Accounting for Lexical Tones When Modeling Phonological Distance

Youngah Do[1] and Ryan Ka Yau Lai[2]

[1] Department of Linguistics, The University of Hong Kong

[2] Department of Linguistics, University of California, Santa Barbara

Author Note

Youngah Do ✉ youngah@hku.hk

Ryan Ka Yau Lai ✉ kayaulai@umail.ucsb.edu

Correspondence concerning this article should be addressed to Youngah Do, Room 918, Run Run Shaw Tower, Centennial Campus, Department of Linguistics, The University of Hong Kong, Pokfulam Road, Hong Kong.

Accounting for Lexical Tones When Modeling Phonological Distance

Abstract

Methods of quantifying distance between sound sequences are known as phonological distance measures. Despite their wide applications across subfields, phonological distance has been calculated mainly with features related to consonants and vowels. This paper establishes new measurements of phonological distance by incorporating lexical tone through experimental approaches and modeling, using Hong Kong Cantonese as a case study. Results show correspondences between the experimental data and predictions from information-theoretic measures, including entropy measure and functional load, suggesting that lexical components playing a more crucial role in phonological distance judgments are also lexically less predictable. Implications on phonological distance measures are discussed.[1]

Keywords: phonological distance, tone, Bayesian modeling, entropy measure, functional load, Cantonese

[1] An early version of this paper was presented at Society for Computation in Linguistics (SCiL) in 2019. In addition to that audience, we thank *Language* reviewers and editors for their great help to improve the paper. Many thanks to Diana Archangeli, Adam Albright, and Arthur Lewis Thompson for helpful questions and comments. This project was supported by The University of Hong Kong's Seed Fund for Basic Research 201611159006 awarded to the first author.

1. INTRODUCTION. Native speakers of English have intuitive knowledge that the word *cat* is phonologically more similar to *cap* than it is to *ban*. On what basis do native speakers make such similarity judgments and how can we systematically measure the degree of (dis)similarity between words? Methods of quantifying distance between sound sequences are known as phonological distance measures and their usefulness spans a wide variety of linguistic subfields. In dialectology, phonological distance measures are used to examine divergence between dialects (e.g. Heeringa 2004, Heeringa et al. 2006, Nerbonne & Heeringa 1997, Tang 2009, Tang & van Heuven 2009 2011 2015). In computational historical linguistics, distance measures help to align and reconstruct cognate words (e.g. Oakes 2000). In psycholinguistics, distance measures are often adopted in studies of bilingualism and diglossia to investigate effects of between-language or between-variety similarity (e.g. Saiegh–Haddad 2004). Some older methods of automatic speech recognition compare reference symbols with hypothesized symbols using distance measures (Fisher & Fiscus 1993). In phonology, distance measures inform the formulation of constraints on alternations (Gildea & Jurafsky 1996) and phonotactics (e.g. Frisch et al. 1997, Pierrehumbert 1993), and specifically in phonotactics, phonological distance is a crucial factor in modeling phonological neighborhood density, the degree to which a sound sequence overlaps with existing words in the lexicon. Models built on phonological distance measures have been applied to spoken word recognition as a predictor in experimental paradigms (Luce et al. 2000, Luce & Pisoni 1998), to the investigation of speech errors (Vitevitch 1997), and to the explanation of some phonological phenomena such as asymmetries between roots and affixes (Ussishkin & Wedel 2002).

The validity of phonological distance-based methods hinges on the quality of the distance measure, that is, the extent to which it resembles human listeners' perceptual distance. Despite the wide application of phonological distance, its methodological approach so far has been predominantly concerned with segmental features (e.g. Heeringa 2004, Nerbonne & Heeringa 1997, Somers 1998) and work incorporating suprasegmental features, like tone or stress, is rare. While this methodological oversight is not particularly relevant to some languages, where the lexical role of suprasegmental features is relatively small (e.g. languages with a positional stress system such as Finnish, Armenian, or Polish), it cannot be overlooked in languages where suprasegmental features are essential in creating lexical contrasts (e.g. tonal languages such as Cantonese or Igbo; or pitch-accent languages like Swedish or Japanese). For example, Malins

and Joanisse (2010) point out that it is uncertain how the neighborhood activation model of Luce and Pisoni (1998) applies to spoken word recognition in Mandarin, because the model does not specify how 'neighbor' is defined in a tone language.

Against this background, this paper aims to find a way to measure phonological distance that best reflects speakers' judgments in languages where suprasegmental features are crucial in creating lexical contrasts. Out of all the suprasegmental features used to create lexical contrasts across languages, we have chosen to focus on tone. While a few studies have considered tonal distance measures with limited discussions on their quality or nature (e.g. Neergaard & Huang 2016 2019, Tang & van Heuven 2009, Yang & Castro 2008, Yao & Sharma 2017), to the best of our knowledge, none have made tonal distance measures their object of study. In order to create a phonological distance measure which accounts for both tones and segments, we implement an experimental and modeling approaches to Hong Kong Cantonese (Cantonese hereafter) as our case study. Among lexically contrastive suprasegmental features, tone can involve relatively rich representations, including level and contour tones. Cantonese, with its multiple level tones (Tone 1, high-level; Tone3, mid-level; Tone 6, low-level) and contour tones (Tone 2, high rising; Tone 5, low rising, Tone 4, falling), allows us to design a methodological approach based on both pitch-based and contour-based tone contrasts.

To find out the optimal phonological distance measure that matches native speakers' judgments, Section 2 first defines a variety of variables which can be used as metrics to compare phonological distance among segments and tones. Section 3 presents a phonological distance judgment experiment, from which native speakers' judgment data between sound sequences is obtained. Through the comparisons of experimental results with theoretically predicted distances, we investigate how best to predict speakers' distance judgments. To do so, specifically the following three topics will be explored: (a) relative contributions of segmental and tonal distances to phonological distance judgments; (b) phonological distance metrics of segments and tones that can best predict speakers' judgment data; and (c) relative contributions of syllable components (onset, nucleus, coda, tone) in calculating phonological distance. To further consider each of the syllable components' contribution to phonological distance judgments, Section 4 attempts a lexical analysis and shows correspondences between the experimental data and predictions from information-theoretic measures. We employ two types of information-theoretic measures of syllabic components, entropy measures and functional loads, and show that syllabic

components playing a more important role in phonological distance judgments are lexically less predictable. Section 5 discusses methodological implications of the current study.

2. DISTANCE MEASURES.[1] This section provides an overview of the distance measures that will be tested against our experimental data in Section 3. Segmental distance measures will be presented first, followed by tonal distance measures. We then discuss how we apply the measures to calculate phonological distance in Cantonese.

2.1. SEGMENTAL DISTANCE.

PHONEMIC DISTANCE. As an initial step to determine how similar two segmental sequences are to one another, we first measure the distance between phonemes. A simple approach is to classify pairs of phonemes as either the same (e.g. /b/ and /b/) or different (e.g. /b/ and /p/), with no elaboration on the kind or extent of the difference (e.g. Heeringa et al. 2006, Tang & van Heuven 2015, inter alia). The binarity of this approach, thus, does not take the gradient differences between phonemes into account, for example, /b/ is equidistant from both /p/ and /l/. Two other methods of measuring phonemic distance take a more nuanced approach by using binary phonological features. The first method finds the number of feature values (e.g. [±voice], [±nasal]) that are different from phoneme to phoneme. The number of different feature values is normalized by dividing by the total number of phonological features, as shown in 1.[2] This method is called HAMMING DISTANCE measure (e.g. Gildea & Jurafsky 1996, Pierrehumbert 1993).

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

The Hamming distance measure does not take into account how phonological features are used to create contrasts between phonemes; it is purely based on counts of different feature values. Contrary to the Hamming distance measure, a second distance measure is based on phonemes' natural classes. As in 2, the number of nonshared natural classes between two phonemes is divided by the total number of shared natural classes and nonshared natural classes of the two phonemes. The formula for the natural class-based measure is in 2,[3] adapted from Frisch and colleagues (1997).

(2) $$Distance_{Natural\ Class} = \frac{Nonshared\ natural\ classes}{Shared\ natural\ classes + nonshared\ unnatural\ classes}$$

As an example of the natural class-based measure, refer to the following system of stops.

	lab	voi
p	+	−
b	+	+
t	−	−
d	−	+

Natural class memberships, defined by finding all possible combinations of feature values, are as follows.

	[+lab]	[−lab]	[+voi]	[−voi]	[+lab, +voi]	[+lab, −voi]	[−lab, +voi]	[−lab, −voi]
p	Y			Y		Y		
b	Y		Y		Y			
t		Y		Y				Y
d		Y	Y				Y	

To calculate the phonemic distance between /b/ and /d/ based on the natural class-based measure, the number of nonshared natural classes between /b/ and /d/ is divided by the number of all natural classes between the two phonemes, that is, shared natural classes and nonshared natural classes, ([+lab], [−lab], [+voi], [+lab, +voi], [−lab, +voi]). Here, the natural class-based phonemic distance is $4/5 = 0.8$, because the two phonemes share only one of the five natural classes. To calculate the phonemic distance between /b/ and /d/ with the Hamming distance measure, the number of nonshared features is divided by the number of all relevant features. The Hamming distance will be $1/2 = 0.5$, because between /b/ and /d/, labial feature value is different but their voicing feature value is same.

In dialectological studies, multivalued features are widely used instead of binary features. In this measure, each feature permits more than two categories or a range of values along a scale. For example, a place feature may be bilabial, coronal or dorsal (categorical), or a place feature may hypothetically assign values 100 for bilabial, 80 for coronal and 20 for dorsal (numeric). When the features are categorical, the Hamming distance measure in 1 can be adopted to count the number of different features. If the values are numeric, other distance measure are required, either EUCLIDEAN or MANHATTAN DISTANCE (Nerbonne & Heeringa 1997[4]). In the Euclidean distance measure in 3, phonological distance is calculated by evaluating the square of the difference between the feature values of two phonemes and taking the square root of the sum. To visualize the concepts, Figure 1 shows that the Euclidean distance is diagonal shown in blue, with the *x*- and *y*-axes assumed to be feature values in a two-feature system. The Manhattan distance measure shown in 4 is similar but it sums up the absolute values of the differences

between the corresponding feature values of a phoneme pair. In Figure 1, the Manhattan distance is shown in red. In the formulas for Euclidean and Manhattan distance below, $f_i(p_j)$ refers to the *i*-th feature value of the *j*-th phoneme and $f_i(p_k)$ refers to the *i*-th feature value of the *k*-th phoneme.

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

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

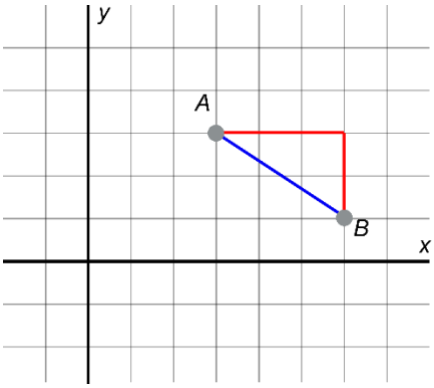


FIGURE 1. Euclidean distance (blue) and Manhattan distance (red) between two points on the Cartesian plane. Here, the *x*- and *y*-axes can be taken as feature values in a two-feature system.

An assumption behind the distance measures in 1 to 4 is that features are weighted equally. For instance, two phonemes differing in the [±voi] feature are assumed to be equally distinctive to the two phonemes differing in the [± continuant] feature. However, this assumption may not be true. There are several ways to assign different 'weights' to different features. First approach considers the weights as free parameters and a model finds the optimal weights to best predict the distance (Kondrak 2002). Theoretically, this could be achieved by introducing the weights as parameters in multivalued feature representations, but the weight of each individual feature would add an additional parameter, which is not ideal for a modeling purpose. A second method is Nerbonne and Heeringa's (1997) information-theoretic approach in which each feature is multiplied by a weight determined by 'information gain'. Roughly speaking, the weight of a feature is determined by calculating how much information a feature gives us about the lexicon. Put another way, features are weighted by their roles in predicting the lexicon, which is calculated by the difference between the amount of 'uncertainty' in identifying a segment in the

lexicon and the average degree of uncertainty left once we determine the value of the feature in question. A problem of Nerbonne and Heeringa's (1997) information-theoretic approach is that it cannot deal with null features. A third approach is a modified information-theoretic approach, which takes into account that certain feature values may be null (Broe 1996). The formula and Broe's modification are presented in Appendix A-2.

In the current study, we apply the distance measures from 1 to 4 to calculate phoneme distances in Cantonese. We always normalize the distances so that the phoneme distances in each syllabic position range from 0 to 1. The exact binary feature set of Cantonese on which the Hamming distance calculation is based is presented in Table A-1.1 of Appendix A-1 with reference to Hayes (2011). When we establish a system of multivalued features (Kessler 1995, Kondrak 2002) in Cantonese, we construct a feature matrix based on Ladefoged's (1975) table, which incorporates primarily articulatory and some acoustic features. The features are shown in Table A-1.2 of Appendix A-1. We also consider information-theoretic weightings, both classic information gain weighting following Nerbonne and Heeringa (1997) and a modified information gain weighting following Broe (1996).

