	1	0	0.5	0.5
5	1	0.5	1	0.5	0	0.5
6	0.5	1	0.5	0.5	0.5	0

Table 7: Hamming distances between onset-contour representations:

	1	2	3	4	5	6
1	0	0.666667	0.666667	1	1	0.666667
2	0.666667	0	0.666667	1	0.666667	1
3	0.666667	0.666667	0	1	0.666667	0.666667

4	1	1	1	0	0.666667	0.666667
5	1	0.666667	0.666667	0.666667	0	0.666667
6	0.666667	1	0.666667	0.666667	0.666667	0

Table 8: Hamming distances between onset-contour-offset representations:

	1	2	3	4	5	6
1	0	0.5	0.5	1	1	0.5
2	0.5	0	1	1	0.5	1
3	0.5	1	0	1	0.5	0.5
4	1	1	1	0	1	1
5	1	0.5	0.5	1	0	1
6	0.5	1	0.5	1	1	0

Table 9: Hamming distances between contour-offset representations:

7.2 SUPPLEMENTARY MATERIALS TO SECTION 3

Word 1	Word 2	Word 1	Word 2	Word 1	Word 2
bing1	bing1	nyun5	nyun6	wing5	wing5
bei2	be1	liu2	leu2	wu6	wyu6
bok6	zyun6	leot6	zing6	wan5	nau5
ban6	poe6	lei4	lyu4	waa4	maa4
pei5	pei4	go1	go3	zoek3	zoek6
paa4	pe5	gong3	zong1	zam6	zaam3
pik1	mun6	ge3	fou4	zap1	jit3
paa2	boi3	gun2	hung5	zyu2	ju3
maa5	maa1	ku1	ku2	coek3	coek4
mong4	mung2	king4	ging3	cam4	sam6
miu5	ding3	kiu5	he1	cyu5	pan1
mat6	mo2	kap6	goeng2	caa1	so5
fu6	fu4	hap6	hap6	sam1	sam1
fo2	ho2	him2	heng2	sap1	sat1
fan3	ngaak3	hing3	jo3	syut3	zam3
fu4	pek4	hek4	si4	soeng4	cung4
dim2	dim3	gwaat3	gwaat2	joeng5	joeng2
dyut6	dyu5	gwat6	gat1	jyun4	joen4
dik1	po3	gwaa1	jok3	jap1	lok6
doek3	suk2	gwing2	ting2	jing2	seng5
ting5	ting1	kwok3	kwok5	aat3	aat3
taam4	taang2	kwang2	kwat4	ngan4	ngang1
tou2	mat6	kwai5	sing3	ngaan5	gen3

tiu3	seu4	kwik1	ge4	ngan6	koet2

Table 11: The table of stimuli for monosyllables

Word 1	Word 2	Word 1	Word 2	Word 1	Word 2
cik1zaak3	sik1zaak3	jyun5suk6	jyun5cu6	faa1ping4	faa1pe4
cik1zaak3	sek3faak3	jyun5suk6	jyun2soek6	faa1ping4	haa2ping4
cik1zaak3	bik2caak6	jyun5suk6	jun2zuk2	faa1ping4	waa5pi5
cik1zaak3	jau1sau3	jyun5suk6	lung5zoe6	faa1ping4	hot1zu4
cik1zaak3	gan1loeng2	jyun5suk6	him3joe6	faa1ping4	hyut1coe1
cik1zaak3	fan5nou6	jyun5suk6	su1ze4	faa1ping4	mui4gwai3
gau2joeng5	gau2joeng5	waa6mui4	waa6mui4	tau4deng2	tau4deng2
gau2joeng5	gau5joeng5	waa6mui4	waa3mui4	tau4deng2	tau4deng4
gau2joeng5	gau4joeng6	waa6mui4	waa4mui3	tau4deng2	tau3deng4
gau2joeng5	gau2lau5	waa6mui4	fe6bu4	tau4deng2	gaai4ti2
gau2joeng5	gaau3liu4	waa6mui4	he6mei5	tau4deng2	doi6te3
gau2joeng5	kou3ja3	waa6mui4	haa4mei3	tau4deng2	tu3ti3
sin3ngaan5	sing3ngaa5	kwan3bik1	kan3bi1	koeng5hang4	kong5sing4
sin3ngaan5	cin5ngaan5	kwan3bik1	gwan2bik1	koeng5hang4	koeng4hong4
sin3ngaan5	cin4gan1	kwan3bik1	gwan1baak5	koeng5hang4	koe3haa6
sin3ngaan5	sou3joeng5	kwan3bik1	gu3go1	koeng5hang4	ho5soi4
sin3ngaan5	got3zau2	kwan3bik1	ku3co3	koeng5hang4	ngat2ge4
sin3ngaan5	lou6zoeng3	kwan3bik1	jing4wan6	koeng5hang4	gin1kyut3
lam4lap6	lam4lap6	zau2hoeng3	zau2hoeng3	sek6gwo2	sek6gwo2
lam4lap6	lam5lap6	zau2hoeng3	zau4hoeng3	sek6gwo2	sek3gwo2
lam4lap6	lam2lap5	zau2hoeng3	zau6hoeng1	sek6gwo2	sek4gwo6
lam4lap6	dang4lek6	zau2hoeng3	zou2sang3	sek6gwo2	jip6kwo2
lam4lap6	muk6lap6	zau2hoeng3	zu2ho2	sek6gwo2	zi3goe2
lam4lap6	muk3lip1	zau2hoeng3	fong1hoeng4	sek6gwo2	zik4si1

