Accounting for lexical tones when modeling phonological distance

1. Introduction

English native speakers can intuitively tell that *cat* is phonologically more similar to *cap* than 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 they have been applied in a variety of linguistic fields. In dialectology, phonological distance measures have been frequently used to examine divergence between dialects (e.g. Nerbonne & Heeringa, 1997; Heeringa, 2004; Heeringa, Kleiweg, Gooskens and Nerbonne, 2006; Tang, 2009; Tang and van Heuven, 2009, 2011, 2015). In computational historical linguistics, the measures have been used to align and reconstruct cognate words (e.g. Oakes, 2000). In psycholinguistics, they are often adopted in studies of bilingualism and diglossia to measure effects of between-language or between-variety similarity (e.g. Saiegh-Haddad, 2004). Some older methods of automatic speech recognition use phonological distance measures to compare reference symbols with a system's hypothesized symbols (Fisher & Fiscus, 1993). In phonology, the distance measures have been applied in formulating constraints on alternations (Gildea & Jurafsky, 1996) and phonotactics (e.g. Pierrehumbert, 1993; Frisch, Broe and Pierrehumbert, 1997). In phonotactics, specifically, phonological distance measures are adopted in modeling 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 & Pisoni, 1998; Luce, Goldinger, Auer and Viteviych 2000), 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 and usefulness of phonological distance-based methods, including neighborhood density models, hinges on the quality of the distance measure, i.e. the extent to which it resembles human listeners' perceptual distance. Despite the wide application of phonological distance, research in this domain has focused predominantly on segmental features (e.g., Nerbonne & Heeringa, 1997; Heeringa, 2004) and work incorporating suprasegmental features is rare. While such inadequacy may not heavily affect the validity of phonological distance measures in 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. For example, Malins & 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 provide a way to measure phonological distance between words in languages where suprasegmental features are crucial in creating lexical contrasts. Our focus is tone languages. Although some studies have utilized tonal distance measures with limited discussions on their quality or nature (e.g. Tang and van Heuven, 2009), to the best of our knowledge, few have taken tonal distance metrics themselves as an object of study. To establish proper measurements of phonological distance incorporating tone, we take experimental and modeling approaches, using Cantonese as an exemplary case study. We believe tone is a good example to demonstrate how to measure phonological distance for the following reasons. It is likely more important to incorporate suprasegmentals in lexical density models of languages where suprasegmentals create lexical contrasts, compared to in languages where they cannot. Among lexically contrastive suprasegmental features, tone can involve relatively rich representations, including level and contour tones. Along this line of reasoning, Cantonese is a good example to demonstrate our methodology, with multiple lexical tones, level and contour ones, allowing us to incorporate tonal distance measures based on both categorical and numerical measures.

This paper is organized as follows. Section 2 first defines a variety of metrics to compare phonological distance among segments and tones. Section 3 presents a similarity judgment experiment, the results of which will be compared to the predictions from the distance metrics introduced in Section 2. Through the comparisons of the experimental results with theoretically predicted distances, we aim to answer the following three questions: (a) relative weightings of segmental and tonal distances in making phonological similarity judgments; (b) best phonological distance measures of segments and tones; and (c) relative weightings of syllable components (onset, nucleus, coda, tone) in calculating phonological distance. To further consider the finding on (c), Section 4 attempts a lexical analysis and shows a correspondence between the observed relative weights of syllabic components and predictions from information-theoretic measures. We employ two types of information-theoretic measures of syllabic components, an entropy measure and functional load, and show that speakers assign greater weight to the syllabic components that are lexically less predictable. Section 5 discusses implications of the current study to phonological distance measures beyond Cantonese.

2. DISTANCE METRICS[1]

This section provides an overview of the distance metrics that we will test against our experimental data in Section 3. Segmental distance metrics will be presented first, followed by tonal distance metrics. We then discuss how we apply the metrics to measure phonological distance in Cantonese.

2.1. SEGMENTAL DISTANCE

Phonemic distance. As a first step to determine phonological distance between sound sequences, we measure the distance between phonemes. The simplest approach is ‘categorical’ to assume no distance between phonemes when they are identical and full distance otherwise (e.g. Tang & van Heuven, 2015; Heeringa, Kleiwing, Gooskens & Nerbonne, 2006; *inter alia*). This approach does not take the gradient differences between phonemes into account, e.g., /b/ is equidistant to /p/ and to /l/. In phonology, there are two other influential methods of measuring phonemic distance, using phonological features. First, Hamming distances between binary feature vectors of phonemes can be computed (e.g. Pierrehumbert, 1993; Gildea and Jurafsky, 1996), by finding the number of binary features (e.g., [±voice], [±nasal]) that differ between the two phonemes. The distance can be normalized by dividing by the total number of features, as in (1)[2].

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

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

This method does not take into account *how* phonological features are used to create contrasts between phonemes; it is purely based on counts of (un)shared features. Thus, we may construct a second distance metric based on the phonemes' natural classes, as in (2)[3], adapted from Frisch, Broe and Pierrehumbert (1997). In the distance measurement in (2), the number of unshared natural classes is divided by the total number of natural classes. When this paper adopts binary feature-based measures for Cantonese, we will always normalize the distance using the formula (1), though, so that all distances range from 0 to 1, ensuring comparability between metrics.

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

In dialectological studies, multivalued features are widely used instead of binary features. These features can be categorical or numeric; for example, a 'place' feature may be bilabial, coronal or dorsal (categorical), or it may hypothetically have values 100 for bilabial, 80 for coronal and 20 for dorsal (numeric). When the values are categorical, the Hamming metric in (1) can still be adopted for the calculation between multivalued feature vectors. If the values are numeric, we can use Euclidean or Manhattan distances (Nerbonne & Heeringa, 1997[4]), both of which calculate numeric differences of phonemes' feature values. Specifically, in the Euclidean distance measure in (3), phonological distance is calculated by evaluating the square of the difference between the feature values of the two phonemes under comparison and taking the square root of the sum. To visualize the concepts, Figure 1 shows that 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 in (4) is similar but it sums up the absolute values of the differences between the corresponding feature values of the phoneme pair. In Figure 1, the Manhattan distance is shown in red. In our study, the two distances are also normalized so that they fall between 0 and 1, by dividing the result by the maximum distance. 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:

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

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

[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.

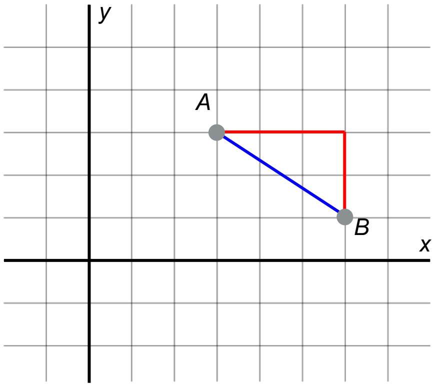

Figure 1: Euclidean distance (blue) and Manhattan distance (red) between two points on the Cartesian plane. Here, the *x*- and *y*-axes can be taken as feature values in a two-feature system.

An underlying assumption behind the distance metrics so far is that features are weighted equally. For instance, two phonemes differing in [±voi] feature are assumed to be equally distinctive to the two phonemes differing in [± continuant] feature. However, this assumption may not be true. There have been several ways to assign different weights to different features. One approach regards the weights as free parameters which are not predetermined by a model but instead a model itself finds weights to optimize the distances' performance (Kondrak, 2002). This could theoretically be achieved by introducing the weights as parameters in our multivalued representation. However, we have refrained from this approach due to its possibility of increasing the complexity of the model, since the weight of each feature is a new parameter. Moreover, it adds complexity to the model-fitting procedure, because insertion and deletion distance is dependent on the distances between phonemes (see section 2.1.2).[5] Instead, we adopt Nerbonne & 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. 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 of interest. Put in a more intuitive way, information gain calculates the difference between the amount of 'uncertainty' in identifying a segment in the lexicon and the average degree of uncertainty left once we figure out the value of the feature in question. Additionally, Broe (1996) proposes modification of the information gain formula, which takes into account that certain feature values may be null. The formula and Broe's modification are presented in 1.2 in Supplementary Materials.

In the current study, we apply the distance metrics from (1) to (4) to measure phoneme distances in Cantonese with an additional consideration of information-theoretic weightings. The exact binary feature set of Cantonese on which the Hamming distance calculation is based is presented in Table 1 in Supplementary Materials 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 2 in Supplementary Materials. We also consider both classic information gain weighting and Broe's modification.