DISTANCE BETWEEN PHONEME SEQUENCES. To measure phonological distance between words, calculating the distance between individual phonemes will not suffice. The distance between phoneme sequences needs to be measured. Measurements used for sequences of equal length (e.g. Hamming distance) will not be adequate here, because two sequences can differ in their length. In such case, the LEVENSHTEIN DISTANCES (Jurafsky & Martin 2014) can be computed. The Levenshtein distance measure finds the optimal sequence of substitutions, deletions, or insertions required to transform one sequence into another while minimizing the total 'cost' of these operations, cost being the distance between phoneme sequences involved. For substitutions, the cost is the phonemic distance defined above in 1 to 4 or a simple all-or-nothing cost (i.e. no change vs. change). The cost for insertions and deletions is assumed to be half the average of all phonemic distances between any two phonemes (Nerbonne & Heeringa 2001). Take an example of a distance from *ka* to *tap*, where a substitution in the onset position, /k-/, and an addition in the coda position, /-p/, are observed. If the optimal distance from /k-/ to /t-/ is 0.5 and if the average phoneme distance cost is 0.8, then the total distance from /ka/ to /tap/ is 0.9, with 0.5 for the /k-/ to /t-/ substitution and 0.4 for the addition of the coda /-p/. While other operations are conceivable, for example, turning /-a-/ into /-p/, then inserting /-a-/ before /-

p/, they incur higher cost and hence are not used as the final distance. The maximum possible distance between two strings, assuming that all operations are substitutions with a cost of 1, is 3.

An assumption so far is that phonemes in a word are linearly ordered. Some measures assume a syllabic structure, as opposed to a linear string, and our study consider these measures as well. The first is the syllable part approach (Bailey & Hahn 2001) where the Levenshtein distance is computed not over individual phonemes but over the onset, nucleus and coda. The distances based on the syllable part approach may differ from the those from a simple linear string-based measure when multiple phonemes are allowed in one syllabic position (e.g. consonant clusters). Evidence showing the psychological reality of rime, primarily from priming studies (Radeau et al. 1998, Turnbull & Peperkamp 2017) support another method, namely the syllable-rime approach. According to the syllable-rime approach, the distance of sound sequences is computed over onsets and rime, combining nucleus and coda together.[5]

Table 1 summarizes the segmental distance measures introduced in this section. Numbers below match the numbers of corresponding formulas in Section 2.1. These measures will be adopted in our study of Cantonese.

Distance measure between phonemes	Featural representation	Abbreviation	Distance measure between phoneme sequences
All-or-nothing	None	Simple	Levenshtein (with the assumptions of linear strings or syllabic structure)
Hamming (1)	Binary	Binary	
Natural class (2)		Natural class	
Hamming (1)	Multivalued	Multivalued (H)	
Euclidean (3)		Multivalued (E)	
Manhattan (4)		Multivalued (M)	

TABLE 1. Summary of the different distance measures investigated in this paper.

2.2. TONAL DISTANCE.

TONAL REPRESENTATIONS. In order to measure distance between tones, first we should touch on the ways in which tones are represented. We do this with respect to Cantonese. For future work calculating phonological distance in other tone languages, the same types of representations can be adopted but the specific numbers of representations and their descriptions should be modified depending on the tonal system of the language concerned.

Of the six tonal representations presented in Yang and Castro (2008), the following five can be adopted for Cantonese: (a) Chao tone letters, (b) autosegmental, (c) onset-contour, (d) onset-contour-offset, and (e) contour-offset representations of tone.[6] The Chao tone letters were Chao's original proposal, except that in the current study Tone 1 has been fixed at Chao tone

letter 55 instead of 53 because 53 is no longer phonologically contrastive in Hong Kong Cantonese (Bauer & Benedict 1997).[7] The autosegmental representation are based on Yip's (1980) framework, describing the tonal phonology of Chinese varieties using a two-tiered system, including Register, which is either upper (+) or lower (−), and Tone, which is either high (H) or low (L). The onset-contour-offset representations ((c) onset-contour; (d) onset-contour-offset; (e) contour-offset) follow standard tone descriptions such as Bauer and Benedict (1997), where the offset is extrapolated using the onsets and Chao tone letters. The six tones in Cantonese are diagramed in Figure 2, and their corresponding tonal representations are shown in Table 2.

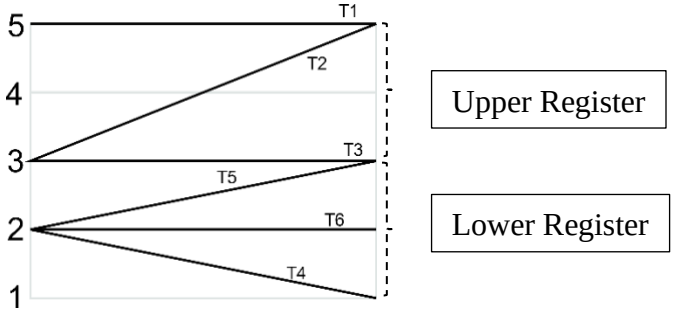


FIGURE 2. A graphical illustration of the Chao tone letter representations of the six Cantonese tones.

Tone	(a) Chao tone letters	(b) Autosegmental		(c–e) 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
4	21	-	LL	L	F	L*
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. 'L*' indicates 'Lowest'.

FROM TONAL REPRESENTATIONS TO TONAL DISTANCE MEASURES. With the tonal representations in Table 2, tonal distances between two tones are calculated. Here, the distance measures introduced in Section 2.1 can also be applied. For all tonal representations in Table 2, the Hamming and Levenshtein distances can be calculated, when tonal differences are assumed categorical (i.e. two tones are same or different). The distances between two 'symbols' are set to be 0 when they are same and 1 when they are different, where each character in the

representations in Table 2 is treated as a 'symbol'—each number representation in a, +/−, high (H), and low (L) in b, high (H), mid (M), low (L), lowest (L*), and rising (R) in (c–e). The distances are then divided by the number of symbols used in the representation. Our distance measures with Cantonese showed that the Hamming and Levenshtein distance measures resulted in no differences for the tonal representations (a) and (b) (see Table A-3.4–A-3.7 in Appendix A3 for the calculated values). Additionally, because each segment sequence has the form of onset, contour and offset, the Hamming and Levenshtein distance measures for the tonal representations (c) to (e) should be same. So we only report the Hamming distance in the following sections. For the Chao tone letters (a), the Euclidean and Manhattan distances can be computed as well because each tone letter can be treated to bear its own numeric value (scalars). Note that this is because the nature of Chao tone letters representation is different from the autosegmental (b) or onset-contour-offset representations (c)–(e), where they involve only categorical values. We calculated tonal distances in Cantonese based on the Hamming, Euclidean, and Manhattan distance measures, shown in Table A-3.1–A-3.3 in Appendix A3. To be consistent with the segmental distances, the tonal distances are always normalized as well so that all tonal distances range from 0 to 1. Since, for any given syllable, the distance of each onset, nucleus, and coda is set 1, thus the maximum segmental distance is the sum of these three distances 3, our assumption here is that tone has as about as much potential contribution to distance as one segment, resulting in tonal distances ranging from 0 to 1.

In this paper, we will also consider distances between disyllables. For such distances, we will calculate segmental and tonal distance by simple addition: The segmental distance of the disyllabic sequence is the sum of the segmental distances of the two individual syllables, and similarly for the tonal distance. Thus, segmental distances will range from 0 to 6, and tonal distances from 0 to 2.

2.3. INTERIM SUMMARY. Section 2 first introduced several ways of evaluating the distance between segments. They include measures assuming binary features, natural class, and multivalued features. Based on these phoneme representations, the distances between phoneme sequences are calculated by applying the Hamming, Euclidean, and Manhattan distance metrics. We have also presented five types of tonal representations, including Chao tone letters, autosegmental, onset-contour, onset-contour-offset, and contour-offset representations. The Hamming, Euclidean, and Manhattan distance metrics are applied to calculate tonal distances for

these tonal representations. Against this background, we will now use Cantonese as a case study to show how phonological distance between words should be calculated for tone languages with the following three topics: (a) the relative contributions of segmental and tonal distances to phonological distance judgments; (b) the optimal segmental and tonal distance metrics that can best predict speakers' distance judgments; and (c) the relative contributions of syllable components (onset, nucleus, coda, tone) to phonological distance judgments. To explore these topics, we first obtain phonological distance judgment data from native speakers. Our experiment presents a pair of items varying in degrees of segmental and tonal distances and asks native speakers to judge the similarities between the two sound sequences.

3. PHONOLOGICAL DISTANCE JUDGEMENT.

3.1. EXPERIMENT.

DESIGN. A question set of 72 monosyllabic and 72 disyllabic sound sequences was created. The stimuli list is provided in Table B1.1–B1.2 in Appendix B-1. When designing the stimuli, two criteria were considered: First, the items are well balanced across different segmental and tonal distances and second, segmental and tonal distances are not correlated among stimuli. Given that one of the three focuses of this experiment is to figure out relative contributions of segmental distance and tonal distance in making phonological distance judgments, it was important to keep the segmental and tonal distances uncorrelated. The stimuli design was based on the natural class-based distance measure for segments following Bailey and Hahn (2001) and the Hamming distance measure between onset-contour-offset tonal representations for tones following Yang and Castro (2008) and Tang and van Heuven (2011). For monosyllables, multiple simulations by picking 10,000 monosyllables from the Hong Kong Cantonese Corpus (Luke & Wong 2015) at random showed that the natural class-based distance measure of segments rarely went above 2.5. Based on this observation, segmental distances were set within the interval of [0, 2.5] and divided into four regions: high (>1.67), mid (≤1.67 but >0.83), low (≤0.83 but nonzero), where each region occupies one third of the interval, plus those with zero distance (i.e. two same items). Similarly, tonal distances were divided into three regions: high (1, farthest apart), low (0.5, middle), and zero (0, no distance). For disyllables, the interval for segmental distances was [0, 5], simply double that of monosyllables. This interval was divided into four regions, high (>3.33), mid (≤3.33 but >1.67), low (≤1.67), and zero (0), to

be consistent with monosyllables. Tonal distance was classified as high (>2), low (≤1), or zero (0), again doubled from monosyllables. For both monosyllables and disyllables, it was ensured that each segmental distance and tonal distance region was selected the same number of times in the stimuli. It was also ensured that each segmental distance and tonal distance region pair was shown the same number of times in the stimuli. Additionally, every possible segment in every position appeared equal amount. In both monosyllabic and disyllabic pairs, the first item within a pair was an existing word in Cantonese, whereas the second item was either an existing word (e.g. *pei4* 皮 'skin', *mui4gwai3* 玫瑰 'rose') or a nonword (e.g. *poe6*, *doi6te3*).[8] When creating the nonwords, phonotactically illegal phonemes in onset, nucleus and coda positions were excluded; for example, no fricatives were in coda position, following a Cantonese phonotactic restriction. However, any other phonotactic constraints were not considered in stimuli design as such constraints will be discovered later through phonotactic modeling.

A native speaker of Hong Kong Cantonese who is not affected by ongoing sound changes in Cantonese, such as the merging of onsets [n-] and [l-] and merging of codas [-t] and [-k], recorded the items. All the items were recorded in a sound-attenuated booth in the first author's institute. The recordings were scaled to 70 dB using the Scale intensity feature in Praat (Boersma & Weenink 2009). 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) and directed towards native speakers of Hong Kong Cantonese.[9] 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 is easier to understand similarity than distance, the participants were asked to rate similarity between the two items ranging from 0 to 100, where 0 means the two items were completely different and 100 means they were identical. The similarities were then converted into distances by subtracting the similarity from 100. Before the judgments test, a screening task was added in forms of AXB tests to ensure that participants could perceptually distinguish between [n] and [l] onsets, which are merging in some Cantonese speakers (Bauer & Benedict 1997), and that they could distinguish between tones 2 and 5, 3 and 6, and 4 and 6, which are also merging in some Cantonese speakers (Mok et al. 2013). This perception test was to make sure that the Cantonese spoken by

participants is fairly homogenous and rarely involves dialectal varieties. If participants submitted an incorrect answer to any of screening questions, the experiment stopped.

PARTICIPANTS. Anonymous participants were recruited after circulating the survey on social media platforms in Hong Kong using snowball sampling. The number of participants who passed the screening task was 61. Twenty-nine participants completed all 144 questions while others submitted incomplete forms (mean completion rate = 97%). The data from all of the participants were used to fit the model regardless of completion, as the model is able to handle variable sample sizes: Participants who did not complete the survey simply have their estimates shrunk to the population-level mean, whereas participants who have answered all of the questions will have subject-level coefficient estimates influenced largely by their own judgments (Gelman & Hill 2007).