Table 12: The table of stimuli for disyllables

7.3 SUPPLEMENTARY MATERIALS TO SECTION 4

Description of monosyllable data

The descriptive data is shown in Figure 1. Each graph represents the data from one subject.[13] Each scatterplot shows the relationship between natural class-based segmental distance (x-axis) and the judged distance (y-axis). Black points are those with tonal distance of 0; dark blue dots are those with tonal distance of 0.5; light blue dots have tonal distances of 1.[14]

[13] We display data only from participants who answered all questions, plus a participant who answered 62/72.

[14] Note that this graph should only be treated as a rough visualization of the data that fails to display a substantial portion of useful information. In particular, there are many cases of overlapping points, but we

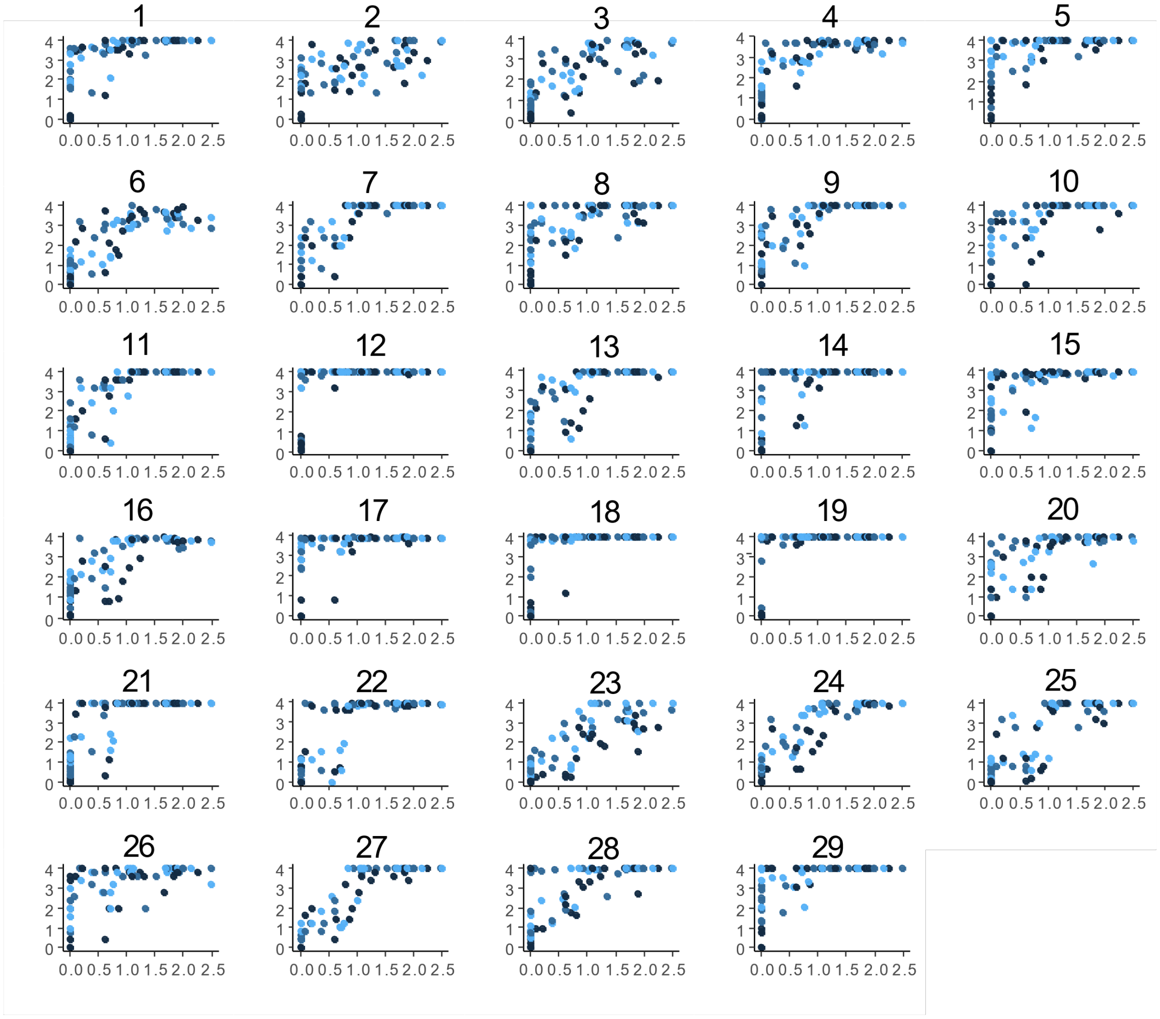

Figure 1: Scatterplots of distance judgements against theoretical segmental distance. Black points are those with tonal distance of 0; dark blue dots are those with tonal distance of 0.5; light blue dots have tonal distances of 1. Numbers indicate participants' numbers.

Consistent with expected behaviour, darker dots (which have lower tonal distance) have greater tendency to lie in the lower part of the y-axis in the plot. However, it is not uncommon to see them in higher parts in the y-axis even if the segmental distance is just slightly higher than 0. This suggests that although tonal distance contributes to the distance judgements, the relationship is not strong.

Moreover, it is clear in all graphs that there is a positive relationship between segmental distance and distance judgements, suggesting that the natural-class-based segmental distances are a good predictor of distance judgements. While this observation is not surprising, there seems to be substantial variation among the participants as to the nature of the positive relationship. Certain participants display a near-categorical approach to judgements, such as the17, 18 and 19, with almost all items judged as full distance if the segmental distance is nonzero. Others' judgements are more gradient; for example, 2 and 3 display gradience in judgement even when the segmental distance is above 2. Some others still display complex relationships intermediate

have not scaled the sizes of the dots according to the number of samples in a position because of insufficient space. Nevertheless, certain clear trends can be gleaned from the graphs.

between the two. For example, 1 displays categorical judgements above 1.5 segmental distance, while 15 is quasi-categorical beyond 1.

Details of the monosyllable model

