Accounting for lexical tones when modeling phonological distance

YOUNGAH DO
The University of Hong Kong

RYAN KA YAU LAI
University of California, Santa Barbara

Methods of quantifying distance between sound sequences are known as phonological distance measures. Despite the wide application across subfields, phonological distance has been calculated mainly with features related to consonants and vowels. This research report establishes new measurements of phonological distance that incorporate 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 measures and functional load, suggesting that lexical components which play a more crucial role in phonological distance judgments are lexically less predictable as well. Implications for phonological distance measures are discussed.*

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

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 * An early version of this paper was presented at the annual meeting of the Society for Computation in Linguistics (SCiL) in 2019. In addition to that audience, we thank Language referees and editors for their great help in improving 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. Correspondence concerning this paper should be addressed to the first author.

Printed with the permission of Youngah Do & Ryan Ka Yau Lai. © 2021.
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 and Igbo, or pitch-accent languages like Swedish and Japanese). For example, Malins and Joannisé (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 research report 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. 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 of 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 that accounts for both tones and segments, we apply 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, tone 3: 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 that incorporates both pitch-based and contour-based tone contrasts.

To determine the optimal phonological distance measure that matches native speakers’ judgments, §2 first defines a variety of variables that can be used as metrics to compare phonological distance among segments and tones. Section 3 presents a phonological-distance judgment experiment, from which data about native speakers’ judgments of distance between sound sequences are obtained. Through comparison of experimental results with theoretically predicted distances, we investigate how best to predict speakers’ distance judgments. To do so, we specifically explore the following three topics: (i) the relative contributions of segmental and tonal distances to phonological distance judgments, (ii) the phonological distance metrics of segments and tones that can best predict speakers’ judgment data, and (iii) the relative contributions of syllable components (onset, nucleus, coda, tone) in calculating phonological distance. To further consider the contribution of each syllable component to phonological distance judgments, §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 load, and show that syllabic components that play a more important role in phonological distance judgments are lexically less predictable. Section 5 discusses methodological implications of the current study.

2. Distance measures. This section provides an overview of the distance measures that are tested against our experimental data in §3. Segmental distance measures are
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 in determining 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 gradient differences between phonemes into account; for example, /b/ is equi-distant 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. This method is called the Hamming distance measure (see e.g. Gildea & Jurafsky 1996, Pierrehumbert 1993).

\[
\text{Distance}_{\text{Hamming}} = \frac{\text{different features between phonemes}}{\text{total number of phonological features}}
\]

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. Contrasting with 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 and nonshared natural classes of the two phonemes. The formula for the natural class-based measure is in 2, adapted from Frisch et al. 1997.

\[
\text{Distance}_{\text{Natural Class}} = \frac{\text{nonshared natural classes}}{\text{shared natural classes} + \text{nonshared unnatural classes}}
\]

As an example of the natural class-based measure, refer to the system of stops in Table 1.

<table>
<thead>
<tr>
<th>LAB</th>
<th>VOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>[p]</td>
<td>+</td>
</tr>
<tr>
<td>[b]</td>
<td>+</td>
</tr>
<tr>
<td>[t]</td>
<td>−</td>
</tr>
<tr>
<td>[d]</td>
<td>−</td>
</tr>
</tbody>
</table>

Table 1. Sample feature values for a system of stops.

Natural class memberships for these stops, defined by finding all possible combinations of feature values, are given in Table 2.

---

2 Null features are usually thought to be different from both positive and negative values (Pierrehumbert 1993). We adopt this assumption in this study, except when using Broe’s information-gain weighting (see Appendix A in the supplementary materials, available at http://muse.jhu.edu/resolve/117).