Distance between phoneme sequences. To measure phonological distance between words, calculating individual phoneme distances will not suffice. The distances between words can be computed using an algorithm that quantifies how (dis)similar two 'strings' or 'sequences' are to one another. We compute the segmental distances between a pair of sound sequences using the Wagner-Fischer

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

algorithm for Levenshtein distances (Jurafsky & Martin, 2014). Specifics are as follows. It finds the optimal sequence of deletions, insertions, or substitutions required to transform one string into another while minimizing the total cost of these operations; this cost is the distance between phoneme sequences. For substitutions, the cost was the phonemic distance defined above in (1) – (4) or a simple all-or-nothing cost (i.e. traditional vanilla Levenshtein distance). For insertions and deletions, the cost is set at half the substitution cost between two phonemes (i.e., the average of phonemic distance), following Nerbonne & Heeringa (2001). Take an example of a distance from *ka* to *tap*, where a substitution in onset and an addition in coda are observed. If the optimal distance from /k/ to /t/ is 0.5 and the average phoneme distance cost is 0.7, then the total distance from /ka/ to /tap/ would be 1.2: 0.5 for the /k/ to /t/ substitution plus 0.7 for the addition of coda /p/. While other sequences of operations are conceivable, e.g. turning /k/ into /t/, deleting the /a/ then replacing /p/ with /a/, they incur higher cost and hence are not used as the final distance.

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

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

Table 1: Summary of the different distance metrics investigated in this paper

2.2. TONAL DISTANCE

Tonal representations. In order to measure distances between tones, we must first introduce ways in which tones are represented. We introduce tonal representations with Cantonese. For research on other tone languages, the same representations can be adopted but the specific numbers of representations and their descriptions should be modified depending on the tonal system of a language concerned.

Of the six tonal representations presented in Yang & Castro (2008), we retained the following five which can be replicated in Cantonese: (a) the 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 mostly absent in Hong Kong Cantonese (Bauer & Benedict, 1997), the focus of our case study. 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. In this framework, Tone consists of two binary features, H or 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. For six tones in Cantonese diagramed in Figure 2, their corresponding tonal representations are shown in Table 2.

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

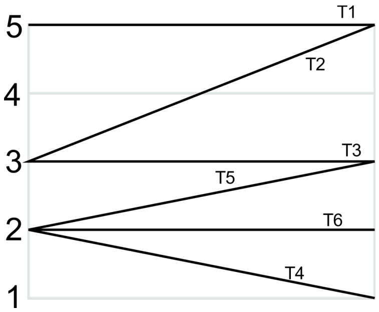

Figure 2: A graphical illustration of the Chao tone letter representations of the six Cantonese tones.

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	VL
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. 'VL' indicates 'Very Low'.

From tonal representations to tonal distance metrics. With the tonal representations in Table 2, we now consider how to measure phonological distances between tones. Hamming and Levenshtein distances are calculated for all tonal representations from (a) to (e) in Table 2, given their categorical nature. We set the distance between two symbols to be 1 when they are different and 0 when they are same, where each character in the representations above is treated as a 'symbol' - each number in (a), +/-, H, and L in (b), H, M, L , and R, in (c-e)[7]. Our distance measures with Cantonese showed that the Hamming and Levenshtein distance measures resulted in no differences for any tonal representations from (a) to (e) (see Table 6-9 in Supplementary Materials for the calculated values). Thus, in the following sections we only report the Hamming distances. For the Chao tone letters (a), Euclidean and Manhattan distances are computed as well because each tone letter bears its own numeric value, as opposed to (b) – (e) representations. That is, we treat the Chao tone letters as integer-valued vectors and evaluated the distances between them in Euclidean space. The calculated tonal distances in Cantonese based on Hamming, Euclidean, and Manhattan distance measures are shown in Table 3 – 5 in Supplementary Materials.

In Section 2, using the segmental distance metrics in Table 1 and tonal distance metrics in Table 2, we obtained distance measures of words in Cantonese. The distance measures will be compared against native speakers' phonological distance judgment data presented in Section 3. Before comparing the metrics with empirical data, we first review previous studies that compared phonological distance metrics.

2.3. PREVIOUS STUDIES ON DISTANCE METRIC COMPARISON

In various subfields of linguistics, proposals have been made to identify the ideal metric of phonological distance. Previous studies comparing different distance metrics, however, have largely

[7] The only exception was 'VL', which means 'very low' thus was simply treated as one single symbol.

focused on segmental features of languages. Somers (1998) considered algorithms to align phonemes in child language with their adult counterparts and evaluated the performance (in terms of the quality of the alignments with real and simulated child-language data) of three segmental feature sets defining similarity metrics: binary articulatory features, Ladefoged (1971)-style multivalued features (similar to the Ladefoged (1975)-style features we implement in this study), and a perceptual distance based on frication and pitch. They reported that the perceptual distance performed worst, though no formal comparison among the metrics was provided. Heeringa (2004) compared the simple all-or-nothing distance, a binary feature system, two multi-valued feature systems, and a variety of phonetic distance measures to compare different Norwegian dialects. It was found that the simple system using the all-or-nothing distance measure works best. Nerbonne and Heeringa (1997) evaluated the performance of several distance metrics in dialect comparison by comparing the results of the different distances against traditional dialectologists' groupings. They compared the dialect distance results to compare Euclidean, Manhattan, and 'Pearson' distance (a measure of the linear correlation between the two variables compared) between multivalued features, with or without information gain weighting, and with one-segment or two-segment representations of diphthongs, along with a simple Levenshtein baseline treating distance between phones as all-or-nothing. It was found that the Manhattan distance between multivalued features without information gain weighting and with two-phone representation of diphthongs worked best.

To our knowledge, Yang and Castro (2008) is the only study whose main focus was to compare tonal distance metrics, which were derived from different tone representations in varieties of Bai and Zhuang. Their results revealed that tonal representations with contour information work best: higher Pearson correlation coefficients between mutual intelligibility and the representations with contour information than those with Chao tone letter, autosegmental or target representations. Despite the novelty of exploring tonal distance metrics, their approach has a disadvantage of not considering the potential confounding effect of segmental distance. The experiment involved measuring the intelligibility of texts spoken in different dialects to speakers of other dialects; thus, tonal and segmental distance can potentially be correlated in their texts. Yet they only assessed the simple Pearson correlation coefficient between the various tonal distance metrics and mutual intelligibility, without partialing out the effect of segmental distance. Therefore, the reported small differences in Pearson correlation may not be necessarily due to the quality of the tonal distance metrics alone. Tang and van Heuven (2011) also looked at the association between three tonal distance metrics and mutual intelligibility among several Mandarin, Wu, Gan, Xiang, Min, Hakka and Yue dialects, including Levenshtein distances between Chao tone letters[8] and onset-contour representations as well as a 'perceptually weighted' distance. Though Tang and van Heuven (2011) did not directly compare the metrics, their point estimates of Pearson correlation coefficients seem to suggest that the representation with contour information outperforms the other two measures, the results consistent with Yang and Castro (2008).

Yang and Castro (2008) additionally compared relative importance of segments and tones by fitting multiple regression models with segmental and tonal distances as independent variables. They concluded that tones may be more important than segments. Unfortunately, they only provide *t*-statistics and *p*-values. These values give us information about the strength of *evidence* for tonal and segmental effects on intelligibility. However, such values cannot provide the strength of *the effects themselves*, which is better represented by point and interval estimates of the regression coefficients. Standard errors were not reported, so we were unable to recover the coefficients in their model or calculate confidence intervals for them. Therefore, it is unclear how great the difference between tone and segmental distance really are from their reported figures. Also, it is often found in perception studies that different people's weightings of different cues may wildly differ (Yu and Zellou, 2018), while Yang and Castro's modelling method (fixed-effects linear regression that does not contain by-subject effects) did not consider such variation into their analysis.

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

2.4. INTERIM CONCLUSION

In Section 2, we have introduced several ways of evaluating segmental distance which assume different ways of calculating distances between phonemes. They include the proportion of unshared natural classes, the Hamming distance between their binary and multivalued feature vectors, and the Euclidean and Manhattan distance between multivalued numeric feature vectors. We have also presented five different types of tonal representations, including Chao tone letters, the autosegmental, onset-contour, onset-contour-offset, and contour-offset representations. Our literature review reveals that no systematic investigation has been conducted on phonological distance measures incorporating segments and tones, presumably due to the lack of research focus on tonal distance metrics. Against this background, we will now use Cantonese as a case study to show how phonological distance between words can be calculated for tone languages. We aim to figure out distance metrics that best reflect speakers' judgements. To do this, 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 items.