Conceptually, the decision to treat the data as censored may appear strange, as 4 is already the full distance. However, in terms of fitting the data, using the censored-response model solves the problem of subjects from whom the judgements are at first gradient at the lower segmental distance zones, then becomes categorical above a certain point. Visually, fitting a purely linear model to these data without any modifications results in lack of fit. The censoring model resolves this problem by assuming the distance still increases beyond the categorical threshold in some underlying way.[15]

The intercept as well as the coefficients indicating the effects of tonal and segmental distance are subject-level effects in the full model, i.e. the slopes and intercepts vary by subject and are assumed to follow a normal distribution, with the variance-covariance subject between the slopes and the intercept unknown (to be inferred during model-fitting).[16] This allows the model to reflect variation in the panels shown in the two sets of scatterplots.

We also assume an item effect in the model, so that the intercept is affected by both the participant and the item. This item effect arises from fine phonetic differences between the two recordings provided. The differences vary from item to item and may be perceived differently from speaker to speaker, hence the intercept is affected by both participant and item effects.

The intercept had a Student's *t* prior with three degrees of freedom, location parameter 4 and shape parameter 10; the standard deviations of the group-level effects and the residual standard deviation had Student's t priors with three degrees of freedom, location parameter 0 and shape parameter 10; and the correlations among the subject-level parameters had LKJ Cholesky priors.

The following table lists all the models we built, including the full model and various reduced models. The first column assigns a label to the models using Roman numerals. The second column indicates whether the distance judgements are assumed to be right-censored. The third column indicates which subject-level effects are present, i.e. which parameters are assumed to vary by the subject. The fourth column indicates whether the intercept may vary by item. The full model, along with reduced models of various forms, are reported in Table 13 with their Watanabe-Akaike Information Criterion (WAIC):

Model	Censor	Subject-level effects present	Item-level	WAIC
I	No			7273.6
II	No	Intercept		6778.6
III	No	Intercept	Intercept	5360.6
IV	No	Intercept, segdist	Intercept	4980.4
V	No	Intercept, tonedist	Intercept	5344.4
VI	No	Intercept, segdist, tonedist	Intercept	4948.0
VII	Yes			7113.6

[15] In the literature, a similar way to deal with patterns of this shape is to log-transform the distances (Heeringa, 2004). This takes into account that a small amount of phonological change in the beginning leads to a large distance. We do not adopt this because the graphs shown in Figure 3 seem to favour the censoring approach.