3.2. RESULTS. We first explore descriptive patterns in the data to inform our modelling decisions. Distances (i.e. judgments from participants) ranged from 0 to100; these were divided by 25 to make them range from 0 to 4. This was to make the range of distance judgement same as that of theoretical distances if tone, onset, nucleus, and coda each gets a distance of 1. Figure 3 and Figure 4 show results from monosyllables and disyllables respectively. Each scatterplot represents the data from a participant who completed the test. Each scatterplot shows the relationship between the judged distance from a participant (*y*-axis) against the theoretical segmental distance calculated using natural class-based distance measures for segments (*x*-axis, in the range from 0 to 3, with onset, nucleus, and coda each gets a distance of 1) and against the theoretical tonal distance, calculated using Hamming distance measures for the onset-contour-offset representation; black points are items with tonal distance of 0; dark grey dots are items with tonal distance of 0.5; light grey dots have tonal distances of 1.[10]

FIGURE 3. Scatterplots of distance judgments on monosyllables against theoretical segmental distance. Light grey points are those with tonal distance of 0; dark grey dots are those with tonal distance of 0.5; black dots have tonal distances of 1. Numbers indicate participants' numbers.

FIGURE 4. Scatterplots of distance judgments on disyllables against theoretical segmental distance. Light grey points are those with tonal distance of 0; black dots have tonal distances of

2; intermediate shades indicate values in between the two extremes. Numbers indicate participants' numbers.

As shown in Figure 3 and Figure 4, there seems to be a rough correlation between the judged distance from participants (*y*-axis) and theoretically predicted segmental distance (*x*-axis): Segmentally distinctive items were judged more different. It is less clear, at a descriptive level, whether tonally more distinctive patterns (black > dark grey > light grey) were also judged as more or less different. Crucially, scatterplots both from monosyllables and from disyllables show that the strength of the relation between the judged distances and the theoretical distances varies greatly among participants: Some judged categorically while others judged in a more gradient fashion, and the thresholds to perceive the maximal distance differ among individuals.

The descriptive observations above informed our modeling decisions: We chose a multilevel model that allows an item-level random intercept as well as subject-level random slopes for tonal and segmental distances. The use of multilevel modelling allows the partial pooling of data from different items and from different participants so that the model can consider variability in the data as well as can produce high-variance estimates (Barth & Kapatsinski 2018, Gelman & Hill 2007). Also, instead of a frequentist approach, we chose Bayesian modelling (Gelman & Hill 2007, Nicenboim & Vasishth 2016). Bayesian models allow us to use 'priors' on various parameters to make it easier for the fitting algorithm to converge, which is frequently hard with data including large variations as in our case. Finally, in this model, the distance judgments were treated as a right-censored variable (Gelman et al. 2014:225–226), which assumes that there are some underlying distances which may exceed 4 (1 for onset, nucleus, coda, and tone each) but the data is truncated if the number goes beyond it, the setting of which can be justifiable from the raw data in Figures 3 and 4. The models were fit using the R package brms version 2.4.0 (Bürkner 2017a 2017b), which provides a lme4-like interface to the Stan language (Carpenter et al. 2017). Since we have little evidence for relevant priors on the topic, we relied on default priors provided by the package.[11]

Model specifications may vary, thus we first need to identify the optimal model specifications, from which we report our results. For this, we relied on the WATANABE-AKAIKE INFORMATION CRITERION (WAIC) values. Roughly speaking, the lower the WAIC values the better the model's performance. Comparisons of the WAIC values of the fully specified model with various reduced models showed that the full model (i.e. the model containing the item-level

random intercept, all subject-level random effects, as well as the censoring assumption) is the best. Therefore, results in the following sections are based on the full model. Detailed justification of the model specifications, as well as the model comparison procedure, are provided in Appendix C.

Recall our three specific objectives in the experiment. To understand how native speakers make phonological distance judgments, we aim to (a) find out relative contributions of segmental distance and tonal distance, (b) identify the ideal distance metrics to predict native speakers' distance judgements, and (c) determine relative contributions of onset, nucleus, coda, and tone within a syllable.

To answer (a) and (b), we fit the full model to different segmental and tonal distance measures presented in Section 2, comparing their predictive power. To answer (c), we run an additional model that separates onset, nucleus, and coda distances. Apart from the case where we identify the ideal distance metrics (question (b) above), all of the models throughout the result section are based on natural class-based distance measure for segments and the Hamming distance measure between onset-contour-offset tonal representations for tones. This was to make it consistent with our stimuli design, which was created with these two distance measures. Results are reported following the order of (a)–(c) above.

RELATIVE CONTRIBUTIONS OF SEGMENTAL DISTANCE AND TONAL DISTANCE. The first objective is to find out relative contributions of segmental and tonal distances to phonological distance judgments. If the population-level coefficient for segmental distance exceeds that of tonal distance, the result indicates that segmental distance is weighted higher than tonal distance, and vice versa. First, for monosyllables, examining the model parameters suggests that segmental distance is weighted more than tonal distance in predicting the distance judgement data. The population-level estimates[12] of the coefficient of segmental distance (μ_γ) is estimated at 1.50 (SE: 0.14, 95% CI: (1.23, 1.77)), much higher than that of tonal distance (μ_δ), estimated at 0.77 (SE: 0.22, 95% CI: (0.34, 1.19)). To check if this difference is significant, we employed the brms package using posterior draws. This shows how reliable the result is, by providing a 95% credible interval of the difference between the coefficient of segmental distance and that of tonal distance. The result showed that the credible interval of the difference ($\mu_\gamma - \mu_\delta$) excludes zero, (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 segmental distance is, on average, weighted more than tonal

distance. In other words, segmental changes contribute a lot to the judgement of phonological distance; tonal changes are less important.

Second, for disyllables, no strong evidence was found that the population-level coefficients of segmental and tonal distances (μ_γ and μ_δ) are different; the former was estimated at 1.67 (SE: 0.16, 95% CI: (1.37, 2.00)), while the latter was estimated at 1.34 (SE: 0.26, 95% CI: (0.81, 1.85)). A 95% credible interval of the difference between the two ($\mu_\gamma - \mu_\delta$) included zero, (−0.23, 0.91) (point estimate: 0.34; SE: 0.28), indicating that the weight difference between segmental and tonal distances is not significant.

An important caveat here is that in our study, tonal distances range from 0 to 1, whereas the segments of each syllable ranges from 0 to 3, because each segment is within the range from 0 to 1. Thus, we assume that tone contributes the same amount of potential distance as a single segment, rather than a whole syllable. If we were to standardize segmental distances between syllables to also be between 0 and 1, then the coefficient for segmental distance would be greater than tonal distance for both monosyllables and disyllables.

COMPARISON OF DISTANCE METRICS. The second objective is to identify the distance metrics that can best reflect native speakers' phonological distance judgments. The full model was fit to all of logically possible combinations of segmental and tonal distance measures this study considers (see Section 2). The WAIC values were computed and compared for each of these models. Again, the lower the WAIC values, the better the model matches the judgment data.

First, the results for monosyllables are in Table 3. The lowest WAIC values, indicating the best model fit, were achieved with the Hamming distance between multivalued features for segments with the onset-contour-offset representation-based measure for tones (4682.1 in bold face in Table 3). When segmental distance itself is considered, the models with multivalued features using the Hamming distance measure consistently showed the best performance (horizontal grey highlights in Table 3). On the tonal side, the models assuming the representations with contour information (i.e. onset-contour, onset-contour-offset and contour-offset representations) consistently performed best (vertical grey highlights in Table 3).

	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.

Second, the results for disyllables are in Table 4. As shown, the Hamming distance between multivalued features for segments with the contour-offset representation-based measure for tones performed best (7153.0 in bold face in Table 4). This result is similar with that of monosyllables. As to the segmental distance itself, the general tendency is same with monosyllables: The multivalued feature representations were best, especially with the Hamming distance (grey horizontal highlights in Table 4). Of the tonal distances, the contour-offset representation performed well, although the WAIC values were not substantially lower than other models. This result differs from that of monosyllables.

	Chao (H)	Chao (M)	Chao (E)	Autosegmental	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 disyllable model using different segmental and tonal distances without information gain weighting.

Note that the results for segmental distance measures are consistent for monosyllables and disyllables, but they are different for tonal distance: The models with tonal representations with contour information performed best for monosyllables but their performance was not significantly different from other models for disyllables. We hypothesized that this is because the (onset)-contour-(offset) representation in our modeling of disyllables overlooked the change in pitch level across the two syllables. We thus created several extensions of the tonal representations for disyllables. 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 intersyllable pitch-level change, then attached this to the onset-contour-offset representation. In the second type (avg O-C-O+: type 2 in Table 5), we took the ‘average’ pitch of the onset and offset of the two syllables, with extra 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). Take, for example, the tone sequence 1-2. Their two O-C-O representations are HLH and MRH. In O-C-O+ (type 1), the intersyllable pitch-level change would be falling since H is higher than M. In O-C-O+ (type 2), the 'average' pitches of the onset and offset are 4 and 3.5 respectively, so the pitch level change is still falling. In the C-O+ (type 3), the two offsets are H and H, so the pitch-level change is level.

As shown in Table 5, type 1 (O-C-O+) did not result in much improvement, while type 2 (avg O-C-O+) resulted in much lower WAICs than the original onset-contour-offset representation. Type 3 (C-O+) also resulted in much lower WAICs than the original contour-offset representation, resulting in one of the best models (bold faced in Table 5). Based on this observation, we conclude that for disyllables, the best distance metric to predict distance judgments involved the Hamming distance between the 'modified' contour-offset representation of the tones reflecting the change in pitch level between the two syllables 'as a whole' (as in type 2 and type 3 in Table 5), but not simply between offset of a preceding syllable and the onset of the following syllable (type 1 in Table 5).

	O-C-O	O-C-O+ (type 1)	avg 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**

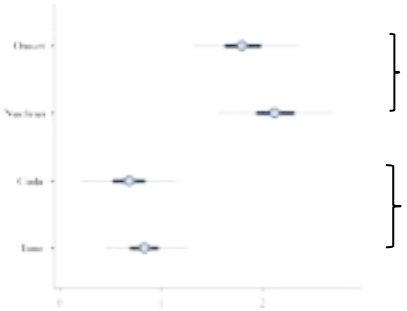
TABLE 5. WAIC values of the monosyllable model using different segmental and tonal distances without information gain weighting, using newly developed tonal representations.

To further compare the results with purely acoustic-based distance measures, we fitted three additional models, namely models based on cochleagrams, MEL FREQUENCY and formant tracks. First, a model that directly calculates acoustic distance from the audio recordings was built. Acoustic distance was calculated by obtaining cochleagrams of each of the recordings using Praat with the default parameters, from which the Euclidean distances between the cochleagrams were calculated. The problem of different numbers of samples was resolved similarly to the method described in Heeringa (2004).[13] This model performed far worse than any of the phonological models in Table 3, at WAIC value 5070.2. Second, a similar calculation

was conducted using Mel frequency cepstral coefficients (Rabiner & Juang 1993), and the model performed even worse at WAIC 5097.3. Third, a model with formant tracks (Heeringa et al. 2009) were the worst at WAIC 5120.5. The results were similar for disyllables: Models assuming purely acoustic-based distance measures performed far worse, with a WAIC of 7510.5 for cochleagrams, 7510.3 for Mel frequency cepstral coefficients and 7628.7 for formant tracks. See Appendix B for the details.

After applying information gain weighting to both the segmental and tonal distances, both classic (Nerbonne & Heeringa 1997) and modified (Broe 1996) versions, the results did not substantially improve. This is consistent with Nerbonne and Heeringa's (1997) results. Adding 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 Table B-5.1–B5.2 in Appendix B-5 for monosyllables and in Table B-10.1–B-10.2 in Appendix B-10 for disyllables.

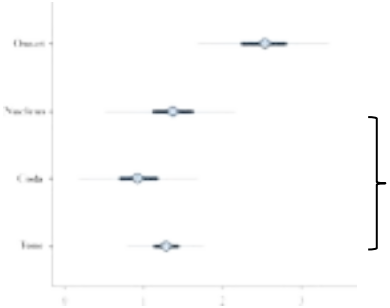
RELATIVE CONTRIBUTION OF SYLLABLE COMPONENTS. The third objective is to determine relative contributions of onset, nucleus, coda, and tone to phonological distance judgments. For this, we separated the segmental distance into onset, nucleus, and coda distances then fitted the full model. First, the results for monosyllables are in Figure 5 with the coefficient estimates, SE, and 95% credible intervals of onset, nucleus, coda, and tone. As shown, onset and nucleus are weighted significantly more than coda and tone, suggesting more crucial role of onset and nucleus than coda and tone in judging phonological distance. The differences between onset and nucleus and between coda and tone were estimated at −0.32 (SE: 0.4, 95% CI: (−1.09, 0.45) and −0.15 (SE: 0.34, 95% CI: (−0.84, 0.49)) respectively, both including zero in their credible intervals. This result suggests no significant difference between the role of onset and nucleus or between coda and tone. However, nucleus was weighted significantly higher than coda, with an estimated difference of 1.44 and 95% credible intervals excluding zero (SE: 0.4, 95% CI: (0.67, 2.25)).[14] The results suggest that the hierarchy of syllabic components' contributions is onset, nucleus > coda, tone when making phonological distance judgements of monosyllables.



	Estimates	SE	95% Credible interval
Onset	1.80	0.27	(1.30, 2.36)
Nucleus	2.12	0.29	(1.30, 2.36)
Coda	0.68	0.68	(0.20, 1.16)
Tone	0.84	0.84	(0.20, 1.16)

FIGURE 5. Estimates of the weightings of onset, nucleus, coda and tone along with 95% and 50% credible intervals.