3. PHONOLOGICAL DISTANCE JUDGMENT

3.1. EXPERIMENT

Design. We created a question set of 72 monosyllabic and 72 disyllabic sound sequences. The stimuli list is provided in Table 10-11 in Supplementary Materials. When designing the stimuli, we considered two criteria: (a) the items are well balanced across different segmental and tonal distances and (b) the two are not correlated among stimuli. Given that one of the core questions of this experiment is to figure out relative weightings of segments vs. tones, it was important to keep the distances of the two uncorrelated. To evaluate segmental distance, we chose a natural class distance measure following Bailey and Hahn (2001), where natural class distances were computed with deletions and insertions set at half of the average substitution cost. Multiple simulations by picking 10,000 monosyllables from the Hong Kong Cantonese Corpus (Luke and Wong, 2015) at random showed that segmental distances rarely went above 2.5. Based on this observation, we divided segmental distances into four types within the interval of [0, 2.5]: 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. As to the tonal distances, the Hamming distances between onset-contour-offset tonal representations were measured following Yang and Castro (2008) and Tang and van Heuven (2011). We divided tonal distance into three types: high (1, farthest apart), low (0.5, middle) and zero (0, no distance). The design for disyllables was similar. To make it consistent with monosyllables, we divided segmental distance into four levels: high (>3.33), mid (≤3.33 but >1.67), low (≤1.67) and zero (0). This is simply double that of the situation for monosyllables, with each region occupying one-third of the interval [0, 5]. Tonal distance was classified as high (>2), low (≤1), or zero (0), doubled from monosyllables. For both monosyllables and disyllables, we ensured that each segmental distance level and tonal distance level were selected the same number of times in our stimuli design. Moreover, each segmental distance-tonal distance pair was also shown the same number of times in the stimuli. We also made sure that that every possible segment in every position appeared at least once. In both monosyllabic and disyllabic pairs, the first item of each 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 non-word (e.g. *poe6*, *doi6te3*) for a general interest.[9] When creating the non-words, we

[9] 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 (Bailey and 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

excluded absolutely illegal segments in onset, nucleus and coda positions; for example, no fricatives were in coda position, which is phonotactically illegal in Cantonese. However, we did not consider any other phonotactic constraints as we view them as constraints to be discovered later through the phonotactic models based on results of the current study.

A native speaker of Hong Kong Cantonese who is not affected by ongoing sound changes in Cantonese, such as the merging of onsets [n-] and [l-] and codas [t] and [k], recorded the test items. All the items were recorded in a sound-attenuated booth in the authors' institute. The recordings were scaled to 70 dB using the Scale intensity feature in Praat (Boersma and 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 Cantonese.[10] Each experimental item was placed on a separate page. On each page, the participants heard the two audio recordings and judged their similarity using a slider. As we believed that it would be easier to understand similarity than distance, the participants were asked to rate similarity between the two items ranging from 0 to 100, where 0 means the two syllables were completely different and 100 means they were identical. The similarities were then converted into distances by subtracting the similarity from 100. Before the judgment test, a screening task was added in forms of AXB tests to ensure that participants could perceptually distinguish between [n] and [l] onsets, which are merging in some Cantonese speakers (Bauer and 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, Zuo and Wong, 2013). This test was to make sure that the Cantonese spoken by participants was fairly homogenous and rarely involves dialectal varieties. If participants submitted an incorrect answer to any of screening questions, the experiment stopped.

Participants. In total, data were collected from 61 anonymous participants after circulating the survey on social media platforms in Hong Kong using snowball sampling. Twenty-nine participants completed all 144 questions while others submitted incomplete forms. 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 judgements (Gelman and Hill, 2007).

3.2. RESULTS FOR MONOSYLLABLES

3.2.1. Descriptive data

We first explore descriptive patterns in the data through scatterplots of segmental and tonal distances against distance judgements to inform our modelling decisions. Before descriptive analysis, the judged distances, which were originally in the range of 0-100, were scaled to lie between 0 and 4 by dividing by 25 for the simplicity of interpretation; the maximum tonal distance ranges from 0-1 and maximum segmental distance ranges from 0-3, assuming 0-1 for each segment, so they sum up to four. Each graph in Figure 3 represents the data from each participant who completed the test. Each scatterplot shows the relationship between the judged distance from a participant (*y*-axis ranged from 0 to 4) against natural class-based segmental distance (*x*-axis ranged from 0 to 3). The judgments were also plotted against tonal distance (long, mid, zero distance); black points are items

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.

[10] There could be differences between Hong Kong Cantonese speakers and those who speak Cantonese overseas as a heritage language. Unfortunately, we did not ask all participants to be speakers of Hong Kong Cantonese specifically, although the survey was mainly distributed in Hong Kong through social media channels where we expect most participants to be from Hong Kong.

with tonal distance of 0; dark grey dots are items with tonal distance of 0.5; light grey dots have tonal distances of 1.[11]

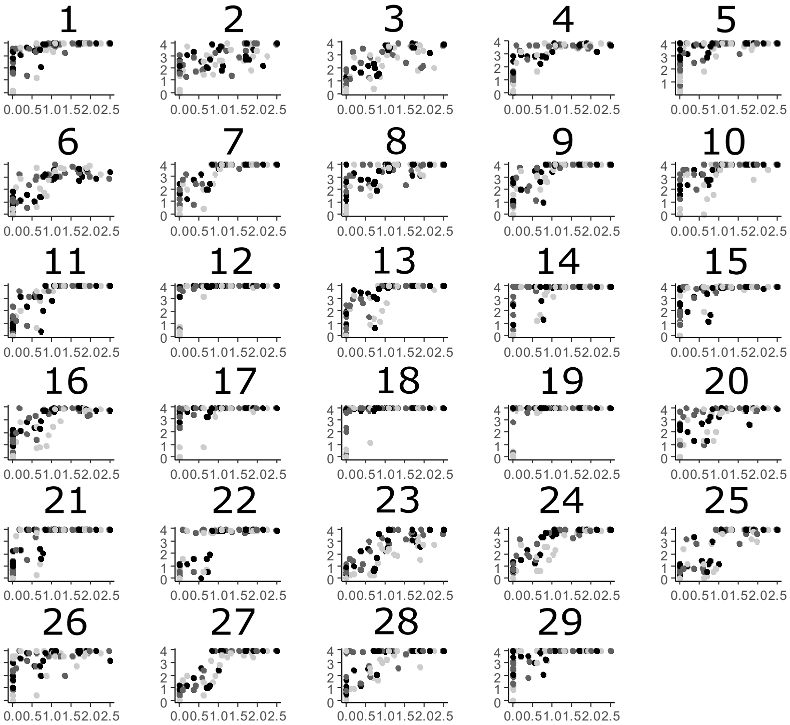

Figure 3: Scatterplots of distance judgements against theoretical segmental distance. Light grey points are those with tonal distance of 0; dark grey dots are those with tonal distance of 0.5; black dots have tonal distances of 1. Numbers indicate participants' numbers.

As shown in Figure 3, there seems to be a rough correlation between the judged distance from participants (*x*-axis) and theoretically predicted segmental distances (*y*-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 more different. Crucially, Figure 3 shows that the strength of the relation between distance judgements and the theoretical distances varies greatly among participants: some are categorical judges while others are more gradient, and the thresholds to perceive the maximal distance differ among individuals. Based on this observation, we chose a multilevel model that allows an item-level random intercept as well as subject-level random slopes for tonal and segmental distance. Instead of a frequentist approach, we opted for Bayesian multilevel modelling (Gelman & Hill, 2007; Nicenboim & Vasishth, 2016) for the following reasons. Different from frequentist analysis, Bayesian multilevel 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 with large variations as in our case. Additionally, the use of multilevel modelling, similar to frequentist mixed-effects models, allows the partial pooling of data from different items and participants. This approach avoids ignoring variability in the data (as is done in complete pooling) or ignoring information in the data to produce high-variance estimates (as is done in no-pooling models) (Gelman & Hill, 2007; Barth and Kapatsinki, 2018). Finally, in this full model, the distance

[11] 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.

judgements are treated as a right-censored variable (Gelman, Carlin, Stern and Rubin, 2014, pp.225-226). This assumes that there is some underlying distance which may exceed 4 (max 3 for segments + max 1 for tones) but the data is truncated if the number goes beyond it, the setting of which can be justifiable from the raw data in Figure 3. The specifics of the full Bayesian model adopted in our study is given below.