[16] Following conventions in the Bayesian paradigm (Gelman and Hill, pp. 2-3, 225), we do not use the terms 'fixed effect' and 'random effect', which are roughly equivalent to group-level (i.e. subject-level or item-level) and population-level effects.

VIII	Yes	Intercept		6289.5
IX	Yes	Intercept	Intercept	5119.4
X	Yes	Intercept, segdist	Intercept	4817.6
XI	Yes	Intercept, tonedist	Intercept	5063.6
XII (Full)	Yes	Intercept, segdist, tonedist	Intercept	4727.1

Table 13: WAIC values of the full monosyllable model along with various reduced models.

The lowest WAIC suggests the full model (model XII) has the best predictive power of the distance judgment. This suggests that the results can be predicted best when we assume that the intercept varies by both the participant and the item, that the weighting of the tones and segments are both allowed to vary by the participant, and that the distance judgements are right-censored.

Parameter estimations in the monosyllable model

We now present the model parameters of the monosyllable model in graphical form. Figure 2 displays the point estimate of each parameter on the x-axis along with its 50% (dark blue) and 95% (light blue) uncertainty intervals:[17]

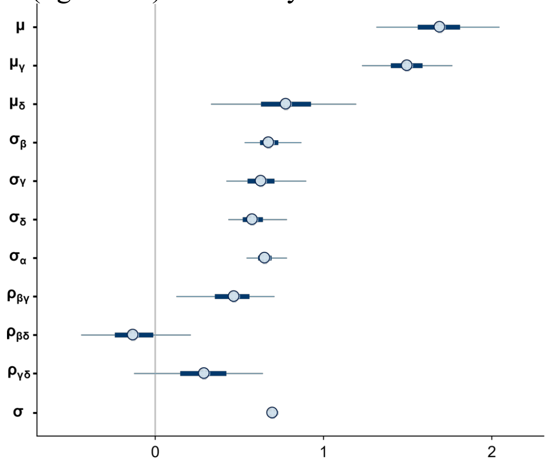

Figure 2: Estimates of the model parameters along with 95% and 50% credible intervals. μ is the overall (population-level) intercept, σ is the residual standard deviation, μ_γ and μ_δ are the mean coefficients of segmental and tonal distance, ρ_{AB} indicate the population correlation between A and B, and σ_A indicates the standard deviation of A. Note that the x-axis provides the numerical values of different types of parameters (intercept, population-level and group-level intercepts, standard deviations and correlation coefficients), and one must take care not to compare across these different types.

Note that the overall intercept μ is around 1.69 (SE: 0.19, 95% CI: (1.31, 2.05)). Under the current model, if we assume that the phonological distance metrics used capture all the information about phonological distance, this can be interpreted as the average inherent phonetic distance perceived in the recordings. The item-level and subject-level standard deviations (σ_α and σ_β), which quantify the variability in this perceived phonetic difference across items and subjects, are respectively are estimated at 0.65 (SE: 0.06; 95% CI: (0.55, 0.78)) and 0.68 (SE: 0.08; 95% CI: (0.53, 0.87)), suggesting that there is a fair amount of variation across both items and subjects. The population-level parameters on tonal and segmental weightings are discussed in the main text. The inter-subject variation in segment and tone weightings is quantified by the subject-level standard deviations σ_γ and σ_δ, which are respectively estimated at 0.61 (SE: 0.10, 95% CI:

[17] As we are not testing for any particular hypotheses, the intervals have not been corrected for multiple comparisons.

(0.48, 0.81)) and 0.67 (SE: 0.14, 95% CI: (0.48, 0.81)). Note thought that the two values cannot be compared directly because of the mean differences. In order to compare them, we need to consider the coefficients of variation. The corresponding coefficients of variation are 0.39 (SE: 0.06, 95% CI: (0.29, 0.54)) and 0.88 (SE: 1.12, 95% CI: (0.48, 1.81)). The difference between the two is estimated at -0.49 (SE: 1.12, 95% CI: (-1.14, -0.07)), so we have weak evidence that the variation in segmental weighting is less than the variation in tonal weighting.