Second, for the model for disyllables, we assumed equal weighting of two syllables within an item; onset, nucleus, and coda in both syllables were treated equally. The results are in Figure 6 with the coefficient estimates, SE, and 95% credible intervals of onset, nucleus, coda, and tone. As shown, onset is weighted significantly more than nucleus, coda, and tone, suggesting the most important role of onset. Based on posterior draws, the differences between onset and nucleus, nucleus and coda, and coda and tone weighting were 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, with the 95% credible intervals all include zero. This suggests that there is no strong evidence that the nucleus, coda, and tone differ in weighting. However, we do have weak evidence that onset is weighted heavier than nucleus, since the 95% credible interval for their difference excludes zero, (0.02, 2.26).[15] The results suggest that the hierarchy of syllabic components' contributions is onset > nucleus, coda, tone when making phonological distance judgments of disyllables.



	Estimates	SE	95% Credible interval
Onset	2.53	0.43	(1.68, 3.36)
Nucleus	1.38	0.41	(0.51, 2.18)
Coda	0.68	0.38	(0.18, 1.70)
Tone	1.29	0.25	(0.78, 1.78)

FIGURE 6. Estimates of the weightings of onset, nucleus, coda and tone along with 95% and 50% credible intervals.

The results of native speaker distance judgments, comparing monosyllables and disyllables, are summarized in Table 6.

	Monosyllables	Disyllables
Segmental distance vs. Tonal distance	Segmental > Tonal	Segmental ≈ Tonal
Best distance metrics	Multivalued features (Hamming) for segments + Contour tone	Multivalued features (Hamming) for segments + Modified contour tone
Syllabic components' contribution	Onset, Nucleus > Coda, Tone	Onset > Nucleus, Coda, Tone

TABLE 6. Distance judgments data of monosyllables and disyllables.

3.3. DISCUSSION.

TONAL AND SEGMENTAL DISTANCES. We compared relative contributions of segmental and tonal distances to phonological distance judgments in Cantonese. We provided evidence that segmental distance is more crucial than tonal distance in making distance judgments of monosyllabic items in Cantonese. The results echo those of perception studies. Among studies investigating Cantonese, Cham (2003) compared Cantonese speakers' perception of Thai segments and tones by phonological awareness tests where participants selected an odd one from among three syllables. Cham found that Cantonese speakers performed better in segment awareness tasks than in tone awareness tasks, implying that tones are perceptually less salient than segments for Cantonese speakers. The current results, however, contrast with Yang and Castro's (2008) findings that segments were equally important as tones in Zhuang and less important in Bai. Considering that Yang and Castro's main focus was phonological distance measures, it is worth considering the potential sources of the differences between their study and our own. The contrasting results could be due to differences in the task performed (direct distance judgments vs. mutual intelligibility). Or, there may exist typological difference in the relative contribution of segments and tones, which needs future research on cross-linguistic comparisons. Note though that the results of disyllables did not support those of monosyllables in our study; segmental distance did not contribute more than tonal distance in making phonological distance judgments. We want to point out that we do not have strong evidence to the contrary either, as shown by no overlap of their credible intervals. In other words, the results from disyllables are less clear. Such unclear pattern among disyllables can be attributed to the fact that the disyllabic test items may be less representative of the lexicon than monosyllables. Recall that our test included the same number of monosyllables (n = 72) and disyllables (n = 72). 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.

Another thing to notice is that some participants predominantly responded with maximal phonological distances both for monosyllables and for disyllables (e.g. participants 12, 14, 17, 18, and 19 in Figures 3 and 4). Different sources of gradient judgments have been suggested, including performance factors (see review in Schütze 1996) and speakers' internalized linguistic knowledge (Hayes 2000), but regardless of the claims, previous studies consistently support gradience in speakers' judgments. This raises a question on the high frequency of maximal distances observed from this study. In the current study, the judged distance is based on two distance sources, namely segmental and tonal distances. Therefore, a perceived difference between two items reflects the combination of two distances, which, in principle, can more frequently result in maximal distances compared to pure segment-based distance or tone-based distance judgments.

Finally, note that an overarching assumption of our study was that tone is considered separate from segments, hence tonal and segmental distances are computed independently as inputs to the final phonological distance judgments. It is possible to assume that the tone is tied to the nucleus instead. However, even if we consider nucleus-tone ties, the effect of nucleus and tone would still be additive, as far as the distance between nucleus-tone combinations is determined using the usual Levenshtein distance. Therefore, the result would be similar to the current model except nucleus is forced to be weighted same as each element of the tone. For example, the distance between uHL and aMF would still be the segmental distance between [u] and [a] summed up with tonal distances between H and M and between L and F.

METRIC COMPARISONS. For segmental distances, we have demonstrated that multivalued features are better representations of phonemes for predicting distance judgments than binary distinctive features. It was also found that purely acoustic-based distance measures are far worse than any models based on abstract phonological features. This result can be interpreted in two ways. First, one might speculate that abstract phonological knowledge must be at play in making phonological distance judgments. This interpretation aligns with conclusions drawn by previous studies like Somers (1998) and Heeringa (2004). 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 to predict the observed distance judgments. Unlike the binary features, the multivalued features distinguish between allophones and allows for gradient features, but at the

same time do not take into account minor, nonsystematic phonetic detail as the cochleagrams, Mel frequency, or formant tracks do.

For tonal distances, we showed that representations with a contour component worked best for both monosyllables and disyllables. This implies that tone contours are important for phonological distance judgement in Cantonese, consistent with the results from the investigations of other tone languages by Yang and Castro (2008) and Tang and van Heuven (2011). This also aligns with work in tone perception in Cantonese, where tonal directions are found to be an important perceptual cue (e.g. Khouw & Ciocca 2007, Xu et al. 2006, inter alia), which is sometimes more important than tonal height (Gandour 1981).[16]

We have also shown that the information gain weighting did not help improving models' predictions for any types of distance metrics. This is consistent with the results from Nerbonne and Herringa (1997), which show distances between multivalued features without information gain weighting work best for determining dialect distance. We want to note that the lack of effectiveness of information gain weighting does not necessarily imply that the features are equally weighted, because information gain is one possible type of weighting scheme and potentially other schemes might improve the predictive power of phonological distance judgments. We leave this for future research.

RELATIVE CONTRIBUTIONS OF ONSET, NUCLEUS, AND CODA. We further split segments into onset, nucleus, coda, and tone to investigate relative contributions of syllable components to phonological distance judgments. For monosyllables, we have shown that onset and nucleus are more crucial than coda and tone. The fact that onset and nucleus are more important than tone may align with previous tonal perception studies, which showed that spoken word recognition is more challenging when tone differences are involved (e.g. Cutler & Chen 1997, Keung & Hoosain 1979), suggesting lower perceptual sensitivity to tone differences than segmental differences. For disyllables, nucleus shows less important role, contrary to its highest contribution to distance judgments for monosyllables. For monosyllables, the nucleus is the 'central' part of the word, while its role is weakened in a disyllabic word due to an additional transitional property incurred between syllables. Similarly, vowels are more important in monosyllables because of their acoustic prominence while their saliency weakens in disyllables. Also, in the monosyllabic conditions, participants may not process the stimuli as actual words, as most Cantonese monosyllables are bound morphemes that need to appear with other syllables to

form polysyllabic words. If so, acoustic properties become an even more decisive factor in monosyllables. In contrast, since at least one of the stimuli in each disyllable-disyllable pair was always an existing lexical word, the provided context may have weakened the 'vowel advantage'. This idea is consistent with the results from Ye and Connine's (1999) perceptual experiment, where the presence of context removes the 'vowel advantage'. Note though that a similar vein of research in the word reconstruction paradigm (Cutler et al. 2000, van Ooijen 1996) found that vowels are more mutable than consonants, contrary to some previous results on a tone language (Wiener & Turnbull 2016) and our study. Considering that the word reconstruction paradigm necessarily involves lexical access, it may be the case that the acoustic prominence of vowels is overridden by the lexical knowledge. The considerations as such allow us to speculate why the role of nucleus differs for monosyllables and for disyllables, but we still cannot account for the full hierarchy of syllabic components' contributions in making phonological distance judgments.

4. PHONOLOGICAL DISTANCE AND LEXICAL PREDICTABILITY. The aim of Section 4 is to find out why speakers rely more on certain syllabic components than others when making phonological distance judgments. For example, why do Cantonese speakers rely more on onset than coda when judging phonological distance between two items? We hypothesize that the relative contributions of syllabic components observed in the phonological distance judgment test (i.e. onset, nucleus > coda, tone for monosyllables and onset > nucleus, coda, tone for disyllables) is due to their lexical predictability; the more predictable a syllabic component is in the lexicon, the less important it becomes in determining phonological distance. The idea behind this hypothesis is that phonological distance is fundamentally relevant to distinguishing between lexical items, so thus speakers may not rely heavily on lexically highly predictable elements when evaluating phonological distance. For example, if coda is always nasals, thus is easily predictable, the phonological properties of coda do not contribute much to judging how different two items are. Instead, speakers will become more sensitive to lexically uncertain parts (e.g. onset) when judging items' phonological distance. This idea aligns with previous work in semantics where information content has been used in evaluating semantic distances (Budanitsky & Hirst 2001, Jiang & Conrath 1997, Resnik 1995). Through a lexical analysis, this section employs two types of information-theoretic measures of syllabic components to analyze their

lexical predictabilities, namely entropy and functional load. The results show correspondences between the predictions from the lexical analysis and the relative contributions of the syllabic components reported from the phonological distance judgment test in Section 3.

4.1. ENTROPY ANALYSIS. A simple way of measuring the amount of predictability is entropy. Roughly speaking, entropy is the quantity representing 'fuzziness' or 'the lack of predictability'. When calculated using base 2 logarithms, the formula for entropy is as follows.

(5) $-\sum_{i=1}^{n} p_i \log_2 p_i$

Where p_i is the probability of the *i*-th possible outcome and *n* is the total number of possible outcomes of a random variable. In the formula 5, the entropy is a lower bound on the expected number of 'bits', that is, representation in terms of '1's and '0's, that are needed to encode information. As an example, let us compare two toy languages with the following probability distributions of nuclei.

(6) Language A: /a/ 50%, /u/ 25%, /i/ 25%

Language B: /a/ 50%, /u/ 25%, /i/ 12.5%, /o/ 12.5%

Given the number of nuclei and their probabilities, nuclei are overall more predictable in Language A than Language B. For Language A, therefore, the entropy should be relatively low. When the formula in 5 is applied to Language A, the entropy is $-0.5\log_2 0.5 - 0.25\log_2 0.25 - 0.25\log_2 0.25 = 1.5$. Nuclei in Language A in binary digits are encoded as '0' for /a/, '10' for /u/ and '11' for /i/; in this case the expected number of bits is $0.5 \times 1 + 0.25 \times 2 + 0.25 \times 2 = 1.5$, matching the entropy. For Language B, the entropy should be higher, meaning nuclei are less predictable. When the formula in 5 is applied, the entropy is $-0.5\log_2 0.5 - 0.25\log_2 0.25 - 0.125\log_2 0.125 = 1.75$. Nuclei in binary digits are encoded as '0' for /a/, '10' for /u/, '110' for /i/ and '111' for /o/; in this case the expected number of bits is $0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75$, again matching the entropy.

We use entropy to predict syllabic components' relative contribution in phonological distance judgements. The weights of syllabic components in the monosyllable and disyllable models are plotted below against estimated sample entropies. Recall that the hierarchy of contributions in the distance models was onset, nucleus > coda, tone for monosyllables and it was onset > nucleus, coda, tone for disyllables. As plotted on the *x*-axis in Figure 7, the overall entropy hierarchy is onset > nucleus > coda > tone for both monosyllables and disyllables, with onset showing much higher entropy than others. The overall relationship between syllabic

components' roles in distance judgments and their entropies seems quite strong for disyllables, but nucleus is an outlier in the monosyllable case. Specifically, the role of nucleus was very crucial in the distance judgments test, while its entropy is relatively low meaning that the lexical properties of nucleus are relatively well predictable.

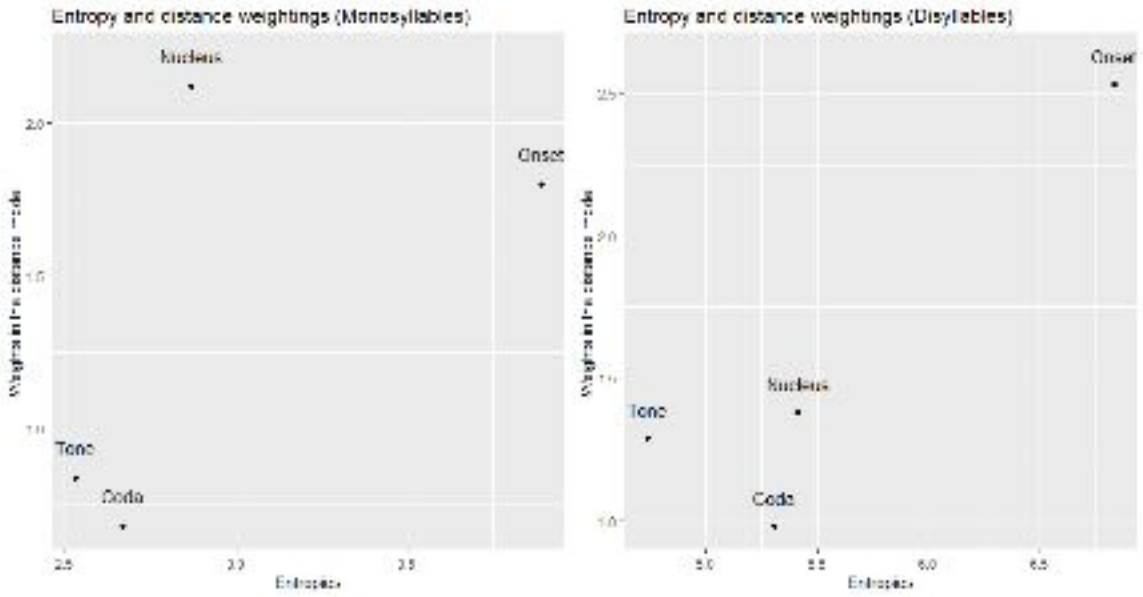


FIGURE 7. Relationship between point estimates of entropy and weighting in distance models, showing their rough correspondence.