3 This was originally a similarity measure. It was converted into a distance measure by subtracting the maximum similarity from 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.
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, \([+\text{lab}], [-\text{lab}], [+\text{voi}], [+\text{lab}, +\text{voi}], [-\text{lab}, +\text{voi}])\). 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/, the labial feature value is different but the voicing feature value is the 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: for example, 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, another distance measure is 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 the Euclidean distance as the diagonal line 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 Fig. 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.

\[
\text{Distance}_{\text{Euclidean}} = \sqrt{\frac{\sum_{j,k} (f_i(p_j) - f_i(p_k))^2}{\max_{j,k} \left[\sum_{i} (f_i(p_j) - f_i(p_k))^2\right]}}
\]

\[
\text{Distance}_{\text{Manhattan}} = \frac{\sum_{j,k} |f_i(p_j) - f_i(p_k)|}{\max_{j,k} \left[\sum_{i} |f_i(p_j) - f_i(p_k)|\right]}
\]

An assumption behind the distance measures in 1–4 is that features are weighted equally. For instance, two phonemes differing in the \([±\text{voi}]\) feature are assumed to be equally distinctive as two phonemes differing in the \([±\text{continuant}]\) feature. However, this assumption may not be true. There are several ways to assign different ‘weights’ to different features. The first approach considers the weights to be free parameters, and a model finds the optimal weights to best predict the distance (Kondrak 2002). Theoretically, Nerbonne and Heeringa also used a distance based on the Pearson correlation between feature vectors, though Heeringa (2004) points out theoretical problems with this approach, and in his perception experiment, the Pearson-based method performed worst by far. Therefore, we have not adopted it.
cally, 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 modeling purposes. 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 with 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 the fact that certain feature values may be null (Broe 1996). The formula and Broe’s modification are presented in appendix §A2 in the supplementary materials.5

In the current study, we apply the distance measures from 1–4 in order 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 appendix Table A1, 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 appendix Table A2. 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 cases, **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

---

5 All appendices referred to throughout the paper are available in the supplementary materials at http://muse.jhu.edu/resolve/117.
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 1–4 or a simple all-or-nothing cost (i.e. no change vs. change). The cost for insertions and deletions is assumed to be half of the average of all phonemic distances between any two phonemes (Nerbonne & Heeringa 2001). Take as an example the 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 a 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 considers 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 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 the rhyme, primarily from priming studies (Radeau et al. 1998, Turnbull & Peperkamp 2017), supports another method, namely the syllable-rhyme approach. According to the syllable-rhyme approach, the distance of sound sequences is computed over onsets and rhymes, combining the nucleus and coda together.6

Table 3 summarizes the segmental distance measures introduced in this section. The numbers indicate the corresponding formulas in §2.1. These measures are adopted in our study of Cantonese.

<table>
<thead>
<tr>
<th>DISTANCE MEASURE BETWEEN PHONEMES</th>
<th>FEATURAL REPRESENTATION</th>
<th>ABBREVIATION</th>
<th>DISTANCE MEASURE BETWEEN PHONEME SEQUENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-or-nothing</td>
<td>None</td>
<td>Simple</td>
<td>Levenshtein (with the assumptions of linear strings or syllabic structure)</td>
</tr>
<tr>
<td>Hamming (1)</td>
<td>Binary</td>
<td>Binary</td>
<td></td>
</tr>
<tr>
<td>Natural class (2)</td>
<td></td>
<td>Natural class</td>
<td></td>
</tr>
<tr>
<td>Hamming (1)</td>
<td>Multivalued</td>
<td>Multivalued (H)</td>
<td></td>
</tr>
<tr>
<td>Euclidean (3)</td>
<td>Multivalued (E)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manhattan (4)</td>
<td>Multivalued (M)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. 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 represen-

6 In priming studies, facilitation effects have been observed across syllabic positions—that is, a phoneme in one syllabic position facilitates the processing of the phoneme in another (Dufour & Grainger 2019). It has also been 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 measures.
tations and their descriptions should be modified depending on the tonal system of the language concerned.

Of the six tonal representations presented in Yang & 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. 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). The autosegmental representations 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 & 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 4.