It is also worth noting that while the correlation between segmental and tonal distance is estimated at around 0.28 , the 95% credible interval extends well beyond 0 (SE: 0.20, 95% CI: (-0.12, 0.64)). This indicates that we have insufficient evidence that they positively correlated. If it turns out that the population correlation were positive, however, it would suggest that people whose judgements are affected more heavily by segments are also affected more heavily by tones in general. Figure 3 plots the estimated tonal weightings against the estimated segmental weightings:

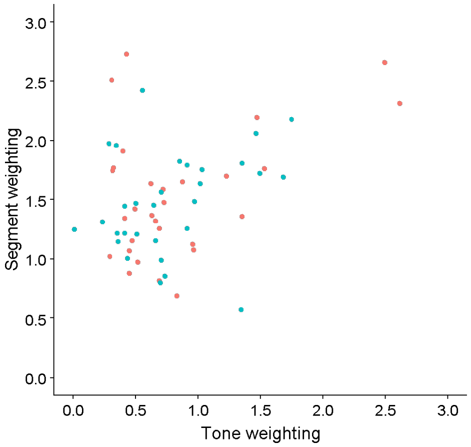

Figure 3: A graph showing estimated tonal and segmental weightings of each participant. Blue dots indicate participants who have completed less than 25% of the experiment. Note that we do not have strong evidence of a positive correlation between the two, despite the appearance of the graph.

WAIC values with information gain weighting

The WAIC values are shown below. They are either comparable or inferior to those without information gain weighting.

	Chao (H)	Chao (M)	Chao €	Autosegmental	O-C	O-C-O	C-O
Natural class	4772.4	4806.6	4802.1	4805.8	4757.1	4735.5	4738.0
Binary (H) (naïve weighting)	4773.9	4808.2	4806.2	4804.2	4762.5	4738.6	4742.1
Binary (H) (Broe weighting)	4776.6	4808.0	4803.4	4807.7	4762.3	4743.1	4743.1
Multivalued €	4735.4	4770.2	4767.4	4767.4	4721.0	4964.9	4692.3
Multivalued (M)	4757.9	4775.6	4770.7	4771.7	4771.7	4722.1	4699.1
Multivalued (H)	4724.5	4755.0	4751.7	4756.2	4708.3	4685.3	4682.4

Table 14: WAIC values of the monosyllable model using different segmental and tonal distances with information gain weighting.

7.4 Supplementary materials to Section 5

Description of disyllable data

We now turn to the exploration of the results for disyllabic items. The graphs of Figure 4 plot the segmental distance of each item (x-axis) against the judged distance (y-axis), and the colour of the dots represent tonal distance: The brighter the shade, the greater the tonal distance.