It is necessary to check whether the above entropy differences correspond to their actual differences or they are just artefacts of our sample. Thus, we computed confidence intervals[17] for the entropy differences to ensure that the results in Figure 7 are meaningful differences and not simply due to sampling error. Since no standard formula is available for confidence intervals of differences between the simple entropy measures, we derived our own using the asymptotic properties of the probability estimates along with the delta method; details are given in Appendix C. In Figure 8, two types of estimates are reported; confidence interval and point estimates, with the point estimates being in the middle of confidence intervals. As shown, the 95% confidence intervals are all very narrow with the lower bounds far away from zero in most cases. This indicates that the entropy differences among the four syllabic components are very significant, justifying the observed entropy hierarchy of onset > nucleus > coda > tone. Based on this observation, we argue that the overall correspondences between entropy measures and phonological distance judgments are meaningful.

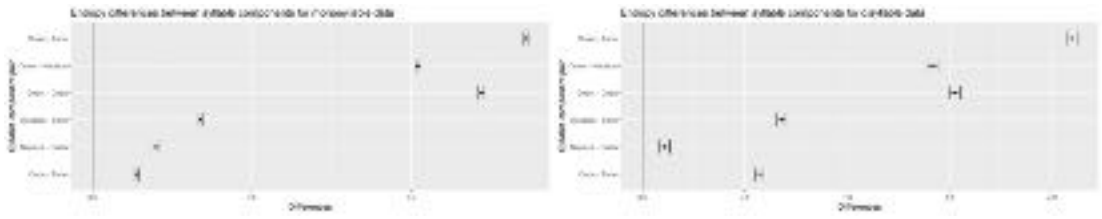


FIGURE 8. Point and interval estimates of the differences between the entropies of various syllable components. Point estimates are represented as circle dots whereas the two limits of the confidence intervals are indicated by short vertical lines.

4.2. FUNCTIONAL LOAD ANALYSIS. The above calculations of entropies do not take into account properties of the other syllabic components when calculating each of their entropy. For example, the lexical properties of onset, nucleus, and tone were ignored when calculating the entropy of coda. This may not be desirable due to phonotactics. If two syllabic components are highly dependent, say we can fully predict tone from coda, then even if there is a huge uncertainty of tone itself, tone becomes less important in making distance judgments because cues from coda enables us to determine tone. An information-theoretic measure that takes this dependency consideration into account is functional load, that is, how important each component is in maintaining contrasts in the language as a whole. The functional load of a component c is computed by comparing the entropy $H(L)$ of the entire language L to the entropy $H(L'_c)$ of a hypothetical language state L'_c where all contrasts in that component are completely neutralized (Carter 1987, Hockett 1966, Oh et al. 2015, Surendran & Levow 2004).

(7) $FL_c(L) = \frac{H(L)-H(L'_C)}{H(L)}$

Using the formula in 7, we computed functional loads for onset, nucleus, coda and tone, and plotted the weights of each syllabic components in the monosyllabic and disyllabic words' models against their functional loads. The results are in Figure 9. As plotted on the *x*-axis, the functional load hierarchy is onset > tone > nucleus > coda for both monosyllables and disyllables. When this functional load hierarchy is compared with the order of the syllabic components' contributions in the distance judgments (onset, nucleus > coda, tone for monosyllables; onset > nucleus, coda, tone for disyllables), they are roughly corresponding except for tone. Specifically, functional loads of tone are higher than those of nucleus and coda (onset > tone > nucleus > coda), while the contribution of tone in the distance judgments was relatively minor. The prediction from functional load for tone is also in contrast with simple

entropy calculations, where the predicted entropy hierarchy was onset > nucleus > coda > tone. This can be because nucleus and coda have more co-occurrence restrictions in Cantonese, and therefore neutralizing one and not the other will have less of an effect on the language, leading to higher functional load, whereas simple entropy calculation only looks at each individual component, and are therefore not affected by such phonotactic factors. Importantly, except for tone, the results again roughly match our phonological distance judgment data.

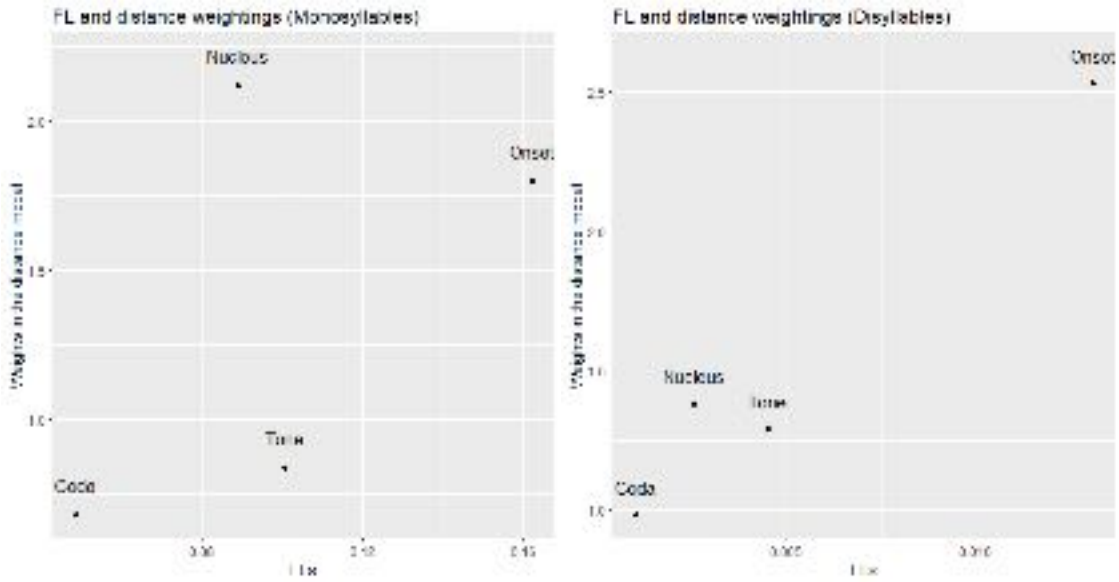


FIGURE 9. Relationship between point estimates of functional load and weighting in distance models, showing a correspondence.

To check the reliability of the results in Figure 9, we calculated confidence intervals for the differences between the functional loads. Note that all but the difference between nucleus and coda in disyllables do not cover zero, suggesting meaningful evidence overall except for one (nucleus-coda differences in disyllables). In other words, the functional load hierarchy remain as onset > tone > nucleus > coda for monosyllables but it should be modified as onset > tone > nucleus, coda for disyllables. Note that the confidence intervals are very narrow among monosyllables but not among disyllables. This indicates that we have very strong evidence for the entropy differences among monosyllables but the evidence is weaker for disyllables. Based on this observation, we argue that the overall correspondences between functional load measures and phonological distance judgments are meaningful especially among monosyllables while their relation is weaker for disyllables.

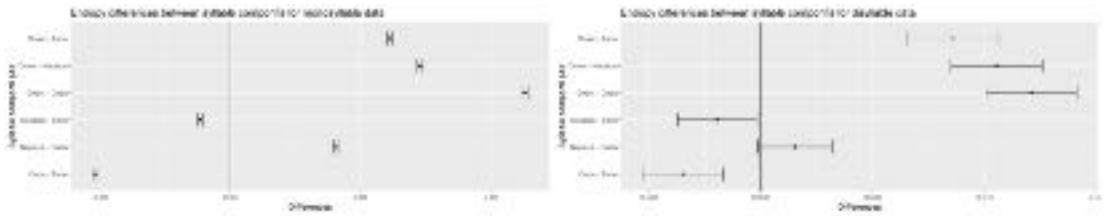


FIGURE 10. Point and interval estimates of the differences between the functional loads of various syllable components. Point estimates are represented as circle dots whereas the two limits of the confidence intervals are indicated by short vertical lines.

From the examinations of simple entropies, we would expect the weight hierarchy of onset > nucleus > coda > tone for both monosyllables and disyllables. From the examination of functional loads, we would expect the weight hierarchy of onset > tone > nucleus > coda for monosyllables, and onset > tone > nucleus, coda for disyllables. Considering that our phonological distance modeling results overall match the predictions from entropies and functional loads, we conclude that measures of lexical predictability have a partial power to account for the weightings of syllabic components in phonological distance measures, although it cannot predict the full range of speakers' phonological distance judgments.

5. DISCUSSION AND CONCLUSION. This study showed the relative contributions of segmental and tonal distances when making phonological distance judgments in Cantonese. It further showed that onset is consistently contributing more than coda and tone (though the role of nucleus is relatively unclear) to phonological distance judgments, and that these results are partially explained by information-theoretical quantities deduced from lexical frequencies. We have also shown that the distance measures for Cantonese that best match native speakers' judgments are based on multivalued, phonetically-based (but not purely phonetic) segmental representations and tonal representations that incorporate information on contours, both within and between syllables.

Beyond its implications to the nature of phonological distance in tone languages, our modelling work has shown how to set up and find optimal measures of phonological distance that can best predict native speakers' judgments. This was done by choosing empirically best-supported distance measure (e.g. in our case the multivalued features), by empirically determining weights for different components of a syllable, and by incorporating random effects to allow for individual variation. Models of language cognition that depend on such measures

can thus be potentially improved by incorporating these insights. The experimental and simulation results in the current paper are from a case study of Cantonese but we believe that our study provides sufficient methodological groundwork to investigate phonological distance measures in other tone languages. Even for tone languages with complex tonal processes, such as complicated tone alternations, we believe our methodology is still applicable as far as tonal representations at a surface-level are correctly identified. This is because phonological distance measures here are mainly about surface representations of segments and tones, not directly related to the processes involved in deriving surface phonemes or tones from their underlying representations. We also believe that this study can open doors to wider explorations of neighborhood models incorporating tonal features, since good neighborhood models can be built only with solid phonological distance measurement methods. Ultimately, the methods presented in this paper should allow for better modelling of phonotactics, speech errors, spoken word recognition and other aspects of phonological cognition in tone languages, which has been relatively overlooked in the current literature.

Appendix A

A-1. SEGMENTAL REPRESENTATIONS.

The distinctive binary values for Cantonese phonemes are presented in Table A-1.1. The exact number of phonemes may be debatable; we assume that each symbol in Jyutping, a standard phonological transcription system for Cantonese (Tang et al. 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 A-1.1. Distinctive binary values for Cantonese phonemes.

The multivalued phonological features for Cantonese segments are presented in Table A-1.2. 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 (Bauer & Benedict 1997, Clumeck et al. 1981, Tse 2005, Zee 1999). For example, for voicing, we follow Ladefoged 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
[p͡]	60	0	0	100	100	100	0	0	0	0	100	49	0	50
[t͡]	60	0	0	85	5	100	0	0	0	20	100	49	0	50
[k͡]	60	0	0	60	5	100	0	0	0	0	100	49	0	30

TABLE A-1.2. Multivalued values for Cantonese phones.

A-2. 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 **S** be a random vector of features with components $\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_I}$, where each component represents a feature. Each possible value of **S**, denoted s_i, is thus a phoneme. Suppose there are *J* phonemes in the language. Then the entropy of **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 **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 **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) $H(\mathbf{S}|\mathbf{f_i} = v) = -\sum_{j=1,\, \mathbf{f_i}=v \text{ when } \mathbf{S}=s_j}^{J} P(\mathbf{S} = s_j|\mathbf{f_i} = v) \log\big(P(\mathbf{S} = s_j|\mathbf{f_i} = v)\big)$

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 and 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.[2]

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

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 (2006:20–21). A version of the proof, adopted to our current setting, is as follows.

$$\begin{aligned}(6)\quad IG(\mathbf{f_i}) &= -\sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j)\right) - \sum_{v \in V} H(\mathbf{S} \mid \mathbf{f_i} = v) P(\mathbf{f_i} = v) \\ &= -\sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j)\right) \\ &\qquad + \sum_{v \in V} \sum_{j=1,\, \mathbf{f_i} = v \text{ when } \mathbf{S} = s_j}^{J} P(\mathbf{S} = s_j \mid \mathbf{f_i} = v) \log\left(P(\mathbf{S} = s_j \mid \mathbf{f_i} = v)\right) P(\mathbf{f_i} = v) \\ &= -\sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j)\right) \\ &\qquad + \sum_{v \in V} \sum_{j=1,\, \mathbf{f_i} = v \text{ when } \mathbf{S} = s_j}^{J} P(\mathbf{S} = s_j) \log\left(P(\mathbf{S} = s_j \mid \mathbf{f_i} = v)\right) \\ &= \sum_{v \in V} \sum_{j=1,\, \mathbf{f_i} = v \text{ when } \mathbf{S} = s_j}^{J} P(\mathbf{S} = s_j) \left[\log\left(\frac{P(\mathbf{S} = s_j \cap f_i = v_i)}{P(f_i = v_i)}\right) - \log\left(P(\mathbf{S} = s_j)\right)\right] \\ &= \sum_{j=1}^{J} P(\mathbf{S} = s_j) \log\left(\frac{1}{P(f_i = v_j)}\right)\end{aligned}$$

The second last line is equal to the last line because the outer sum in 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

[2] 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 $P(\mathbf{S} = s_j \cap f_i = v_i) = P(\mathbf{S} = s_j)$.