![Figure 2](image-url)

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

<table>
<thead>
<tr>
<th>TONE</th>
<th>(a) CHAO TONE LETTERS</th>
<th>(b) AUTOSEGMENTAL REGISTER</th>
<th>(c–e) ONSET-CONTOUR-OFFSET TONE</th>
<th>ONSET</th>
<th>CONTOUR</th>
<th>OFFSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>+</td>
<td>HH</td>
<td>H</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>+</td>
<td>LH</td>
<td>M</td>
<td>R</td>
<td>H</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>+</td>
<td>LL</td>
<td>M</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>−</td>
<td>LL</td>
<td>L</td>
<td>F</td>
<td>L*</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>−</td>
<td>LH</td>
<td>L</td>
<td>R</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>−</td>
<td>HH</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
</tbody>
</table>

**Table 4.** 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 4, tonal distances between two tones are calculated. Here, the distance measures introduced in §2.1 can also be applied. For all tonal representations in Table 4, the Hamming and Levenshtein distances can be calculated when tonal differences are assumed to be categorical (i.e. two tones are the same or different). The distances between two ‘symbols’ are set to be 0 when they are the same and 1 when they are different, where each character in the representations in Table 4 is treated as a ‘symbol’—each number representation in (a); +/−, high (H), and low (L) in (b); high (H),

---

7 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—tones 2 and 5.

8 In other varieties of Cantonese, like Guangzhou Cantonese, 53 is still phonologically contrastive.
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 study with Cantonese showed that the Hamming and Levenshtein distance measures resulted in no differences for the tonal representations in (a) and (b) (see appendix Tables A6–A9 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 in (c)–(e) should be the same. So we report only 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 as bearing its own numeric value (scalars). Note that this is because the nature of Chao tone letter representation is different from the autosegmental (b) or onset-contour-offset representations (c–e), which involve only categorical values. We calculated tonal distances in Cantonese based on the Hamming, Euclidean, and Manhattan distance measures, shown in appendix Tables A3–A5. 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 to 1, and the maximum segmental distance is thus the sum of these three distances (3), our assumption here is therefore that tone potentially contributes about as much to distance as one segment, resulting in tonal distances ranging from 0 to 1.

In this paper, we also consider distances between disyllables. For such distances, we 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 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 classes, 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, and 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 now use Cantonese as a case study to show how phonological distance between words should be calculated for tone languages, investigating the following three topics: (i) the relative contributions of segmental and tonal distances to phonological distance judgments, (ii) the optimal segmental and tonal distance metrics that can best predict speakers’ distance judgments, and (iii) 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 distance and asks native speakers to judge the similarity between the two sound sequences.

3. Phonological distance judgment.

3.1. Experiment.

Design. A set of seventy-two monosyllabic and seventy-two disyllabic sound sequences was created. The stimuli list is provided in appendix Tables B1–B2. Two criteria were considered in designing the stimuli. 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 determine the 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 done 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 identical 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, we ensured that each segmental distance and tonal distance region was selected the same number of times in the stimuli. We also ensured that each paired segmental distance and tonal distance region was shown the same number of times in the stimuli. Additionally, every possible segment in every position appeared an equal number of times. 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). Phono-tactically illegal phonemes in onset, nucleus, and coda positions were excluded when creating the nonwords; for example, no fricatives were in coda position, following a Cantonese phonotactic restriction. However, no other phonotactic constraints were 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 the merging of codas [-t] and [-k], recorded the items. All of 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 toward native speakers of Hong Kong Cantonese. Each

9 The present study is part of an ongoing project to build a model of Cantonese phonotactics. The results of this paper will be primarily used to build a generalized neighborhood 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 showed participants two recordings in each trial, including one real word and one word that may or may not be real, and asked them to judge the distance between the two.

10 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 identify 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 in which we would expect most participants to be from Hong Kong.
experimental item was placed on a separate page. On each page, 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, participants were asked to rate the similarity
between the two items on a scale from 0 to 100, where 0 means the two items were
completely different and 100 means they were identical. The similarities were then con-
verted into distances by subtracting the similarity from 100. Before the judgment test, a
screening task was added in the form of AXB tests to ensure that participants could per-
ceptually 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 ensure that the Cantonese spoken by participants was
fairly homogenous and rarely involved dialectal varieties. If participants submitted an
incorrect answer to any of screening questions, the experiment stopped.