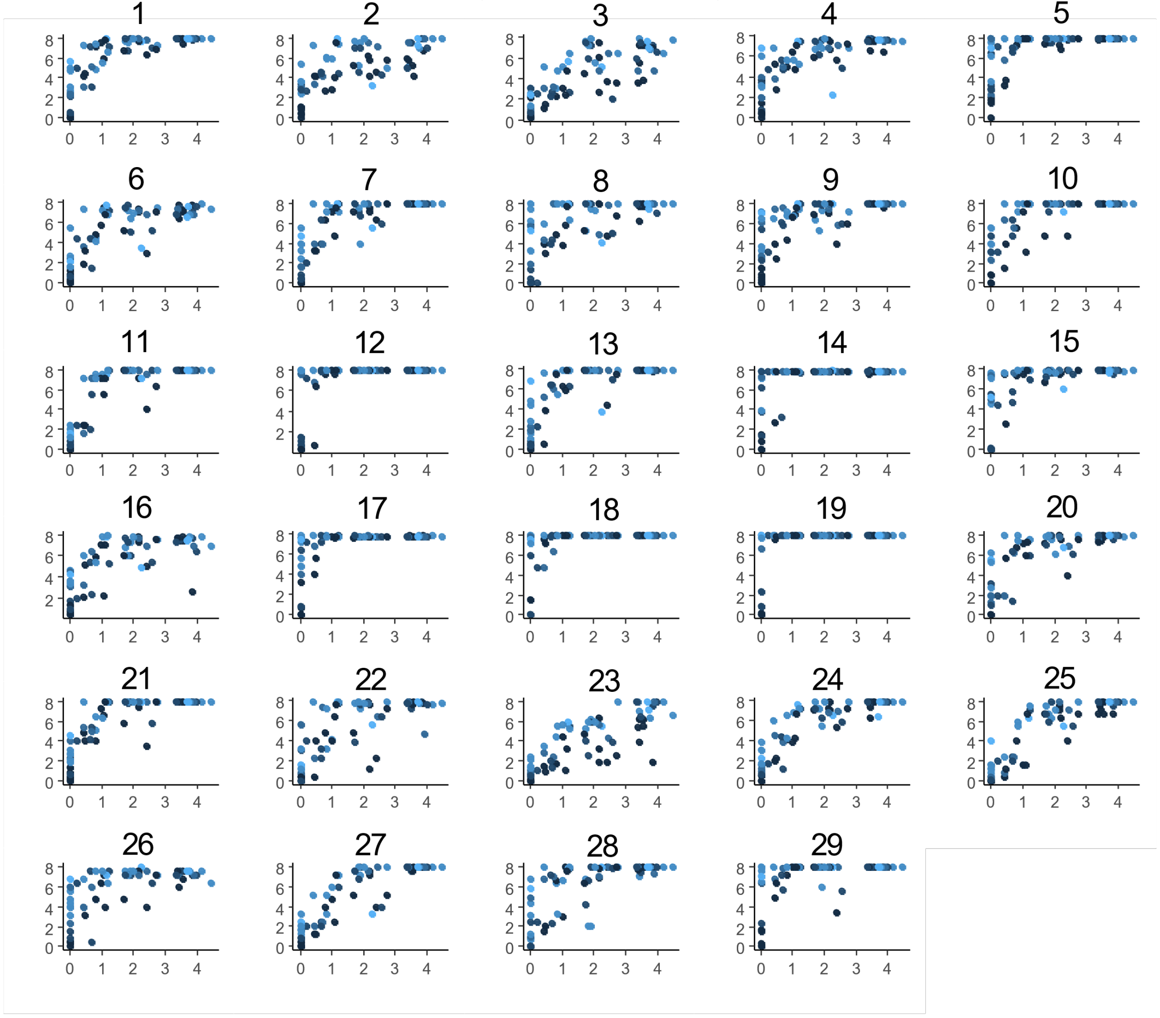

Figure 4: Scatterplots of distance judgements against theoretical segmental distance. Black points are those with tonal distance of 0; dark blue dots are those with tonal distance of 0.5; light blue dots have tonal distances of 1. Numbers indicate participants' numbers.

As before, there is a clear positive correlation between the segmental distance and the judged distance, and it is also clear that darker dots tend to occupy lower sections of the graphs than lighter dots, conforming to our expectations.

However, a major difference between this set of graphs and the previous ones is that the slopes seem to be gentler. In the monosyllabic data, many participants gave gradient judgements for the left half of the graph, but judged all distances to be maximum beyond a certain point. However, this situation is rare among the disyllables; most gradient judges gave gradient judgements all the way. For example, refer to the graph of the first participant. In the monosyllabic graph, all the judgements were of maximum distance after a certain point (at which the distance was 1.5). By contrast, in the disyllabic graph, the participant gave below-maximum judgements almost all the way; there did not seem to be a point after which all the judgements were categorical. This suggests a lighter weighting of segmental distance in the disyllabic case, compared to the monosyllabic case. The graphs also suggest a potential non-linear relationship for some participants. Note, however, that categorical judges such as participants 17-19 still gave categorical judgements for almost all items with segmental distance above 1.

Parameter estimations in the disyllable model

Again, the full model has the optimal WAIC as one can see below:

Model	Censor	Subject-level effects present	Item-level	WAIC
I	No			10873.5
II	No	Intercept		10402.0
III	No	Intercept	Intercept	8547.5
IV	No	Intercept, segdist	Intercept	8251.7
V	No	Intercept, tonedist	Intercept	8546.8
VI	No	Intercept, segdist, tonedist	Intercept	8221.5
VII	Yes			9816.5
VIII	Yes	Intercept		9028.1
IX	Yes	Intercept	Intercept	7633.2
X	Yes	Intercept, segdist	Intercept	7319.1
XI	Yes	Intercept, tonedist	Intercept	7563.1
XII (Full)	Yes	Intercept, segdist, tonedist	Intercept	7194.5

Table 15: WAIC values of the full disyllable model along with various reduced models.