(5) $Y_{ij} \sim N\left(\mu + \alpha_i + \beta_j + \gamma_j t_i + \delta_j s_j, \sigma^2\right), i = 1, \dots, 72, j = 1, \dots, n$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\begin{bmatrix} \beta_j \\ \gamma_j \\ \delta_j \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu_\gamma \\ \mu_\delta \end{bmatrix}, \begin{bmatrix} \sigma_\beta^2 & \rho_{\beta\gamma}\sigma_\beta\sigma_\gamma & \rho_{\beta\delta}\sigma_\beta\sigma_\delta \\ \rho_{\beta\gamma}\sigma_\beta\sigma_\gamma & \sigma_\gamma^2 & \rho_{\gamma\delta}\sigma_\gamma\sigma_\delta \\ \rho_{\beta\delta}\sigma_\beta\sigma_\delta & \rho_{\gamma\delta}\sigma_\gamma\sigma_\delta & \sigma_\delta^2 \end{bmatrix} \right)$$

$$Y_{ij}^* = \begin{cases} Y_{ij} \ if \ Y_{ij} \leq 4 \\ 4 \ \ if \ Y_{ij} > 4 \end{cases}$$

where Y_{ij}^* is the *j*th participant's response to the *i*th item, μ is the overall (population-level) intercept, α_i and β_j are respectively item-level and subject-level intercepts centred at zero, μ_γ and μ_δ are the mean coefficients of segmental and tonal distance, and γ_j and δ_j are their subject-level counterparts. ρ_{AB} indicates the population correlation between A and B, and σ_A indicates the standard deviation of A. The models were fit using the R package brms version 2.4.0 (Bürkner, 2017; Bürkner, in press), 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.[12]

Model specifications may vary, and we first need to identify the best model, from which we report our results. For this, we relied on the Watanabe-Akaike Information Criterion (WAIC) values. Roughly speaking, lower WAIC values indicate better match with data. Comparisons of the WAIC values of the full model with various reduced models showed that the full model is the optimal model, i.e. containing the item-level random intercept, all subject-level random effects, as well as the censoring assumption. Results in the following sections, therefore, are based on the full model. Detailed justification of the model specification, as well as the model comparison procedure, are provided in 3.1 in Supplementary Materials.

Recall our three questions in the experiment. To understand how native speakers make phonological similarity judgments, we aim:

(a) to find out relative weightings of segments and tones
(b) to identify the ideal distance metrics to reflect native speakers' similarity judgments
(c) to determine relative weights of onset, nucleus, coda, and tone within a syllable

To answer (a) and (b), we fit the full model to different tonal and segmental distances presented in Section 2, comparing the predictive power to find the optimal distance. To answer (c), we also run a model that separates onset, nucleus, and coda distance to see if the syllabic components will differ in weighting. Apart from the models we use to compare different tonal and segmental distances (section 3.2.2 and 3.3.2), all of the models throughout the results below are based on natural class distance and Hamming distances between onset-contour-offset tonal representations. This was to make it consistent with our stimulus design, which was created with these two distance measures in

[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-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.

mind, ensuring a wider spread among different possible values of the two distances and that the two distances are not correlated in the design. It also makes the results more comparable with those in the disyllable section, where the optimal metric may be different from in monosyllables. Results are reported following the order of three questions (a)-(c) above.

3.2.2. Relative weightings of segments and tones

Examining the estimated values of the model parameters and their uncertainty estimates suggests that the segmental distance plays a more crucial role than the tonal distance in predicting distance judgement data. Below are the details.

When interpreting the results of the parameter estimates, if the population-level coefficient for segmental distance exceeds that of tonal distance, i.e. $\mu_\gamma - \mu_\delta > 0$, then we have strong evidence that individual segments are weighted higher than the tone, and vice versa. To see this, let us consider operations where a syllable changes to another syllable with the identical syllabic structure (e.g., [nip6] to [mit4], keeping onset-nucleus-coda-tone structure). The form of the model, ignoring random effects, the intercept and the Gaussian error term, would be $Y_{ij} = \mu_\delta \cdot (\mathrm{dist}_{\mathrm{onset}} + \mathrm{dist}_{\mathrm{nuc}} + \mathrm{dist}_{\mathrm{coda}}) + \mu_\gamma \cdot \mathrm{dist}_{\mathrm{tone}} = \mu_\delta \cdot \mathrm{dist}_{\mathrm{onset}} + \mu_\delta \cdot \mathrm{dist}_{\mathrm{nuc}} + \mu_\delta \cdot \mathrm{dist}_{\mathrm{coda}} + \mu_\gamma \cdot \mathrm{dist}_{\mathrm{coda}}$. Since μ_δ is multiplied to each of the three segment costs, we would expect that, the coefficients for tone and segment would be about the same, under the assumption that each segment and tone were equally weighted. Thus, strong evidence for $\mu_\gamma - \mu_\delta > 0$ suggests that segments are indeed weighted heavier than tones. We have found that the population-level estimates[13] of the coefficient of segmental distance (μ_γ) was higher than that of tonal distance (μ_δ); μ_γ is estimated at 1.50 (SE: 0.14, 95% CI: (1.23, 1.77)), which is around twice of μ_δ, estimated at 0.77 (SE: 0.22, 95% CI: (0.34, 1.19)). A 95% credible interval of the difference between the two ($\mu_\gamma - \mu_\delta$), as calculated by the brms package using posterior draws, is (0.25, 1.19) (point estimate: 0.72; SE: 0.24; evidence ratio that $\mu_\gamma - \mu_\delta > 0$: 570.43), indicating very strong evidence that segments are, on average, weighted heavier than tones.

3.2.3. Comparison of phonological distance measures

The full model was fit to all of logically possible combinations of segmental and tonal representations this study considers (see Section 2). The WAIC values were computed and compared for each of these models as in Table 3. We find that the lowest WAIC values, indicating the best model fit, were achieved with the Hamming distances between multivalued features for phonemes with the onset-contour-offset representation for tones (4682.1 in bold face in Table 3). When phoneme distance metrics themselves were concerned, the models with multivalued feature distance metrics using Hamming distances consistently showed the lower WAIC values, regardless of tonal representations (horizontal grey row in Table 3). On the tonal side, the representations with contour information (i.e., onset-contour, onset-contour-offset and contour-offset representations), consistently outperformed the other tonal distance measures, regardless of the segmental distances (vertical grey rows 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

[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 Supplementary Materials 3.

Multivalued (M)	4755.5	4778.8	4774.2	4770.5	4717.8	4697.4	4700.8
Multivalued (H)	4737.1	4759.4	4752.2	4753.7	4702.2	**4682.1**	4683.5

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

After applying information gain weighting to both the segmental and tonal distances, the results did not substantially improve. See Table 15-16 in Supplementary Materials for details. In fact, for the natural class-based distance, the WAIC values increased. This is consistent with Nerbonne and Heeringa's (1997) results.

For comparison, we also fitted a model that, instead of segmental and tonal distances on conceptual grounds, directly calculates acoustic distance from the audio recordings. We calculated them by obtaining cochleagrams of each of the recordings using Praat with the default parameters, then calculating the Euclidean distances between the cochleagrams. The problem of different numbers of samples was resolved similarly to the method described in Heeringa (2004).[14] This purely acoustic distance performed far worse than any of the phonological models in Table 3, at WAIC value 5070.2, showing that phonological knowledge is useful for determining distance judgements. A similar calculation using Mel frequency cepstral coefficients (Rabinet and Juang, 1993) performed even worse at WAIC 5097.3, and formant tracks (Heeringa et al., 2009) were the worst at 5120.5. Details of the acoustic measures were provided in Supplementary Materials 2.

3.2.4. Relative weightings of syllable components

To investigate relative weightings of syllable components, we fitted a model that separates the segmental distance into onset, nucleus, coda, and tonal distances. When fitting this model, we again used the natural class distance for segmental distance and onset-contour representation for tonal distance for consistency. As shown in Figure 4, the results show that onset and nucleus are weighted significantly higher than coda and tone. Specifically, the coefficients of onset, nucleus, coda, and tone were estimated at 1.80 (SE: 0.27; 95% CI: (1.30, 2.36)), 2.12 (SE: 0.29, 95% CI: (1.30, 2.36)), 0.68 (SE: 0.68, 95% CI: (0.20, 1.16), and 0.84 (SE: 0.84, 95% CI: (0.20, 1.16)) respectively. The difference between onset and nucleus and between coda and tone are respectively estimated at -0.32 (SE: 0.4, 90% CI: (-1.09, 0.45) and -0.15 (SE: 0.34, 90% CI: (-0.84, 0.49)), suggesting no significant differences. However, we have strong evidence that nucleus is weighted much heavier than coda, with an estimated difference of 1.44 (SE: 0.4, 95% CI: (0.67, 2.25)).[15]

[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.

[15] 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, 90% CI: (-0.98, 0.61) and -0.21 (SE: 0.37, 90% 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)).

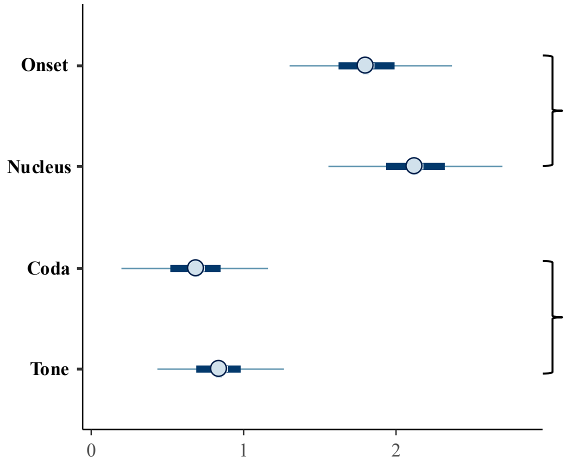

Figure 4: Estimates of the weightings of onset, nucleus, coda and tone along with 95% and 50% credible intervals.