Participants. After the survey was circulated on social media platforms in Hong
Kong, anonymous participants were recruited using snowball sampling. The number of
participants who passed the screening task was sixty-one. Twenty-nine participants
completed all 144 questions, while others submitted incomplete forms (mean comple-
tion rate = 97%). The data from all of the participants were used to fit the model regard-
less 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 answered all of the questions will have sub-
ject-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 modeling
decisions. Distances (i.e. judgments from participants) ranged from 0 to 100; these were
divided by 25 to make them range from 0 to 4. This was to make the range of distance
judgments the 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. In Fig. 3, each scatterplot shows the relationship between the judged
distance from a participant (y-axis) against the theoretical segmental distance calcu-
lated using natural class-based distance measures for segments (x-axis, ranging from 0
to 3, with onset, nucleus, and coda each getting a distance of 1) and against the theoret-
tical tonal distance, calculated using Hamming distance measures for the onset-contour-
offset representation; light gray points are items with a tonal distance of 0; dark gray
dots are items with a tonal distance of 0.5; black dots have tonal distances of 1.11

As shown in Fig. 3 and Fig. 4, there seems to be a rough correlation between the
judged distance from participants (y-axis) and theoretically predicted segmental dis-
tance (x-axis): segmentally distinctive items were judged to be more different. It is less
clear, at a descriptive level, whether tonally more distinctive patterns (black > dark gray
> light gray) were also judged to be 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 for perceiving the maximal distance differ among individuals.

11 Note that this graph should be treated only 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 posi-
tion because of insufficient space. Certain trends are nonetheless clearly discernible.
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 modeling allows the partial pooling of data from different items and from different participants so that the model can consider variability in the data and can produce high-variance estimates (Barth & Kapatsinski 2018, Gelman & Hill 2007). Also, instead of a frequentist approach, we chose Bayesian modeling (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 that include large variations, as in our case. Finally, in this model, the distance judgments were treated as a right-censored variable (Gelman et al. 2014:225–26), which assumes that there are some underlying distances that may exceed 4 (1 each for onset, nucleus, coda, and tone) but the data are truncated if the number goes beyond it, the setting of which can be justified by the raw data in Figs. 3 and 4. The models were fit using the R package brms, version 2.4.0 (Bürkner 2017a,b), which provides an 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.12

12 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-
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 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, and 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 (i) discover relative contributions of segmental distance and tonal distance, (ii) identify the ideal distance metrics to predict native speakers’ distance judgments, and (iii) determine relative contributions of onset, nucleus, coda, and tone within a syllable.

To address (i) and (ii), we fit the full model to different segmental and tonal distance measures presented in §2, comparing their predictive power. To address (iii), we ran an

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.
additional model that separates onset, nucleus, and coda distances. Apart from the case where we identify the ideal distance metrics (question (ii) above), all of the models throughout the results section are based on the 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 (i)–(iii) above.

**Relative contributions of segmental distance and tonal distance.** The first objective is to discover the 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, an examination of the model parameters suggests that segmental distance is weighted more than tonal distance in predicting the distance judgment data. The population-level estimates of the coefficient of segmental distance ($\mu_\gamma$) is estimated at 1.50 ($SE = 0.14$, 95% CI [1.23, 1.77]), much higher than that of tonal distance ($\mu_\delta$), 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 judgment 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 ($\mu_\gamma$ and $\mu_\delta$) 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 range 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 best reflect native speakers’ phonological distance judgments. The full model was fit to all of the logically possible combinations of segmental and tonal distance measures this study considers (see §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 given in Table 5. The lowest WAIC values, indicating the best model fit, were achieved with the Hamming distance between multi-valued features for segments, with the onset-contour-offset representation-based measure.