The model parameters of the full model are as follows:

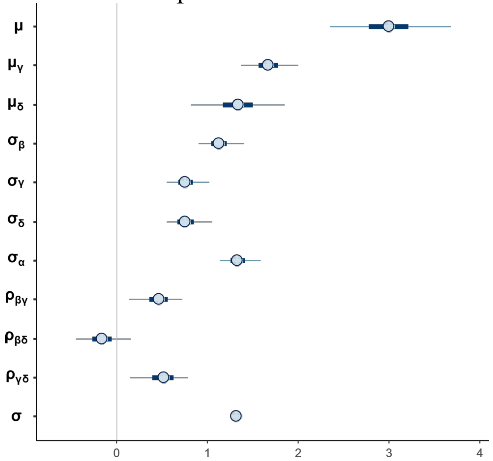

Figure 5: Estimates of the model parameters along with 95% and 50% credible intervals. μ is the overall (population-level) intercept, σ is the residual standard deviation, μ_γ and μ_δ are the mean coefficients of segmental and tonal distance, ρ_{AB} indicate the population correlation between A and B, and σ_A indicates the standard deviation of A. Note that the x-axis provides the numerical values of different types of parameters (intercept, population-level and group-level intercepts, standard deviations and correlation coefficients), and one must take care not to compare across these different types.

The overall intercept μ is around 3.00 (SE: 0.33, 95% CI: (2.35, 3.67)). As mentioned above, this may be interpreted as the average inherent phonetic distance perceived in the recordings. The intercept is smaller compared to the population-level intercept for monosyllables, which is point-estimated at 1.69, since the distance now ranges from 0 to 8 instead of 0 to 4, and halfing the estimated intercept for disyllables gives 1.5, which is smaller than 1.69. This may suggest that phonetic detail matters less when listeners compare disyllables. The item-level and subject-level standard deviations (σ_α and σ_β), which quantify the variability in this perceived phonetic difference across items and subjects, are respectively are estimated at 1.33 (SE: 0.12; 95% CI: (1.13, 1.58)) and 1.13 (SE: 0.13; 95% CI: (0.90, 1.40)), again suggesting that there is a fair amount of variation across both items and subjects.

The inter-subject variation in segment and tone weightings is quantified by the subject-level standard deviations σ_γ and σ_δ, which are respectively estimated at 0.76 (SE: 0.12, 95% CI: (0.55, 1.02)) and 0.67 (SE: 0.76, 95% CI: (0.13, 1.05)). Again, the two values cannot be compared directly because their means differ. The corresponding coefficients of variation are 0.46 (SE: 0.08, 95% CI: (0.33, 0.64)) and 0.59 (SE: 0.15, 95% CI: (0.37, 0.95)). The difference between the two is estimated at -0.13 (SE: 1.12, 95% CI: (-0.49, 0.13)), so unlike in the case of monosyllables, do not have evidence that one is greater or less than the other.

The correlation between segmental and tonal distance is estimated at around 0.50, although the 95% credible interval extends well beyond 0 (SE: 0.50, 95% CI: (0.15, 0.79)), providing evidence that they are positively correlated. Therefore, if someone's judgements are more heavily affected by tonal distance, they are likely to be more heavily affected by segmental distance as well. The following scatterplot shows the relationship between tonal and segmental weighting:

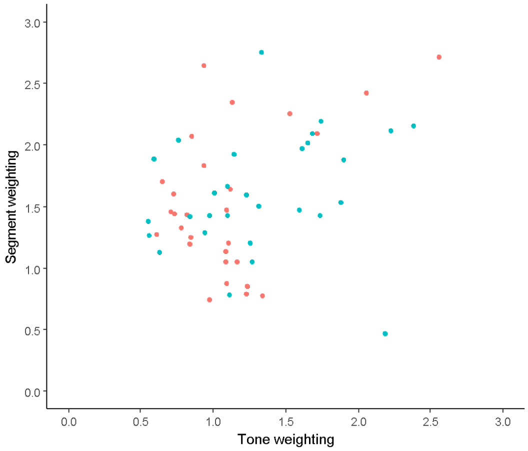

Figure 4: A graph showing estimated tonal and segmental weightings of each participant. Blue dots indicates participants who have not completed the experiment.

WAIC values with information gain weighting

The WAIC values are shown below. They are inferior to those without information gain weighting.

	Chao (H)	Chao (M)	Chao (E)	Autoseg-mental	O-C	O-C-O	C-O
Natural class	7244.3	7221.1	7227.7	7271.6	7226.7	7217.5	7208.3
Binary (H) (naïve weighting)	7229.2	7199.3	7210.0	7252.4	7204.0	7197.6	7189.3
Binary (H) (Broe weighting)	7223.7	7200.7	7211.3	7254.3	7205.0	7198.8	7190.3
Multivalued (E)	7163.0	7171.3	7183.31	7227.9	7182.4	7172.0	7156.2.
Multivalued (M)	7198.3	7170.8	7180.1	7229.2	7177.1	7167.9	7157.9
Multivalued (H)	7244.3	7221.1	7227.7	7271.6	7226.7	7217.5	7208.3

Table 16: WAIC values of the monosyllable model using different segmental and tonal distances with information gain weighting.