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 and Heeringa's information gain weighting to account for the fact that certain features may have null values. Let $V' = V \backslash \{0\}$, that is, 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).

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

Thus only the phonemes for which the feature is nonnull 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 $0 + 1 + 1 = 2$.

A disadvantage of using the above discrete formulas, for the case of multivalued features, was that the numerical values of the features were ignored. To resolve this issue, we also considered calculating information gain for the Euclidean and Manhattan distances between multivalued features. Analogous to the above formulas, the continuous information gain was

$$(8) \quad IG(\mathbf{f_i}) = \int_{\mathbb{R}^J} f_{\mathbf{S}}(f_1 = v_1, \ldots, f_J = v_J) \log\left(\frac{1}{f_{f_i}(f_i=v_j)}\right) \mathrm{d}v_1 \ldots \mathrm{d}v_j$$

The question then becomes how $f_{\mathbf{S}}$, the joint probability density function of the features, is calculated. We estimated this by fitting a GAUSSIAN MIXTURE MODEL to the features in the R package mclust (Fraley et al. 2012). We then evaluated the integral using MONTE CARLO INTEGRATION. Denoting the joint density function of the *i*-th component Gaussian as $f_{\mathbf{S},i}$, for each Gaussian we simulated 1000 fake data from $f_{\mathbf{S},i}$ (Robert & Casella 2010), calculated $\log\left(\frac{1}{f_{f_i}(f_i=v_j)}\right)$ for each and averaged over them. We then obtained the entire integral be multiplying the average for each Gaussian by the probability of that Gaussian.

A-3. 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 A-3.1. 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 A-3.2. 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 A-3.3. 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 A-3.4. 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 A-3.5. 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 A-3.6. 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 A-3.7. Hamming distances between contour-offset representations.

Appendix B

B-1. TABLES OF STIMULI.

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 B-1.1. 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 B-1.2. The table of stimuli for disyllables.

B-2. 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 judgments 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.[3]

The intercept as well as the coefficients indicating the effects of tonal and segmental distance are subject-level effects in the full model, that is, 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).[4] 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 judgments are assumed to be right-censored. The third column indicates which subject-level effects are present, that is, which parameters are assumed to vary by the subject. The fourth column indicates whether the intercept may vary by

[3] 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 and 4 in the draft seem to favour the censoring approach.

[4] Following conventions in the Bayesian paradigm (Gelman & Hill 2007:2–3 225), 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.

item. The full model, along with reduced models of various forms, are reported in Table B-2.1 with their Watanabe-Akaike information criterion (WAIC).

Model	Censor	Subject-level effects present	Item-level	WAIC	10-fold CV IC
I	No			7273.6	7273.9
II	No	Intercept		6778.6	6776.5
III	No	Intercept	Intercept	5360.6	5372.4
IV	No	Intercept, segdist	Intercept	4980.4	5010.3
V	No	Intercept, tonedist	Intercept	5344.4	5391.1
VI	No	Intercept, segdist, tonedist	Intercept	4948.0	4972.7
VII	Yes			7113.6	7112.8
VIII	Yes	Intercept		6289.5	6310.0
IX	Yes	Intercept	Intercept	5119.4	5129.6
X	Yes	Intercept, segdist	Intercept	4817.6	4841.8
XI	Yes	Intercept, tonedist	Intercept	5063.6	5078.1
XII (Full)	Yes	Intercept, segdist, tonedist	Intercept	4727.1	4757.9

TABLE B-2.1. 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 judgments. 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 judgments are right-censored.

B-3. PARAMETER ESTIMATIONS IN THE MONOSYLLABLE MODEL.

We now present the model parameters of the monosyllable model in graphical form. Figure B-3.1 displays the point estimate of each parameter on the *x*-axis along with its 50% (dark blue) and 95% (light blue) uncertainty intervals.[5]

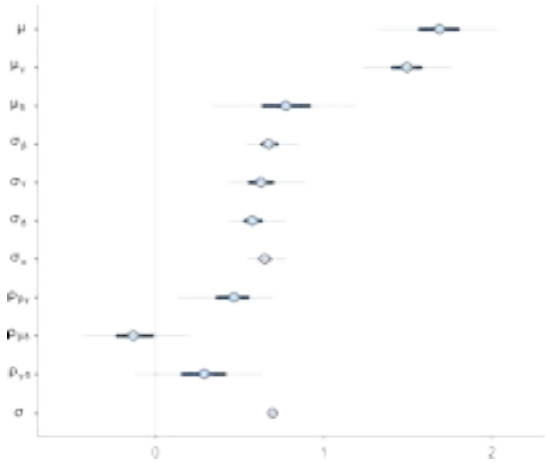


FIGURE B-3.1. Estimates of the model parameters along with 95% and 50% credible intervals. μ is the overall (population-level) intercept, σ is the residual standard deviation, μ_γ and μ_δ are the mean coefficients of segmental and tonal distance, ρ_{AB} indicate the population correlation between A and 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 intersubject variation in segment and tone weightings is quantified by the

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

subject-level standard deviations σ_γ and σ_δ, which are respectively estimated at 0.61 (SE: 0.10, 95% CI: (0.48, 0.81)) and 0.67 (SE: 0.14, 95% CI: (0.48, 0.81)). Note though 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 was positive, however, it would suggest that people whose judgments are affected more heavily by segments are also affected more heavily by tones in general. Figure B-3.2 plots the estimated tonal weightings against the estimated segmental weightings.

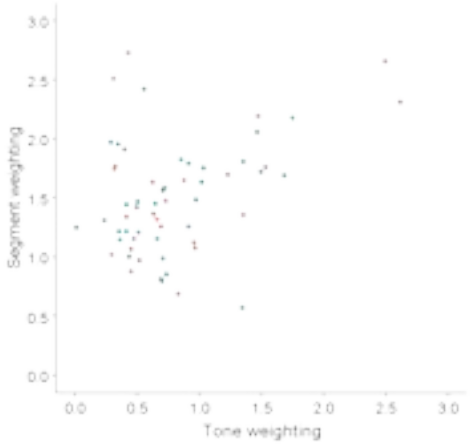


FIGURE B-3.2. A graph showing estimated tonal and segmental weightings of each participant. Blue dots indicate participants who have completed less than 25% of the experiment. Note that we do not have strong evidence of a positive correlation between the two, despite the appearance of the graph.

B-4. CROSS VALIDATION INFORMATION CRITERION VALUES FOR THE UNWEIGHTED MODELS.

To give a more direct measure of out-of-sample predictive power, we have computed 10-fold cross validation information criterion results in addition to the WAIC values supplied in the main text.

	Chao (H)	Chao (M)	Chao (E)	Autosegmental	O-C	O-C-O	C-O
Simple	4808.1	4814.2	4823.4	4831.9	4786.9	4748.5	4765.3
Natural class	4810.1	4860.8	4860.9	4801.1	4860.8	4786.9	4762.0
Binary (H)	4807.7	4849.1	4846.0	4841.9	4786.7	4772.4	4799.8
Multivalued (E)	4771.6	4787.6	4780.9	4791.6	4757.3	4719.6	4731.8
Multivalued (M)	4782.0	4788.9	4817.4	4808.0	4757.1	4728.1	4726.8
Multivalued (H)	4769.6	4793.9	4813.3	4778.5	4741.6	4713.2	4709.9

TABLE B-4.1. 10-fold cross validation information criterion values of the monosyllable model using different segmental and tonal distances.

A value of 5088.1 was obtained for the purely acoustic distance.

B-5. WAIC/CV VALUES WITH INFORMATION GAIN WEIGHTING.

The WAIC values are shown below. They are either comparable or inferior to those without information gain weighting. Note that, when calculating the natural class distances, we adopt a slightly different approach to allow for information-theoretic weighting: The denominator is ALL of the natural classes present in the language, rather than the union of the natural classes to which the two phonemes belong.

	Chao (H)	Chao (M)	Chao (E)	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 (E, discrete)	4735.4	4770.2	4767.4	4767.4	4721.0	4964.9	4692.3
Multivalued (E, continuous)	4736.8	4773.4	4768.2	4766.5	4720.2	4694.3	4694.4
Multivalued (M, discrete)	4757.9	4775.6	4770.7	4771.7	4771.7	4722.1	4699.1
Multivalued (M, continuous)	4740.2	4775.7	4770.7	4771.3	4724.6	4700.4	4699.7
Multivalued (H)	4724.5	4755.0	4751.7	4756.2	4708.3	4685.3	4682.4

TABLE B-5.1. WAIC values of the monosyllable model using different segmental and tonal distances with information gain weighting.

	Chao (H)	Chao (M)	Chao (E)	Autosegmental	O-C	O-C-O	C-O
Natural class	4803.7	4822.9	4841.8	4834.8	4818.6	4790.0	4765.9
Binary (H) (naïve weighting)	4823.4	4877.6	4860.8	4875.7	4804.2	4804.7	4809.2
Binary (H) (Broe weighting)	4845.7	4874.0	4859.6	4850.4	4832.6	4806.8	4792.5
Multivalued (E, discrete)	4749.1	4806.5	4788.4	4797.7	4764.0	4733.6	4742.0
Multivalued (E, continuous)	4771.4	4783.8	4816.5	4833.4	4757.1	4742.1	4695.3
Multivalued (M, discrete)	4773.6	4810.0	4794.4	4786.5	4768.3	4725.8	4758.0
Multivalued (M, continuous)	4777.7	4811.8	4774.3	4820.8	4764.4	4708.6	4709.1
Multivalued (H)	4758.0	4805.7	4780.5	4790.7	4744.6	4731.2	4732.3

TABLE B-5.2. 10-fold cross-validation information criterion values of the monosyllable model using different segmental and tonal distances with information gain weighting.

B-6. Details of the acoustic representations.

Cochleagrams were created using default settings in Praat (time step 0.01s, frequency resolution 0.1 Bark, window length 0.03s, forward-masking time 0.03s). MFCC was extracted with the default values, that is, 12 coefficients, window length 0.015s, time step 0.005s, first filter frequency 100 mel, distance between filters 100 mel, and maximum frequency set at 0 mel. Formants were extracted using To Format (burg) with time step 0,1s, maximum number of formants of 5, maximum formant at 5500 Hz, window length 0.025s, and pre-emphasis 50 Hz.

B-7. Details of 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	10-fold CV IC
I	No			10873.5	10872.8
II	No	Intercept		10402.0	10409.5
III	No	Intercept	Intercept	8547.5	8580.9
IV	No	Intercept, segdist	Intercept	8251.7	8269.9
V	No	Intercept, tonedist	Intercept	8546.8	8523.1
VI	No	Intercept, segdist, tonedist	Intercept	8221.5	8253.2
VII	Yes			9816.5	9815.2
VIII	Yes	Intercept		9028.1	9032.4
IX	Yes	Intercept	Intercept	7633.2	7660.7
X	Yes	Intercept, segdist	Intercept	7319.1	7354.4
XI	Yes	Intercept, tonedist	Intercept	7563.1	7577.2
XII (Full)	Yes	Intercept, segdist, tonedist	Intercept	7194.5	7237.3

Table B-7.1. WAIC values of the full disyllable model along with various reduced models.

B-8. Parameter estimations in the disyllable model.

The model parameters of the full model are as follows.

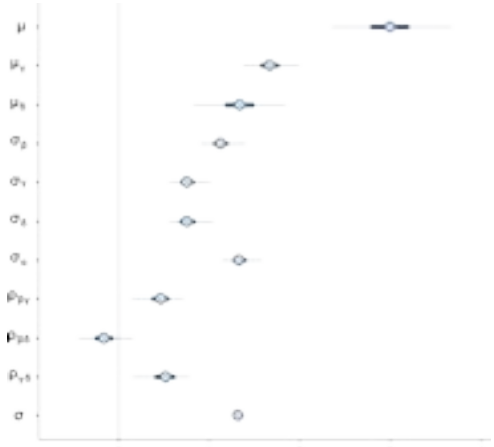


Figure B-8.1. Estimates of the model parameters along with 95% and 50% credible intervals. μ is the overall (population-level) intercept, σ is the residual standard deviation, μ_γ and μ_δ are the mean coefficients of segmental and tonal distance, ρ_{AB} indicate the population correlation between A and 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 intersubject 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, and 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 judgments 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.