To summarize the native speaker distance judgment results for monosyllables, we found that (1) segments are on average weighted heavier than tones, (2) the Hamming distances between multivalued features for phonemes with the tonal representations including contour representation performed best, and (3) onset and nucleus are weighted more than coda and tone when syllabic components are considered.

3.3. RESULTS FOR DISYLLABLES

3.3.1. Descriptive data

Scatterplots of the data in Figure 5 show a highly varied range of judgements among participants as in monosyllabic items; some participants are fully categorical judges while others are gradient judges (e.g., participant 19 vs. participant 3)[16] with different thresholds for maximal distance. For this reason, we have retained the same model as in the previous section. Again, the full model was found to have the best WAIC compared to reduced models. Therefore, our reports below are based on the full model (See 3.5-3.6 in Supplementary Materials for model specifications and model comparisons).

[16] Among disyllabic items, there seems to be a sharper discrepancy between the fully categorical judges and gradient judges, suggesting that a usual random-effects model with a (monomodal) Gaussian random slope may be insufficient. Thus, we attempted to model the situation by assuming that the tone and segmental distances come from a Gaussian mixture model with different means but we were not able to generate a model without divergent transitions in the MCMC chains.

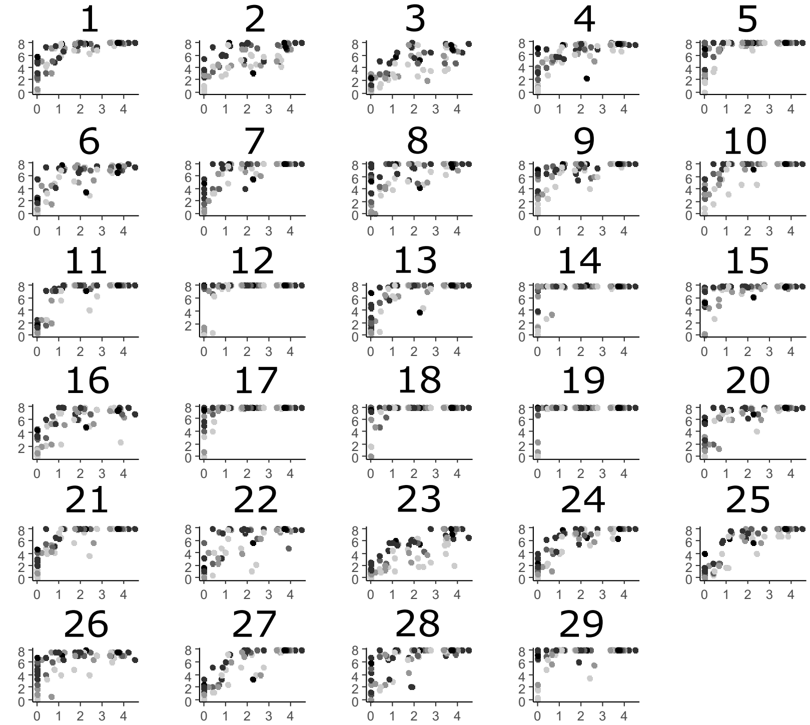

Figure 5: Scatterplots of distance judgements against theoretical segmental distance. Light grey points are those with tonal distance of 0; black dots have tonal distances of 2; intermediate shades indicate values in between the two extremes. Numbers indicate participants' numbers.

3.3.2. Relative weightings of segments and tones

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

3.3.3. Comparison of phonological distance measures

To compare different distance metrics, we applied the full model to all logically possible combinations of tonal and segmental distance metrics, as we have done with monosyllables. The results are in Table 4. As shown, the Hamming distance between multivalued features with contour-offset performed best (7153.0 in bold face in Table 4). As to the segmental distances themselves, the general tendency is consistent with monosyllables: multivalued feature representations were the best, especially with the Hamming distance (grey horizontal row in Table 4). It was additionally found that the phonology-based distances, such as natural class distance or Hamming distance on binary features, performed even worse than the baseline all-or-nothing distance among disyllables. Of the tonal distances, the contour-offset representation seems to perform well, but there is not much difference with the model using Hamming distances between Chao tone letters. 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 20-21 in Supplementary Materials.

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

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

Note that metrics that worked best are same for segments across monosyllables and disyllables, but they are different for tones: Different from monosyllables, the tonal representations with contour information failed to outperform the Chao tone letters in the modelling of 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 inter-syllable 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 very low denoted by '1' and high denoted by '4', then determined whether the average pitch was rising, falling or level. Then we added this to the onset-contour-offset representation. Finally, we determined the pitch level change between the two offsets and added the result to the contour-offset representation (C-O+: type 3 in Table 5). 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 inter-syllable 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+ representation, the two offsets are H and H, so the pitch-level change is level.

As shown in Table 5, the type 1 (O-C-+) did not result in much improvement, while the type 2 (avg O-C-O+) resulted in much lower WAICs than the original onset-contour-offset representation. The 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 judgements involved the Hamming distance based on multivalued features between the segment strings and the Hamming distance between the *modified* contour-offset representation of the tones reflecting the change in pitch level between the two syllables *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.

Again, purely acoustic 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 Supplementary Materials 2.5 for the details.

3.3.4. Relative weighting of syllable components

To explore the relative weights of syllable components among disyllables, we fitted a version of the model that separates segmental distance into onset, nucleus and coda distances, as we did for monosyllables. In our modeling, we assumed equal weighting of two syllables within an item; onset, nucleus, and coda in both syllables were treated equally. When fitting this model, we used the natural class distance for segmental distance and onset-contour representation for tonal distance, for reasons we have stated in 3.2.1. This makes the models comparable between monosyllables and disyllables.

The coefficient of onset, nucleus, and coda was estimated at 2.53 (SE: 0.43; 95% CI: (1.68, 3.36)), 1.38 (SE: 0.41, 95% CI: (0.51, 2.18)), and 0.68 (SE: 0.38, 95% CI: (0.18, 1.70) respectively, and that of tone was at 1.29 (SE: 0.25, 95% CI: (0.78, 1.78)). Based on posterior draws, the difference between onset and nucleus, nucleus and coda, and coda and tone weighting is estimated at 1.15 (SE: 0.68, 95% CI: (-0.2, 2.46)), 0.43 (SE: 0.62, 95% CI: (-0.81, 1.68)), and -0.35 (SE: 0.43, 95% CI: (-1.19, 0.48)) respectively. Clearly, we do not have strong evidence that the nucleus, coda, and tone differ in weighting. However, we do have weak evidence that onset is weighted heavier than nucleus, since a 95% credible interval is for their difference (0.02, 2.26).[17] The relative weightings of onset, nucleus, coda and tone are provided below:

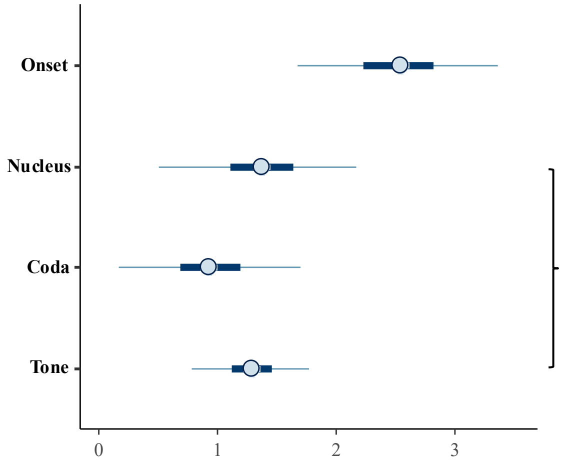

Figure 6: Estimates of the weightings of onset, nucleus, coda and tone along with 95% and 50% credible intervals.

To summarize, the results of native speaker distance judgments of disyllables are as follows: (1) neither segments nor tones were weighted significantly higher than the other; (2) the natural class distance between the segmental strings with the Hamming distance between the 'modified' contour-offset representation (i.e., the tones reflecting the change in pitch level between the two syllables as a whole) performed best; (3) onset is weighted heavier than other syllabic components, with other components weighted similar with each other.

[17] 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).

Table 6 below compares the results from monosyllables and disyllables.