13 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.
for tones (4682.1, in boldface in Table 5). When segmental distance itself is considered, the models with multivalued features using the Hamming distance measure consistently showed the best performance (horizontal gray highlights in Table 5). 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 gray highlights in Table 5).

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

Table 5. WAIC values of the monosyllable model using different segmental and tonal distances without information-gain weighting. (H): Hamming, (E): Euclidean, (M): Manhattan distances.

Second, the results for disyllables are given in Table 6. As shown, the Hamming distance between multivalued features for segments, with the contour-offset representation-based measure for tones, performed best (7153.0, in boldface in Table 6). This result is similar to that of monosyllables. As to the segmental distance itself, the general tendency is the same as with monosyllables: the multivalued feature representations were best, especially with the Hamming distance (gray horizontal highlights in Table 6). Of the tonal distances, the contour-offset representation performed well, although the WAIC values were not substantially lower than those of other models. This result differs from that of monosyllables.

<table>
<thead>
<tr>
<th></th>
<th>CHAO (H)</th>
<th>CHAO (M)</th>
<th>CHAO (E)</th>
<th>AUTOSEG</th>
<th>O-C</th>
<th>O-C-O</th>
<th>C-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>7168.7</td>
<td>7172.0</td>
<td>7185.4</td>
<td>7237.2</td>
<td>7177.2</td>
<td>7176.4</td>
<td>7168.5</td>
</tr>
<tr>
<td>Natural class</td>
<td>7185.7</td>
<td>7194.0</td>
<td>7201.0</td>
<td>7247.7</td>
<td>7194.5</td>
<td>7189.9</td>
<td>7179.2</td>
</tr>
<tr>
<td>Binary (H)</td>
<td>7191.2</td>
<td>7204.5</td>
<td>7213.0</td>
<td>7249.9</td>
<td>7200.7</td>
<td>7193.2</td>
<td>7188.0</td>
</tr>
<tr>
<td>Multivalued (E)</td>
<td>7161.6</td>
<td>7172.6</td>
<td>7181.1</td>
<td>7226.6</td>
<td>7175.1</td>
<td>7164.9</td>
<td>7158.8</td>
</tr>
<tr>
<td>Multivalued (M)</td>
<td>7162.0</td>
<td>7175.1</td>
<td>7181.1</td>
<td>7226.6</td>
<td>7177.9</td>
<td>7168.5</td>
<td>7158.5</td>
</tr>
<tr>
<td>Multivalued (H)</td>
<td>7163.5</td>
<td>7173.4</td>
<td>7181.3</td>
<td>7227.5</td>
<td>7178.5</td>
<td>7165.7</td>
<td>7153.0</td>
</tr>
</tbody>
</table>

Table 6. 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 7), we used the offset of the first syllable and the onset of the 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 7), 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 7). 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 C-O+ (type 3), the two offsets are H and H, so the pitch-level change is level.

As shown in Table 7, 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 (boldfaced in Table 7). 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 7), but not simply between the offset of a preceding syllable and the onset of the following syllable (type 1 in Table 7).

<table>
<thead>
<tr>
<th></th>
<th>O-C-O</th>
<th>O-C-O+ (TYPE 1)</th>
<th>avg O-C-O+ (TYPE 2)</th>
<th>C-O</th>
<th>C-O+ (TYPE 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>7176.4</td>
<td>7177.8</td>
<td>7164.6</td>
<td>7168.5</td>
<td>7152.8</td>
</tr>
<tr>
<td>Natural class</td>
<td>7189.9</td>
<td>7188.4</td>
<td>7176.7</td>
<td>7179.2</td>
<td>7162.8</td>
</tr>
<tr>
<td>Binary (H)</td>
<td>7193.2</td>
<td>7180.2</td>
<td>7189.4</td>
<td>7188.0</td>
<td>7169.7</td>
</tr>
<tr>
<td>Multivalued (E)</td>
<td>7164.9</td>
<td>7153.8</td>
<td>7161.3</td>
<td>7158.8</td>
<td>7142.3</td>
</tr>
<tr>
<td>Multivalued (M)</td>
<td>7168.5</td>
<td>7153.0</td>
<td>7163.7</td>
<td>7158.5</td>
<td>7143.8</td>
</tr>
<tr>
<td>Multivalued (H)</td>
<td>7165.7</td>
<td>7153.0</td>
<td>7163.6</td>
<td>7153.0</td>
<td>7138.6</td>
</tr>
</tbody>
</table>