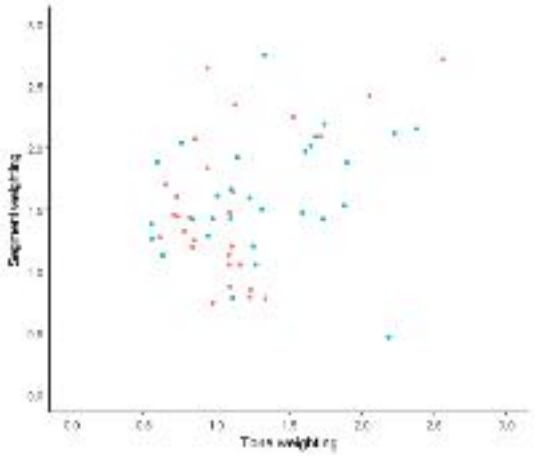


FIGURE B-8.2. A graph showing estimated tonal and segmental weightings of each participant. Blue dots indicate participants who have not completed the experiment.

B-9. CROSS VALIDATION INFORMATION CRITERION VALUES FOR THE UNWEIGHTED MODELS.

	Chao (H)	Chao (M)	Chao (E)	Autosegmental	O-C	O-C-O	C-O
Simple	7208.0	7238.7	7199.3	7240.0	7215.2	7232.2	7172.4
Natural class	7268.6	7273.7	7262.2	7294.4	7263.1	7223.2	7232.5
Binary (H)	7259.7	7255.7	7279.8	7272.9	7206.6	7236.3	7203.5
Multivalued (E)	7174.6	7213.1	7227.5	7270.3	7196.3	7177.7	7196.1
Multivalued (M)	7169.0	7180.7	7227.7	7263.4	7196.7	7194.3	7204.9
Multivalued (H)	7212.9	7192.0	7183.7	7235.1	7220.1	7177.4	7208.7

TABLE B-9.1. 10-fold cross validation information criterion values of the disyllable model using different segmental and tonal distances.

	O-C-O	O-C-O+ (type 1)	O-C-O+ (type 2)	C-O	C-O+ (type 3)
Simple	7232.2	7177.0	7188.6	7172.4	7213.9
Natural class	7223.2	7226.7	7174.8	7232.5	7201.2
Binary (H)	7236.3	7216.9	7204.4	7203.5	7192.9
Multivalued (E)	7177.7	7174.1	7181.4	7196.1	7161.5
Multivalued (M)	7194.3	7200.8	7183.5	7204.9	7181.3
Multivalued (H)	7177.4	7174.9	7167.6	7208.7	7157.4

TABLE B-9.2. 10-fold cross validation information criterion values of the disyllable model using new tonal representations.

B-10. WAIC AND CV VALUES WITH INFORMATION GAIN WEIGHTING.

The WAIC values are shown below. Most of the results are similar or inferior to those without information gain weighting. A few of the results appear to have improved (e.g. with multivalued segmental representations along with Manhattan distances between Chao tone letters) but there were no improvements based on information gain across the board.

	Chao (H)	Chao (M)	Chao (E)	Autosegmental	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, discrete)	7188.9	7158.3	7170.3	7219.5	7172.7	7158.5	7151.9.
Multivalued (E, continuous)	7186.8	7160.7	7170.5	7219.6	7171.7	7160.3	7149.6
Multivalued (M, discrete)	7198.3	7170.8	7180.1	7229.2	7177.1	7167.9	7157.9
Multivalued (M, continuous)	7192.4	7165.2	7174.1	7220.5	7172.6	7162.7	7154.1
Multivalued (H)	7244.3	7221.1	7227.7	7271.6	7226.7	7217.5	7208.3

TABLE B-10.1. WAIC values of the disyllable model using different segmental and tonal distances with information gain weighting.

	Chao (H)	Chao (M)	Chao (E)	Autosegmental	O-C	O-C-O	C-O
Natural class	7268.6	7273.7	7262.2	7294.4	7263.1	7223.2	7232.5
Binary (H) (naïve weighting)	7264.9	7244.9	7248.8	7288.6	7221.7	7234.3	7233.8
Binary (H) (Broe weighting)	7272.3	7234.7	7229.5	7278.0	7241.6	7252.3	7224.9
Multivalued (E, discrete)	7240.0	7215.6	7213.2	7247.5	7227.6	7184.3	7157.4
Multivalued (E, continuous)	7223.0	7220.7	7191.0	7249.2	7197.1	7168.2	7202.1
Multivalued (M, discrete)	7189.1	7230.3	7203.3	7240.2	7209.4	7191.8	7185.3
Multivalued (M, continuous)	7225.2	7215.1	7215.8	7239.0	7202.4	7190.8	7169.2
Multivalued (H)	7259.2	7200.7	7203.0	7258.4	7182.8	7214.0	7173.4

TABLE B10-2. 10-fold cross-validation information criterion values of the disyllable model using different segmental and tonal distances with information gain weighting.

Appendix C

ENTROPY CALCULATIONS.

We used a method to estimate the sampling distributions of the entropies and functional loads obtained as follows. Since the plugin 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. By the multivariate version of the CENTRAL LIMIT THEOREM and the delta method (Casella & 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.

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

where $\alpha(\theta) = \left[\frac{\partial f(\theta)}{\partial \theta_1} \ldots \frac{\partial f(\theta)}{\partial \theta_k}\right]$ 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 offdiagonal 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.[6]

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 $\log p_{k+1} - \log p_i$, whereas the derivative of the entropy of the modified language with respect to p_i is $\log p^*_{k+1} -$

[6] 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.

$\log p_i^*$, where p_i^* is the probability that a random syllable is the same as the i-th 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.

REFERENCES

BAILEY, TODD M., and ULRIKE HAHN. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44. 568–591.

BARTH, DANIELLE, and VSEVOLOD KAPATSINSKI. 2018. Evaluation logistic mixed-effects models of corpus data. *Mixed-effects regression models in linguistics*, ed. by Dirk Speelman, Kris Heylen, and Dirk Geeraerts, 99–116. Cham: Springer.

BAUER, ROBERT S., and PAUL K. BENEDICT. 1997. *Modern Cantonese phonology*. Berlin; New York: Mouton de Gruyter.

BOERSMA, PAUL, and DAVID WEENINK. 2009. *Praat: Doing phonetics by computer (Version 5.1.05.)* [Computer software]. Amsterdam: Institute of Phonetic Sciences.

BROE, MICHAEL. B. 1996. A generalized information-theoretic measure for systems of phonological classification and recognition. *Computational phonology in speech technology: Proceedings of the second meeting of the ACL special interest group in computational phonology*, ed. by Richard Sproat, 17–24. Somerset, NJ: Association for Computational Linguistics.

BUDANITSKY, ALEXANDER, and GRAEME HIRST. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Paper presented at the Workshop on Wordnet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh.

BÜRKNER, PAUL-CHRISTIAN. 2017a. brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software* 80. 1–28. Online: https://www.jstatsoft.org/index.php/jss/article/view/v080i01.

BÜRKNER, PAUL-CHRISTIAN. 2017b. Advanced Bayesian multilevel modeling with the R package brms. Online: https://arxiv.org/pdf/1705.11123.pdf.

CARPENTER, BOB; ANDREW GELMAN; MATTHEW D. HOFFMAN; DANIEL LEE; BEN GOODRICH; MICHAEL BETANCOURT; MARCUS BRUBAKER; JIQIANG GUO; PETER LI; and ALLEN RIDDELL. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76. 1–32. Online: https://www.jstatsoft.org/article/view/2991.

CARTER, DAVID M. 1987. An information-theoretic analysis of phonetic dictionary access. *Computer Speech & Language* 2. 1–11.

CASELLA, GEORGE, and ROGER L. BERGER. 2002. *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

CHAM, HOI YEE. 2003. *A cross-linguistic study of the development of the perception of lexical tones and phones* (Bachelor dissertation, The University of Hong Kong, Hong Kong). Online: https://core.ac.uk/reader/37886270.

CLUMECK, HAROLD; DAVID BARTON; MARLYS A. MACKEN; and DOROTHY A. HUNTINGTON. 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. 210–225.

COVER, THOMAS. M, and JOY A. THOMAS. 2006. *Elements of information theory* (2nd ed.). Hoboken, N.J.: Wiley-Interscience.

CUTLER, ANNE, and HSUAN-CHIH CHEN. 1997. Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics* 59. 165–179. Online: https://link.springer.com/content/pdf/10.3758/BF03211886.pdf.

CUTLER, ANNE; NURIA SEBASTIÁN-GALLÉS; OLGA SOLER-VILAGELIU; and BRIT VAN OOIJEN. 2000. Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition* 28. 746–755. Online: https://link.springer.com/content/pdf/10.3758/BF03198409.pdf.

DUFOUR, SOPHIE, and JONATHAN GRAINGER. 2019. Phoneme-order encoding during spoken word recognition: A priming investigation. *Cognitive Science* 43. e12785.

ELLISON, T. MARK, and SIMON KIRBY. 2006. Measuring language divergence by intra-lexical comparison. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. 273–280. Online: https://dl.acm.org/doi/pdf/10.3115/1220175.1220210.

FISHER, WILLIAM M., and JONATHAN G. FISCUS. 1993. Better alignment procedures for speech recognition evaluation. *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* 2. 59–62.

FRALEY, CHRIS; ADRIAN E. RAFTERY; THOMAS B. MURPHY; and LUCA SCRUCCA. 2012. mclust Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report No. 597*. Online:

https://pdfs.semanticscholar.org/5bbc/022e371259d39cef9c47f453545a95cc36b2.pdf?_ga=2.79420052.1835117691.1586446706-1251682946.1586446706.

FRISCH, STEFAN A.; MICHAEL BROE; and JANET PIERREHUMBERT. 1997. Similarity and phonotactics in Arabic. *Rutgers Optimality Archive 223*. Online: http://roa.rutgers.edu/files/223-1097/roa-223-frisch-2.pdf.

GANDOUR, JACK. 1981. Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics* 9. 20–36.

GELMAN, ANDREW; JOHN B. CARLIN; HAL STEVEN STERN; DAVID B. DUNSON; AKI VEHTARI; and DONALD B. RUBIN. 2014. *Bayesian data analysis* (3rd ed.). Boca Raton: CRC Press.

GELMAN, ANDREW, and JENNIFER HILL. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

GILDEA, DANIEL, and DANIEL JURAFSKY. 1996. Learning bias and phonological-rule induction. *Computational Linguistics* 22. 497–530. Online: https://dl.acm.org/doi/pdf/10.5555/256329.256335.

GOOSKENS, CHARLOTTE, and WILBERT HEERINGA. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16. 189–207.

HAYES, BRUCE. 2000. Gradient well-formedness in optimality theory. *Optimality theory: Phonology, syntax, and acquisition*, ed. by Joost Dekkers, Frank Reinoud Hugo van der Leeuw, and Jeroen Maarten van de Weijer, 88–120. Oxford; New York: Oxford University Press.

HAYES, BRUCE. 2011. *Introductory phonology*. Hoboken: John Wiley & Sons.

HEERINGA, WILBERT. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. (Doctoral dissertation, University of Groningen, the Netherlands). Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.285&rep=rep1&type=pdf.

HEERINGA, WILBERT; KEITH JOHNSON; and CHARLOTTE GOOSKENS. 2009. Measuring Norwegian dialect distances using acoustic features. *Speech Communication* 51, 167–183.

HEERINGA, WILBERT; PETER KLEIWEG; CHARLOTTE GOOSKENS; and JOHN NERBONNE. 2006. Evaluation of string distance algorithms for dialectology. *Proceedings of the Workshop on Linguistic Distances*. 51–62. Online: https://www.aclweb.org/anthology/W06-1108.pdf.

HOCKETT, CHARLES F. 1966. *The quantification of functional load: A linguistic problem* (RM-5168-PR). Rand Corp. Online: https://files.eric.ed.gov/fulltext/ED011649.pdf.

JIANG, JAY J., and DAVID W. CONRATH. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the 10th Research on Computational Linguistics International Conference*. 19–33. Online: https://arxiv.org/pdf/cmp-lg/9709008.pdf.

JONES, ZACK, and CYNTHIA G. CLOPPER. 2019. Subphonemic variation and lexical processing: Social and stylistic factors. *Phonetica* 76. 163–178.

JURAFSKY, DAN, and JAMES H. MARTIN. 2014. *Speech and language processing* (2nd ed.). Harlow: Pearson Education.

KESSLER, BRETT. 1995. Computational dialectology in Irish Gaelic. *EACL '95: Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, ed. by Steven P. Abney and Erhard W. Hinrichs, 60–66. San Francisco, California: Morgan Kaufmann Publishers. Online: https://dl.acm.org/doi/pdf/10.3115/976973.976983.

KEUNG, TSANG, AND RUMJAHN HOOSAIN. 1979. Segmental phonemes and tonal phonemes in comprehension of Cantonese. *Psychologia: An International Journal of Psychology in the Orient* 22. 222–224.