	Monosyllables	Disyllables
Segment vs. Tone weighting	Seg > Tone	Seg ≈ Tone
Best distance metrics	Multivalued (Hamming) Seg + Contour tone	Multivalued (Hamming) Seg + modified contour tone
Syllabic components' weighting	Onset, Nucleus > Coda, Tone	Onset > Nucleus, Coda, Tone

Table 6. Distance judgments data of monosyllables and disyllables

3.4. Discussion

3.4.1. Tonal and segmental weighting

We compared relative weighting of segments and tones when judging phonological distance of words in a tone language, Cantonese. We provide evidence that segments are weighted heavier than tones in Cantonese monosyllabic words in measuring phonological distance. The results echo those from some other studies. Among studies investigating Cantonese, Cham (2003) compared the perception of Thai tones and segments by Cantonese-speaking children and adults by phonological awareness tests where participants selected an odd one from among three syllables. Cham found that Cantonese speakers outperformed in phone awareness tasks than in tone awareness tasks. Despite the different nature of the task performed from the current study, their results may imply that tones are perceptually less salient than phones for Cantonese speakers, and thus the higher weighting of phones for monosyllables in the current study should not be surprising. The current results, however, contrast with Yang and Castro's (2008) findings that segments were as important as tones in Zhuang and less important in Bai. Considering that Yang and Castro's main focus was phonological distance, 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 judgements vs. mutual intelligibility); there may exist typological difference in the relative weighting of tones and segments, which needs future research on cross-linguistic comparisons; or the coefficients in Yang and Castro's model (which they do not report) may not directly support their conclusion. Further investigations are needed to verify whether a cross-linguistic generalization can be drawn about the relative weights of tones and segments in measuring phonological distance of words and if so on what basis the weighting differences are driven.

Note though that our results of disyllables did not support those of monosyllables in our study; segments were *not* weighted heavier than tones in disyllables. We want to point out that we do not have strong evidence to the contrary either, since their 50% credible intervals do not overlap. A less clear pattern among disyllables can be attributed to the fact that the disyllabic test items 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.

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. 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 distance between [u] and [a], between H and M and between L and F summed up.

3.4.2. Metric comparisons

For segmental distances, we have demonstrated that multivalued features are better representations of phonemes for predicting distance judgements than binary distinctive features. It was also found that a purely acoustic or auditory measure of distances work far worse than any of the other features mentioned. This result can be interpreted in two ways. First, it is possible to speculate that articulation is most relevant to distance judgements. This is because most of the multivalued features are articulation-based; the binary distinctive features were designed with reference to articulation, but abstracted away from it (Chomsky and Halle, 1968); and the cochleagram or other purely acoustic representations had no articulatory component at all. This interpretation aligns with conclusions drawn by previous studies like Somers (1998) and Heeringa (2004), as well as a view in phonetics and phonology that speech perception involves processes also used in production (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967). Second, it is also possible to propose that a balance between phonetics and phonology, which is what the multivalued features provide, may be the best. Unlike the binary features, the multivalued features distinguish between allophones and allows for gradient features, but at the same time do not take into account minor, non-systematic 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 judgements in Cantonese, consistent with the results from the investigations of other tone languages by Yang and Castro (2008) and also those of Tang and van Heuven (2011). This also aligns with work in tone perception in Cantonese, where it is found that tonal directions are an important perceptual cue (e.g. Xu, Gandour and Francis, 2006; Khouw and Ciocca, 2007, *inter alia*), and indeed is sometimes found to be somewhat more important than tonal height (Gandour, 1981).[18]

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 Nerbonne & Herringa's (1997) results, which show distances between multivalued features without information gain weighting works best for determining dialect distance among different metrics using multivalued features. We want to note that the lack of effectiveness of information gain weighting does not necessarily imply that the features are equally weighted, since information gain is only one possible type of weighting scheme and potentially other theoretical or empirical schemes might improve the predictive power of phonological distance. We leave this for future research.

3.4.3. Relative weighting of Onset, Nucleus, and Coda

We further split segments into onset, nucleus, coda, and tone to investigate relative weightings of syllable components in phonological distance judgments. For monosyllables, we have shown that onset and nucleus are weighted heavier than coda and tone. The fact that the onset and the nucleus are found to be more important than the tone in the case of monosyllables may align with tonal perception studies, which show lower accuracy and longer response times in spoken word recognition when tone differences are involved, e.g. when a nonword and a word differ only in tone, when a distractor differs from the target only in tone, or when asked to discriminate between two syllables with only a tonal difference (e.g. Keung and Hoosain, 1979; Cutler and Chen, 1997), suggesting that tones contribute less to distinctions than segments.

For disyllables, onset is weighted heavier than nucleus, coda and tone. One possible intuitive explanation for a less important role of nucleus in the disyllable case is as follows. 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. Another possible reason is that vowels are

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

more important in monosyllables because of acoustic prominence while their saliency weakens in disyllables. 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; acoustic properties thus become a more decisive factor in monosyllables. By contrast, since at least one of the stimuli in each disyllable-disyllable pair is always an existing lexical word, the provided context may lead the 'vowel advantage' to disappear, consistent with the results from Ye and Connine's (1999) perceptual experiment, where the presence of context removes the 'vowel advantage'. However, a similar vein of research in the word reconstruction paradigm (van Ooijen, 1996; Cutler, Sebastián-Gallés, Soler-Vilageliu and van Ooijen, 2000) found that in Mandarin monosyllables, vowels are less mutable than consonants, contrary to previous results in non-tonal languages (Wiener and Turnbull, 2006) and our study. Considering that the word reconstruction paradigm necessarily involves lexical access, it may be the case that the acoustic prominence of the nucleus always plays a role in monosyllables, regardless of whether lexical items are activated or not. Further investigations are needed to determine the exact reasons behind the weightings.

4. PHONOLOGICAL DISTANCE AND LEXICAL PREDICTABILITY

A fundamental question we want to take on to Section 4 is to understand why speakers weigh certain syllabic components heavier than others in their phonological distance judgments. For example, why do Cantonese speakers rely more on onset than coda when judging phonological distance between two items? As hinted in the above section, we hypothesize that the relative weights of syllabic components (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 distances are fundamentally relevant to distinguishing between lexical items, so we predict that speakers may not rely heavily on lexically highly predictable elements in evaluating phonological distance. For example, if coda is easily predictable in the lexicon (e.g., coda is restricted to either lenis obstruents or nasals), native speakers will tend to downweight properties of coda in distinguishing two items due to its predetermined lexical properties. If so, high lexical predictability of coda will in turn be reflected in lower relative weights of coda in phonological distance judgments. This idea aligns with previous results in semantics where information content has been used in evaluating semantic distances in semantic networks to avoid weighting all edges equally (Resnik, 1995; Jiang and Conrath, 1997; Budanitsky and Hirst, 2001). Through a lexical analysis, this section employs two types of information-theoretic measures of syllabic components to analyze syllabic components' entropies and functional load. The results show a correspondence between the predictions from the lexical analysis and the relative weights of the syllabic components reported in Section 3.

4.1. ENTROPY ANALYSIS

A simple way of measuring the amount of uncertainty is entropy. Very roughly speaking, entropy is the quantity representing the lack of predictability. When calculated using base 2 logarithms, the formula for entropy is as follows:

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

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 this case, the entropy is a lower bound on the expected number of 'bits', i.e., 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:

(7) Language A: /a/ 50%, /u/ 25%, /i/ 25%
Language B: /a/ 50%, /u/ 25%, /i/ 12.5%, /o/ 12.5%

For 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. Correspondingly, a maximally efficient[19] way of encoding the nucleus of Language A in binary digits is to use '0' for /a/, '10' for /u/ and '11' for /i/; in this case the expected number of bits is $0.5 \times 1 + 0.25 \times 2 + 0.25 \times 2 = 1.5$, exactly matching the entropy. Similarly, for Language B, the entropy is $-0.5\log_2 0.5 - 0.25\log_2 0.25 - 0.125\log_2 0.125 = 1.75$, and correspondingly, maximally efficient method of coding is to use '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. As a crude measure of a component's importance, we may directly use entropy to predict weightings in distance judgements. The weights in the monosyllable and disyllable models are plotted below against estimated sample entropies. Recall that the order of the weights in the distance models was Onset, Nucleus > Coda, Tone for monosyllables and it was Onset > Nucleus, Coda, Tone for disyllables. As shown in Figure 7, the overall relationship seems quite strong for disyllables, but nucleus seems to be an outlier in the monosyllable case.