**Table 7.** WAIC values of the disyllable 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 2004.14 This model performed far worse than any of the phonological models in Table 5, 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) was 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 re-

---

14 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.
sults 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 appendix Tables B5–B6 for monosyllables and in appendix Tables B10–B11 for disyllables.

**Relative contribution of syllable components.** The third objective is to determine the 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, standard error (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 a 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 roles of onset and nucleus or between those of 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]). The results suggest that the hierarchy of syllabic components’ contributions is onset, nucleus > coda, tone when making phonological distance judgments of monosyllables.

![Figure 5. Estimates of the weightings of onset, nucleus, coda, and tone, along with 95% credible intervals, for monosyllables.](image)

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 central 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 [−0.2, 1.52]), and 0.23 (SE = 0.48, 95% CI [0.05, 0.72]), 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]).

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]).
CI $[-0.81, 1.68]$), and $-0.35$ ($SE = 0.43$, 95% CI $[-1.19, 0.48]$), respectively, with the 95% credible intervals all including 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 more heavily than nucleus, since the 95% credible interval for their difference excludes zero: $[0.02, 2.26]$. The results suggest that the hierarchy of syllabic components’ contributions is onset $>$ nucleus, coda, tone when making phonological distance judgments of disyllables.

16 Again, we refit a model using a separate phonemic representation for final stops, with almost no differences in results. The coefficients of onsets, nuclei, and coda were 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 more heavily than nuclei, since a 90% credible interval is $[0.05, 2.16]$.
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 differences in the relative contributions of segments and tones, which needs future research on crosslinguistic 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. The unclear pattern among disyllables can be attributed to the fact that the disyllabic test items may be less representative of the lexicon than the monosyllables are. Recall that our test included the same number of monosyllables (n = 72) and disyllables (n = 72). Due to this setting, a smaller 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 Figs. 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 about the high frequency of maximal distances observed in 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 to be separate from segments, hence tonal and segmental distances are computed independently as inputs to the final phonological distance judgments. It is possible to assume instead that the tone is tied to the nucleus. 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 the 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 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 are. 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 best for predicting the observed distance judgments. Unlike the binary features, the multivalued
features distinguish between allophones and allow 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 judgment 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 on 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).17

We have also shown that the information-gain weighting did not help to improve the
models’ predictions for any type of distance metrics. This is consistent with the results
from Nerbonne & Heeringa 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 just one possi-
ble type of weighting scheme and other schemes might potentially improve the predictive
power of phonological distance judgments. We leave this for future research.

Relative contributions of onset, nucleus, and coda. We further split seg-
ments into onset, nucleus, coda, and tone in order to investigate the relative contribu-
tions 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 percep-
tion studies, which showed that spoken word recognition is more challenging when
tone differences are involved (e.g. Cutler & Chen 1997, Keung & Hoosain 1979), sug-
gest ing lower perceptual sensitivity to tone differences than to segmental differences.
For disyllables, nucleus is shown to have a less important role, in contrast to its impor-
tant contribution to distance judgments for monosyllables. For monosyllables, the nu-
cleus 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 simi-
lar 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 lexical knowledge. These con-
siderations as such allow us to speculate about why the role of nucleus differs for mono-