7.5 SUPPLEMENTARY MATERIALS TO SECTION 6

Estimations of syllable component weightings with lexical effects

Syllable	Component	Lexical?	Estimate	SE	95% CI
1	Onset - Nucleus	Y	0.09	0.52	(-0.91, 1.12)
1	Nucleus – Coda	Y	2.23	0.68	(0.92, 3.54)
1	Coda – Tone	Y	-1.53	0.55	(-2.6, -0.46)
1	Onset - Nucleus	N	-0.34	0.5	(-1.3, 0.66)
1	Nucleus – Coda	N	1.11	0.44	(0.26, 1.98)
1	Coda – Tone	N	0.47	0.37	(-0.27, 1.17)
2	Onset - Nucleus	Y	0.14	1.23	(-2.29, 2.52)
2	Nucleus – Coda	Y	-1.08	2.04	(-5.02, 3.15)
2	Coda – Tone	Y	1.8	1.09	(-0.34, 4.00)
2	Onset - Nucleus	N	1.39	0.58	(0.25, 2.54)
2	Nucleus – Coda	N	-0.69	0.62	(-1.91, 0.55)
2	Coda – Tone	N	0.45	0.4	(-0.35, 1.24)

Table 17: Estimations of syllable component weightings with lexical effects

Entropy calculations

We used a method to estimate the sampling distributions of the entropies and functional loads obtained as follows. Since the plug-in estimate of entropy uses the maximum likelihood estimators (MLEs) of the probabilities, and functions of MLEs are also MLEs, we have estimated the entropies using their MLEs. This means we can apply asymptotic properties of the MLE to estimate the error in our estimates.

We assume that the syllables in the corpus follow independent categorical distributions, i.e. they form a multinomial distribution. By the multivariate version of the Central Limit Theorem and the delta method (Casella and Berger, 2002; Rao, 1973), if a function *f* is differentiable near the true value θ_0 of a parameter θ with *k* components, then $f(\hat{\theta})$ approximately follows:

(8) $N(f(\theta_0), \alpha(\theta_0)I(\theta_0)^{-1}\alpha(\theta_0)^T)$

where $\alpha(\theta) = \left[\frac{\partial f(\theta)}{\partial \theta_1} \dots \frac{\partial f(\theta)}{\partial \theta_k}\right]$ and $I(\theta_0)$ is the information matrix. Here, θ is the vector of probabilities of each possible monosyllable or disyllable, excluding the final one (since it can be calculated by subtracting the rest of the probabilities from 1). It was calculated that the (i, j)-th entry of $I(\theta_0)$ has the form $\frac{n}{p_i} + \frac{n}{1-p_{k+1}}$ for diagonal entries and $\frac{n}{1-p_{k+1}}$ for off-diagonal entries.

For marginal entropies, *f* is the function that gives a vector of entropies with four components, including the onset, nucleus, coda and tone entropies. The (*j*, *i*)-th entry of $\alpha(\theta_0)$ is thus calculated to be $\log\, p^*_{k+1,j} - \log p^*_{i,j}$, where $p^*_{i,j}$ is the probability that a random syllable has the same *j*th component (*j* = 1 means 'onset', etc.) as the *i*th syllable; in particular, in the rows where the *j*th syllable component has the same value as the $(k + 1)$th (i.e. last) syllable, the entry is 0.[18]

For functional loads, *f* is the function that gives a vector of functional loads, again with four components. We first compute the derivatives of the entropy of the whole language and the entropy of the modified language separately, then find the derivative of the functional load using the quotient rule. The derivative of the entropy with respect to p_i is simply $\log p_{k+1} - \log p_i$,

[18] The intuition behind this result is as follows: the probability of the *j*th syllable component having the same value as the last syllable can be obtained by subtracting the probabilities of the other values from 1. So, for estimating the entropy of the *j*th syllable component, the probabilities of each of the syllables containing the *i*th syllable component don't matter.

whereas the derivative of the entropy of the modified language with respect to p_i is $\log p_{k+1}^* - \log p_i^*$, where p_i^* is the probability that a random syllable is the same as the ith syllable in the modified language under consideration. The value of α was then derived from these results.

In constructing the confidence intervals, we estimated the true values of the probabilities using their MLEs, since they are consistent estimators. The sampling distributions of the differences were found by multiplying the estimates of the entropies' and functional loads' distributions with the appropriate matrices.