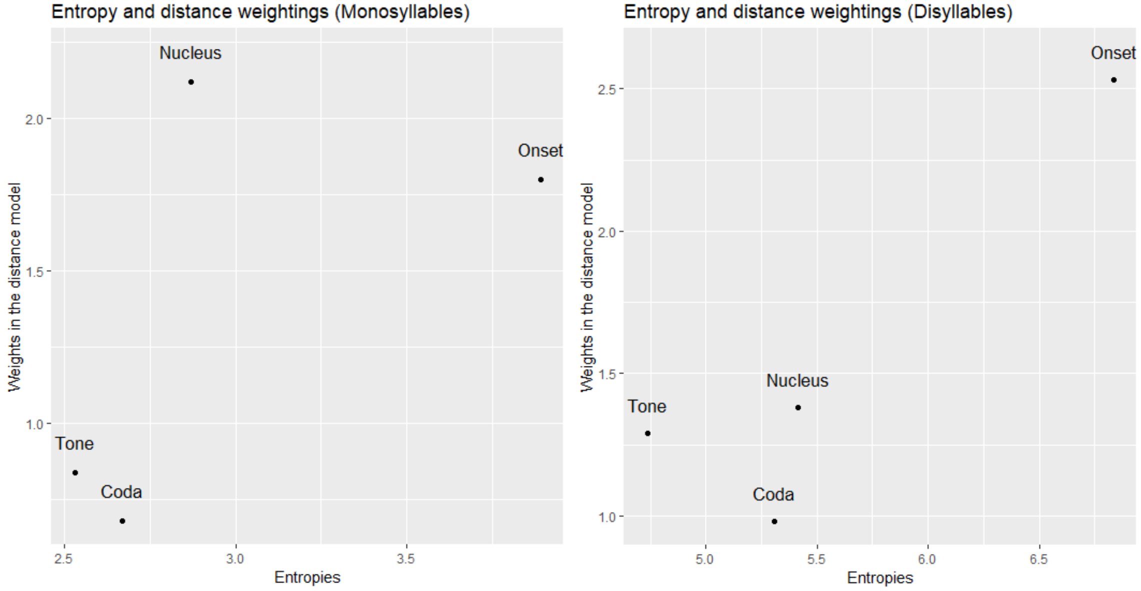

Figure 7: Relationship between point estimates of entropy and weighting in distance models.

The above point estimates of entropy, however, do not give us information about variability in the estimates. It is uncertain whether the differences between the entropies of the various syllable components here correspond to actual differences, and are not just artefacts of our sample. Thus, we computed confidence intervals[20] for the differences between the entropies to ensure that the differences are not simply due to sampling error. Since no standard formula is available for confidence intervals of differences between marginal entropy measures, we derived our own using the asymptotic properties of the probability estimates along with the delta method; details are given

[19] Here, 'maximally efficient' means that the expected number of bits needed is minimized, considering only prefix codes, i.e. no codeword (i.e. representation of an outcome) forms the first part of another codeword. For example, in Language A, '0' for /a/, '1' for /u/ and '11' for /i/ would have lower expected number of bits, but it is not a prefix code since '1' is a prefix of '11'. See Cover and Thomas (2006) for a more formal treatment of the topic.

[20] 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% of the time. 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.

in Supplementary Materials 3.1. We applied a Bonferroni correction with *g* = 6, so each of the monosyllable and disyllable estimates have at least 95% confidence as a whole.

In Figure 8, we report two types of estimates, point estimates and confidence interval estimates, in which the point estimates are in the middle of confidence intervals. As shown, the confidence intervals are all very narrow with the lower bounds far away from zero in most cases. This indicates that we have very strong evidence of the entropy differences. The entropy differences suggest that the entropies of the four elements are ranked onset > nucleus > coda > tone for both monosyllables and disyllables. Note that this is largely consistent with our findings of syllabic components' weights in phonological distance judgments: onset, nucleus > coda, tone for monosyllables, and onset > nucleus, coda, tone for disyllables.

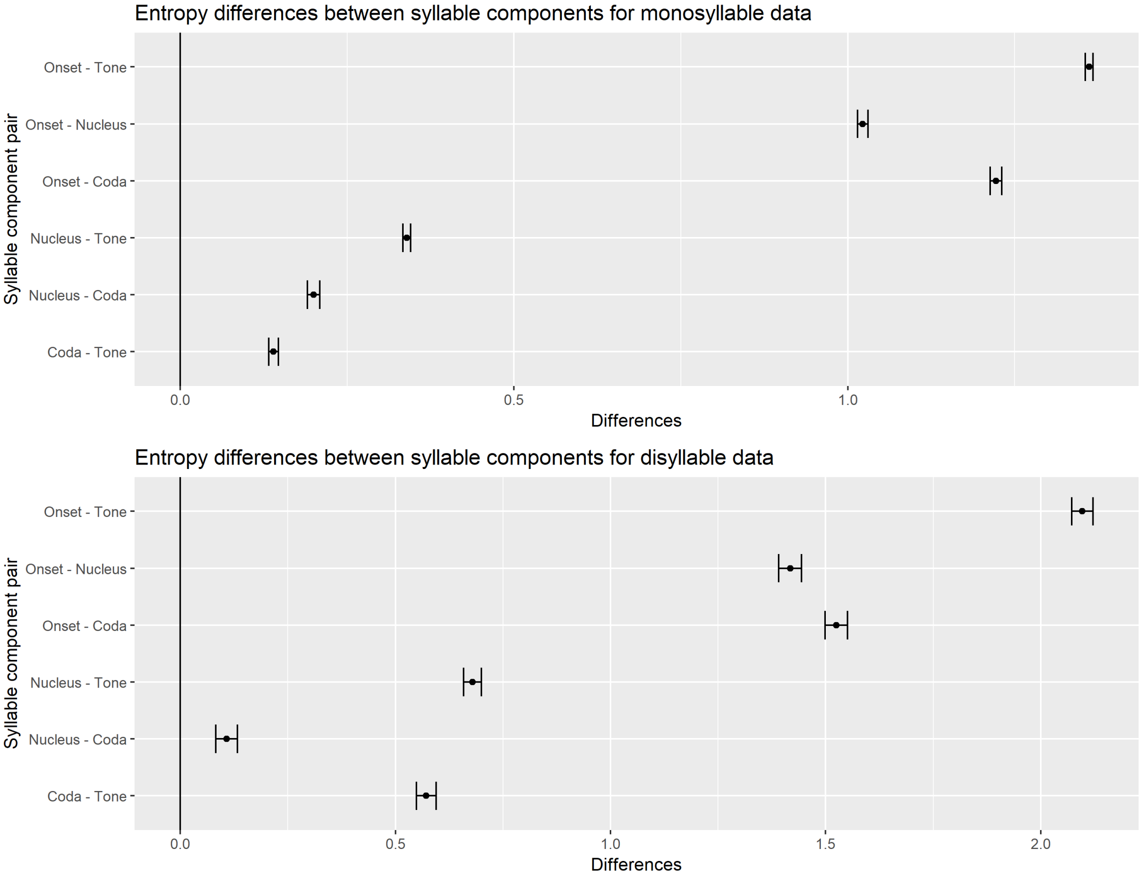

Figure 8: Point and interval estimates of the differences between the entropies of various syllable components. Point estimates are represented as circle dots whereas the two limits of the confidence intervals are indicated by short vertical lines.

4.2. FUNCTIONAL LOAD ANALYSIS

In the above calculations of (marginal) entropies, we do not take into account properties of the other three syllabic components when calculating each of their entropy (e.g., the lexical properties of onset, nucleus, and tone were ignored when calculating the entropy of coda). This may not be a desirable situation because of phonotactics. If two components of the syllable were strongly dependent, say we can completely predict the coda from the nucleus, then even if there were a huge *marginal* uncertainty in the coda, we would expect that the coda is less important because cues from the nucleus can fully determine the coda. An information-theoretic measure that takes this consideration into account is functional load, i.e., how important each component is in maintaining meaning 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 fictional language state L'_c where all contrasts in that component are neutralised (Hockett, 1966; Carter, 1987; Surendran & Levow, 2004; Oh, Coupé, Marsico and Pellegrino, 2015):

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

We computed functional loads for onset, nucleus, coda and tone, and plotted the estimated weights in the two distance models (monosyllables and disyllables) against the FLs:

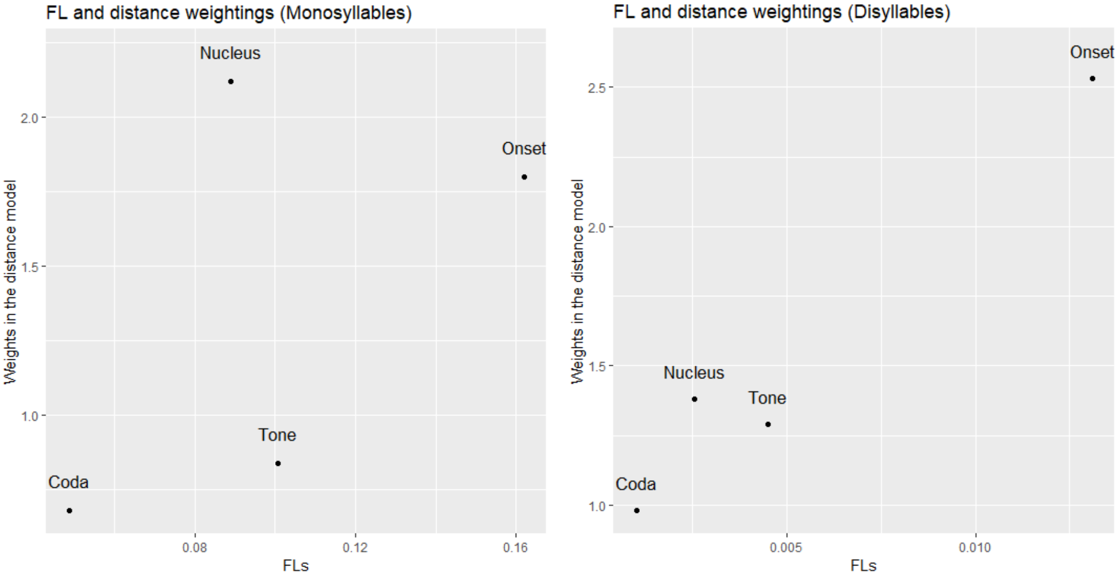

Figure 9: Relationship between point estimates of functional load and weighting in distance models.