KHOUW, EDWARD, and VALTER CIOCCA. 2007. Perceptual correlates of Cantonese tones. *Journal of Phonetics* 35. 104–117.

KONDRAK, GRZEGORZ. 2002. Determining recurrent sound correspondences by inducing translation models. *Proceedings of the 19th International Conference on Computational Linguistics* 1. 1–7. Online: https://dl.acm.org/doi/pdf/10.3115/1072228.1072244.

LADEFOGED, PETER. 1975. *A course in phonetics*. Chicago: Harcourt Brace Jovanovich.

LEWANDOWSKI, DANIEL; DOROTA KUROWICKA; and HARRY JOE. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100. 1989–2001. Online: https://doi.org/10.1016/j.jmva.2009.04.008.

LUCE, PAUL A.; STEPHEN D. GOLDINGER; EDWARD T. AUER; and MICHAEL S. VITEVITCH. 2000. Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics* 62. 615–625. Online: https://link.springer.com/content/pdf/10.3758/BF03212113.pdf.

LUCE, PAUL A., and DAVID B. PISONI. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19. 1–36. Online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467695/pdf/nihms-410790.pdf.

LUKE, KANG KWONG, and MAY L.Y. WONG. 2015. The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics Monograph Series* 25. 312–333. Online: http://compling.hss.ntu.edu.sg/hkcancor/data/LukeWong_Hong-Kong-Cantonese-Corpus.pdf.

MALINS, JEFFREY G., and MARC F. JOANISSE. 2010. The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language* 62. 407–420.

MOK, PEGGY P. K.; DONGHUI ZUO; and PEGGY W. Y. WONG. 2013. Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change* 25. 341–370.

NEERGAARD, KARL D., and CHU-REN HUANG. 2016. Graph theoretic approach to Mandarin syllable segmentation. Paper presented at the 15th International Symposium on Chinese Languages and Linguistics (IsCLL-15), Taiwan.

NEERGAARD, KARL D., and CHU-REN HUANG. 2019. Constructing the Mandarin phonological network: Novel syllable inventory used to identify schematic segmentation. *Complexity* 2019. Article ID 6979830. Online: http://downloads.hindawi.com/journals/complexity/2019/6979830.pdf.

NERBONNE, JOHN, and WILBERT HEERINGA. 1997. Measuring dialect distance phonetically. *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*. 11–18. Online: https://www.aclweb.org/anthology/W97-1102.pdf.

NERBONNE, JOHN, and WILBERT HEERINGA. 2001. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica* 2001. 69–83.

NICENBOIM, BRUNO, and SHRAVAN VASISHTH. 2016. Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass* 10. 591–613.

OAKES, MICHAEL P. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter lLanguages. *Journal of Quantitative Linguistics* 7. 233–243.

OH, YOON MI; CHRISTOPHE COUPÉ; EGIDIO MARSICO; and FRANÇOIS PELLEGRINO. 2015. Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics* 53. 153–176.

PIERREHUMBERT, JANET B. 1993. Dissimilarity in the Arabic verbal roots. *Proceedings of the North East Linguistics Society* 23. 367–381. Online: http://www.phon.ox.ac.uk/jpierrehumbert/publications/arabic_roots.pdf.

QUALTRICS. 2018 *Qualtrics*. Provo, Utah, USA. Online: http://www.qualtrics.com.

RABINER, LAWRENCE R., and BIING HWANG JUANG. 1993. *Fundamentals of speech recognition*. Englewood Cliffs, N.J.: PTR Prentice Hall.

RADEAU, MONIQUE; MIREILLE BESSON; ELISABETH FONTENEAU; and SAO LUIS CASTRO. 1998. Semantic, repetition and rime priming between spoken words: Behavioral and electrophysiological evidence. *Biological Psychology* 48. 183–204.

RAO, C. RADHAKRISHNA. 1973. *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

RESNIK, PHILIP. 1995. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Online: https://arxiv.org/pdf/cmp-lg/9511007.pdf.

ROBERT, CHRISTIAN, and GEORGE CASELLA. 2010. *Introducing Monte Carlo methods with R*. New York, NY: Springer.

SAIEGH–HADDAD, ELINOR. 2004. The impact of phonemic and lexical distance on the phonological analysis of words and pseudowords in a diglossic context. *Applied Psycholinguistics* 25. 495–512.

SCHÜTZE, CARSON T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, Ill.: University of Chicago Press.

SOMERS, HAROLD L. 1998. Similarity Metrics for Aligning Children's Articulation Data. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* 2. 1227–1232. Online: https://www.aclweb.org/anthology/P98-2200.pdf.

SURENDRAN, DINOJ, and GINA-ANNE LEVOW. 2004. The functional load of tone in Mandarin is as high as that of vowels. *Speech Prosody 2004, Nara, Japan, March 23-26, 2004*. 99–102. Online: https://www.isca-speech.org/archive/sp2004/papers/sp04_099.pdf.

TANG, CHAOJU. 2009. *Mutual intelligibility of Chinese dialects: An experimental approach* (Doctoral dissertation, Landelijke Onderzoekschool Taalwetenschap (LOT), Utrecht, the Netherlands). Online: https://openaccess.leidenuniv.nl/bitstream/handle/1887/13963/Tang_diss2009_PDF_final2.pdf?sequence=5.

TANG, CHAOJU, and VINCENT VAN HEUVEN. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119. 709–732.

TANG, CHAOJU, and VINCENT VAN HEUVEN. 2011. Tone as a predictor of mutual intelligibility between Chinese dialects. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII): August 17-21, 2011*. 1962–1965. Online: https://openaccess.leidenuniv.nl/bitstream/handle/1887/18161/Tang_Heuven_ICPhS2011.pdf?sequence=1.

TANG, CHAOJU, and VINCENT VAN HEUVEN. 2015. Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics* 53. 285–312.

TANG, SZE-WING; KWOK FAN; THOMAS HUN-TAK LEE; CAESAR LUN; KANG KWONG LUKE; PETER TUNG; and KWAN HIN CHEUNG. 2002. *Guide to LSHK Cantonese romanization of Chinese characters* (2nd ed.). Hong Kong: The Linguistic Society of Hong Kong.

TSE, HOLMAN. 2005. *The phonetics of VOT and tone interaction in Cantonese* (Master thesis, University of Chicago, Chicago, Illinois). Online: http://d-scholarship.pitt.edu/25258/1/tse_h_2005_ma_thesis.pdf.

TURNBULL, RORY, and SHARON PEPERKAMP. 2017. The asymmetric contribution of consonants and vowels to phonological similarity: Evidence from lexical priming. *The Mental Lexicon* 12. 404–430.

USSISHKIN, ADAM, and ANDREW WEDEL. 2002. Neighborhood density and the root-affix distinction. *Proceedings of NELS* 32. 539–549.

VAN OOIJEN, BRIT. 1996. Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. *Memory & Cognition* 24. 573–583. Online: https://link.springer.com/content/pdf/10.3758/BF03201084.pdf.

VITEVITCH, MICHAEL. S. 1997. The neighborhood characteristics of malapropisms. *Language & Speech* 40. 211–228.

WIELING, MARTIJIN; ELIZA MARGARETHA; and JOHN NERBONNE. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics* 40. 307–314.

WIELING, MARTIJIN; JOHN NERBONNE; JELKE BLOEM; CHARLOTTE GOOSKENS; WILBERT HEERINGA; and R. HARALD BAAYEN. 2014. A cognitively grounded measure of pronunciation distance. *PLoS ONE* 9, e75734. Online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3886970/pdf/pone.0075734.pdf.

WIENER, SETH, and RORY TURNBULL. 2016. Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech* 59. 59–82.

XU, YI, and EMILY Q. WANG. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33. 319–337.

XU, YISHENG; JACKSON T. GANDOUR; and ALEXANDER L. FRANCIS. 2006. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America* 120. 1063–1074.

YANG, CATHRYN, and ANDY CASTRO. 2008. Representing tone in Levenshtein distance. *International Journal of Humanities & Arts Computing: A Journal of Digital Humanities* 2. 205–219.

YAO, YAO, and BHAMINI SHARMA. 2017. What is in the neighborhood of a tonal syllable? Evidence from auditory lexical decision in Mandarin Chinese. *Proceedings of the Linguistic Society of America* 2. 45–1–14. Online: http://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4090/3783.

YE, YUN, and CYNTHIA M. CONNINE. 1999. Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes* 14. 609–630.

YIP, MOIRA JEAN. 1980. *The tonal phonology of Chinese* (Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, Mass.). Online: https://dspace.mit.edu/handle/1721.1/15971.

ZEE, ERIC. 1999. Chinese (Hong Kong Cantonese). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, ed. by International Phonetic Association, 58–60. Cambridge: University Press.

Notes

[1] We will not cover distance measures based on historical sound changes (e.g. Oakes 2000), methods to combine phonological distance to allow for comparison of languages (Ellison & Kirby 2006), or distance metrics that rely on lists of correspondences between different dialects (Wieling et al. 2012 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 & Heeringa 2004) and cochleagrams (Heeringa 2004:79–120); however, one of the phonological distance 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 Appendix A).

[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.

[5] In priming studies, facilitation effects have been observed across syllabic positions—that is, a phoneme in one position facilitates the processing of the phoneme in another syllabic position (Dufour & Grainger 2019). It was also found that priming effects among phonemes are sensitive to stylistic and social properties (Jones & Clopper 2019). Such results suggest that the distance between phoneme sequences should incorporate psychological and social factors as well. While the current study restricts its limit to sounds' phonological properties, such psychological and social factors should be incorporated when building a complete model of phonological distance measure.

[6] We excluded the target representation (Xu & Wang 2001). Xu and Wang propose characterizing 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, that is, tones 2 and 5.

7 In other varieties of Cantonese, like Guangzhou Cantonese, 53 is still phonologically contrastive.

8 The present study is a part of an ongoing project to build a model of Cantonese phonotactics. The results of this paper will be primarily used to build a GENERALISED NEIGHBOURHOOD MODEL (GNM) of Cantonese phonotactics (see Bailey & Hahn 2001). In constructing GNM models for the participants, we aim to use the current results to construct distance metrics. Therefore, in the current experiment, we show participants two recordings in each trial, including one real word and one word that may or may not be existent, and ask them to judge the distance between the two.

9 There could be differences between Hong Kong Cantonese speakers and those who speak Cantonese overseas as a heritage language. Unfortunately, we did not include a way to tell if all participants are Hong Kong Cantonese speakers currently living in Hong Kong specifically, although the survey was mainly distributed in Hong Kong through social media channels where we would expect most participants to be from Hong Kong.

10 Note that this graph should only be treated as a rough visualization of the data. There are many cases of overlapping points, but we have not scaled the sizes of the dots according to the number of samples in a position because of insufficient space. Certain trends are nonetheless clearly discernible.

11 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 half-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 an LKJ prior (Lewandowski et al. 2009) on its Cholesky decomposition.

12 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 Appendix C.

13 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. Note that 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.

14 Note that our model assumes no difference between onsets and codas, which may not always be true. We ran another version of the model where the [spread glottis] feature is neutralized (with value 0) in coda position. However, there were no substantial differences in the results. The coefficients of onsets, nuclei and tone were estimated at 1.87 (SE: 0.26; 95% CI: (1.36, 2.38)), 2.06 (SE: 0.32, 95% CI: (1.45, 2.68)), 0.65 (SE: 0.27, 95% CI: (0.10, 1.19)), and 0.85 (SE: 0.22, 95% CI: (0.43, 1.30)) respectively. The difference between onsets and nuclei and between codas and tones are respectively estimated at −0.2 (SE: 0.4, 95% CI: (−0.98, 0.61)) and −0.21 (SE: 0.37, 95% CI: (−0.92, 0.5)), revealing little difference. The difference between nuclei and codas remained at 1.44 (SE: 0.44, 95% CI: (0.57, 2.32)).

15 Again, we refit a model using a separate phonemic representation for final stops, with almost no differences in results. The coefficient of onsets, nuclei, and coda was estimated at 2.51 (SE: 0.42; 95% CI: (1.71, 3.35)), 1.42 (SE: 0.40, 95% CI: (0.65, 2.23)), and 0.90 (SE: 0.38, 95% CI: (0.15, 1.64)) respectively, and that of tones was estimated at 1.27 (SE: 0.25, 95% CI: (0.79, 1.76)). Based on posterior draws, the differences between onset and nucleus, nucleus and coda, and coda and tone weighting are estimated at 1.09 (SE: 0.64, 95% CI: (−0.16, 2.39)), 0.52 (SE: 0.6, 95% CI: (−0.66, 1.75)), and −0.37 (SE: 0.43, 95% CI: (−1.22, 0.46)) respectively. Again, we do have weak evidence that onsets are weighted heavier than nuclei, since a 90% credible interval is (0.05, 2.16).

16 Gandour's CONTOUR feature indicates whether a tone is contour or level; his DIRECTION feature is what we refer here to as contours.

17 Note that the confidence intervals here are calculated using frequentist principles, in particular the asymptotic distribution of the MLE. They are interpreted as follows: If we repeat the same data collection method 100 times, on average we should expect that confidence intervals all cover the true values 95 times. This is different from the credible intervals we have seen before, calculated using Bayesian principles, where we may say that the parameter's true value has 95% chance of falling into the interval.