17 Gandour’s contour feature indicates whether a tone is contour or level; his direction feature is what
we refer here to as contours.
sylabes 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 §4 is to investigate 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 the 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) are 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 speakers may not rely heavily on lexically highly predictable elements when evaluating phonological distance. For example, if the coda is always a nasal, thus easily predictable, the phonological properties of the coda do not contribute much to judging how different two items are. Instead, speakers will become more sensitive to lexically uncertain parts (e.g. the 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 predictability, 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 §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 ‘lack of predictability’. When calculated using base 2 logarithms, the formula for entropy is as follows.

\[ (5) \quad - \sum_{i=1}^{n} p_i \log_2 p_i \]

Here \( p_i \) 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 in 5, the entropy is a lower bound on the expected number of ‘bits’, that is, representation in terms of 1s and 0s that are needed to encode information. As an example, let us compare two toy languages with the following probability distributions of nuclei.

\[ (6) \]

a. Language A: /a/ 50%, /u/ 25%, /i/ 25%

b. 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 in 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 × 1 + 0.25 × 2 + 0.25 × 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 

\[ -0.5 \log_2 0.5 - 0.25 \log_2 0.25 - 0.125 \log_2 0.125 - 0.125 \log_2 0.125 = 1.75 \]

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 × 1 + 0.25 × 2 + 0.125 × 3 + 0.125 × 3 = 1.75, again matching the entropy.
We use entropy to predict syllabic components’ relative contributions in phonological distance judgments. 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 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 the 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.

It is necessary to check whether the above entropy differences correspond to their actual differences or if they are just artifacts of our sample. Thus, we computed confidence intervals for the entropy differences to ensure that the results in Fig. 7 are meaningful differences and not simply due to sampling error.18 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 intervals and point estimates, with the point estimates being in the middle of the 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.

18 Note that the confidence intervals here are calculated using frequentist principles, in particular the asymptotic distribution of the maximum likelihood estimate. 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 ninety-five 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 a 95% chance of falling into the interval.
4.2. Functional load analysis. The above calculations of entropies do not take into account properties of the other syllabic components. 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 the coda—then even if there is a huge uncertainty of tone itself, tone becomes less important in making distance judgments because cues from the coda enable 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).

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

Using the formula in 7, we computed functional loads for onset, nucleus, coda, and tone, and plotted the weights of each syllabic component 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 roughly correspond, except for tone. Specifically, the 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 could be because the nucleus and coda have more cooccurrence 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 is therefore not affected by such phonotactic factors. Importantly, except for tone, the results again roughly match our phonological-distance judgment data.

To check the reliability of the results in Fig. 9, we calculated confidence intervals for the differences between the functional loads. The results are in Figure 10. Note that all but the interval for the difference between nucleus and coda in disyllables do not include zero, suggesting meaningful evidence overall, with the exception of the nucleus-coda differences in disyllables. In other words, the functional load hierarchy remains onset > tone > nucleus > coda for monosyllables, but it should be modified to 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.

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 entropy and functional load, we conclude that measures of lexical predictability have the partial power to account for the weightings of syllabic components in phonological distance measures, although they cannot predict the full range of speakers’ phonological distance judgments.
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 the onset consistently contributes more than the coda and tone (though the role of the nucleus is relatively unclear) to phonological distance judgments, and that these results are partially explained by information-theoretic 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 for the nature of phonological distance in tone languages, our modeling 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 measures (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 compli-
cated 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 modeling 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.

REFERENCES


CARRINGTON, BOB; ANDREW GELMAN; MATTHEW D. HOFFMAN; DANIEL LEE; BEN GOODRICH; MICHAEL BETANCOURT; MARCUS BRUBAKER; JIQIANG GUO; PETER LI; and ALLEN RIDDLELL. 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76.1–32. DOI: 10.18637/jss.v076.i01.


GELMAN, ANDREW; JOHN B. CARLIN; HAL STEVEN STERN; DAVID B. DUNSON; AKI VE H TARI; and DONALD B. RUBIN. 2014. Bayesian data analysis. 3rd edn. Boca Raton, FL: CRC Press.


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* 25:341–70. DOI: 10.1017/S0954394513000161.