When the results in Figure 10 are compared with the order of the weights in the distance models (Onset, Nucleus > Coda, Tone for monosyllables; Onset > Nucleus, Coda, Tone for disyllables), again nucleus does not show a good match between the judged weights and its functional loads for monosyllables. Except for that, the overall relationship is relatively strong for disyllables. This result is similar to our observation of the judged weights in the distance models against entropies.

As we did for entropies, we additionally 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). 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. For both monosyllables and disyllables, when the differences among pairs are concerned, the entropy hierarchy should be onset > tone > nucleus, coda. Crucially, tone is in fact slightly more important than nucleus and coda, which are relatively close to each other. This is in contrast with simple entropy calculations, where the predicted hierarchy was onset > nucleus > coda > tone. It 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 lower functional load, whereas marginal entropies only look at each individual component, and are therefore not affected by such phonotactic factors. Importantly, the results again roughly match with our phonological distance judgments data in that onset is at the top of the hierarchy.

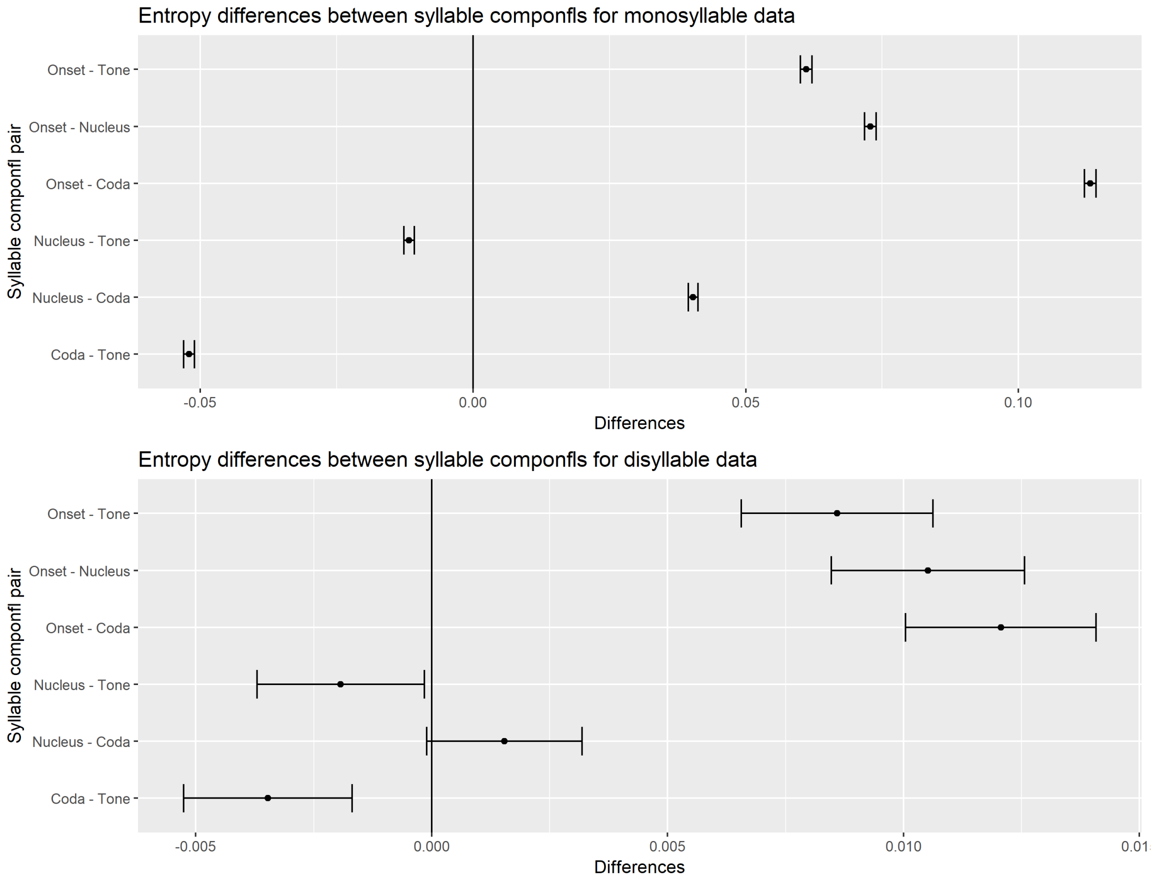

Figure 10: Point and interval estimates of the differences between the functional loads of various syllable components. Point estimates are represented as circle dots whereas the two limits of the confidence intervals are indicated by short vertical lines.

From the examinations of entropies, we would expect the weight hierarchy of onset > nucleus > coda > tone, reflecting the relative sizes of their entropies. From the examination of functional loads, we would expect the weight hierarchy of onset > tone > nucleus, coda. Considering that our modeling results overall match the predictions from entropies and functional loads, we conclude that information-theoretic predictability or functional load has 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. Some other explanation, such as those that we discussed in Section 3.4.3, may be necessary. We will leave it to future studies to figure out what additional components other than items' lexical predictability contribute to phonological distance judgments, and how such components are interacting with the lexical predictabilities.

5. DISCUSSION AND CONCLUSION

This study showed that tones and segments are weighted differently by native Cantonese-speaking participants when making phonological distance judgments. It further showed that onset is consistently weighted heavier than coda and tone (though the role of nucleus is relatively unclear), and that these weighting 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 with 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 Cantonese, our modelling work has shown how to set up and find optimal measures of phonological distance that can predict native speakers' judgements. 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 languages. Even for tonal 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 we consider 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 the 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.

1 References

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

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

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

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

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

Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1. 05) [Computer program].

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

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

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

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

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

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

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

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

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.*

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

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

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

Cutler, A., & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics*, *59*(2), 165-179.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & van Ooijen B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, 28(5), 746–755.

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

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

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

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

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

Gandour, J. (1981). Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics*, 20-36.

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

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

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

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

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

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

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

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

Heeringa, W., Johnson, K., & Gooskens, C. (2009). Measuring Norwegian dialect distances using acoustic features. Speech Communication, 51(2), 167-183.

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

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

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

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

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

Keung, T., & Hoosain, R. (1979). Segmental phonemes and tonal phonemes in comprehension of Cantonese. *Psychologia: An International Journal of Psychology in the Orient.*

Khouw, E., & Ciocca, V. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics*, *35*(1), 104-117.

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

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

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

Ladefoged, P. (1971) *Preliminaries to Linguistic Phonetics.* Chicago: University of Chicago Press.

Ladefoged, P. (1975). *A course in phonetics.* New York, NY: Harcourt Brace Jovanovich, Inc.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100, 1989–2001.

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

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, *19*(1), 1.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & psychophysics*, *62*(3), 615-625.

Luke, K. K., & Wong, M. L. (2015). The Hong Kong Cantonese corpus: design and uses. Journal of Chinese Linguistics, 25(2015), 309-330.

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

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

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

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

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

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

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

Rabiner, L. R., & Juang, B. H. (1993). Fundamentals of speech recognition. Prentice Hall.

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

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

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

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

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

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

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

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

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

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

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

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

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

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, 59(1), 59–82.

Wong, A. W. K., & Chen, H. C. (2008). Processing segmental and prosodic information in Cantonese word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1172.

Ussishkin, A., & Wedel, A. (2002). Neighborhood density and the root-affix distinction. In *PROCEEDINGS-NELS* (Vol. 32, No. 2, pp. 539-550).

van Ooijen, B. (1996). Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. *Memory & Cognition*, 24, 573–583.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, *40*(3), 211-228.

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

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

Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, *120*(2), 1063-1074.

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

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

Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes*, *14*(5-6), 609-630.

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

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

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