Measuring phonological distance in a tonal language: An experimental and computational study with Cantonese

Youngah Do and Ryan Ka Yau Lai
University of Hong Kong
{youngah, kayaulai}@hku.hk

Measures of phonological distance between words are widely used in different fields of linguistics, such as phonology, psycholinguistics, historical linguistics and dialectology. Various studies have compared the quality of different phonological distance measures (e.g. Nerbonne & Heeringa, 1997), but to the best of our knowledge, the only study to do so incorporating tonal information is Yang and Castro (2008); they examine association between the mutual intelligibility of Bai and Zhuang dialects and the segmental and tonal distances between them. Yang and Castro also look at the weighting of tone and segments in determining the intelligibility.

Our current study aims to investigate the following questions about phonological distance in Cantonese. First, we investigate the relative weighting of tonal and segmental distances in determining phonological distance, as well as their interpersonal variation, by constructing Bayesian multilevel models (Nicenboim & Vasishth, 2016). Second, we aim to assess the relative quality of various tonal and segmental distances in the context of Cantonese. While previous studies on phonological distances of tonal languages typically assess the quality of the measures in the context of genetic relationships or intelligibility between dialects, we base our analyses directly on distance judgements from native speakers. Finally, we determine whether different parts of the syllable (onset, nucleus, coda) may also be weighted differently. As this study is part of an ongoing project to model Cantonese phonotactics, the results will be used in a generalized neighbourhood model (GNM) (Bailey & Hahn, 2001).

Among measures of segmental distance, we used the Hamming distance between binary feature vectors of phonemes, the proportion of unshared natural classes between two phonemes (Frisch, Broe and Pierrehumbert, 1997), as well as Hamming, Manhattan and Euclidean distances between multivalued feature vectors of phonemes, based on the phonetically-motivated feature matrix for English in Ladefoged (1975). The distances were scaled to fall in the interval [0, 1] where necessary. We then computed the phonemic distance between words with the Wagner-Fischer algorithm using these phonemic distances as the substitution cost and 0.5 of the average substitution cost as the indel cost. As for tonal distances, we examined five of the six representations of tone discussed by Yang and Castro, including the autosegmental, Chao tone letter, onset-contour (O-C), onset-contour-offset (O-C-O), and contour-offset (C-O) representations. We then computed the Hamming distances between them. As Chao tone letters can also be construed numerically as pitches (i.e. 5 is the highest pitch, 1 is the lowest pitch and 51 would represent a high falling tone), we also computed Euclidean and Manhattan distances between them. For each of these distances, we created a version with weighting based on information gain (Nerbonne & Heeringa, 1997). In the case of binary distinctive features, we tried a version with Broe's (1996) modified formula, which takes into account the existence of null values in binary features.

To determine Cantonese speakers' mental perceptions of phonological distance, we conducted an experiment using the online survey website Qualtrics. Our experiment consisted of 144 items, including 72 monosyllabic and 72 disyllabic ones. Each item consists of a pair of Cantonese pairs of words (e.g. *bei*^{2} vs *be*^{1}). The first word is always an existing word, whereas the second word may be a nonce word. We chose items varying all existing tones and phonemes to ensure that the distance between the pairs are well spread across the possible space of distances, and that segmental and tonal distances are uncorrelated. The words were recorded by a native speaker of Cantonese. For each item, we asked participants to rate the similarity of the two syllables on a scale of 0 to 100 by dragging a bar on the screen. The similarities were then converted into distances by subtracting each similarity rating from 100.

Before constructing our models, to enhance interpretability, the distance judgements were scaled

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 371-372.
New York City, New York, January 3-6, 2019

to fall in [0, 4] for monosyllables and [0, 8] for disyllables, since Cantonese syllables contain only up to three phonemes, and hence the maximum segmental and tonal distances sum up to 4 and 8 for monosyllables and disyllables respectively. We then constructed the Bayesian model with the following likelihood specification:

(1) $Y_{ij} \sim N(\mu + \alpha_i + \beta_j + \gamma_j t_i + \delta_j s_j, \sigma^2)$

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

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

where α_i, β_j $\gamma_j$and δ_j are the item-level intercept and the subject-level intercept, segmental weighting and tonal weighting respectively. Moreover, visualisation of the data suggested that the distances may be treated as right-censored, i.e. the underlying distance may go above the maximum, but is truncated to 4 or 8 if this occurs. By fitting this full model in the R package brms using default uninformative and weakly informative priors, along with various reduced models, we determined that the full model is optimal using the Widely Applicable Information Criterion (WAIC) (Nicenboim & Vasishth, 2016). We then fitted the model to different measures of segmental and tonal distance and compared their WAICs. Unlike Yang and Castro's approach of computing simple correlation coefficients, our approach allows for interpersonal variability and simultaneous comparison of tonal and segmental distance.

We found strong evidence that on average, segments are weighted heavier than tone for monosyllables (95% CI of $\mu_\gamma - \mu_\delta$: (0.25, 1.19)), but no such tendency was found from among disyllables (95% CI: (-0.23, 0.91)). It was found that adding random slopes greatly improved our model WAIC, which suggests substantial interpersonal variation in the weightings.

Of the tonal representations, O-C, O-C-O and C-O representations were the best metrics for predicting monosyllable judgements, but their quality resembled that of Chao tone letters for disyllables. After extending the O-C-O and C-O representations to indicate change in pitch between the two syllables, however, C-O stood out as the best representation in the disyllabic case. This is consistent with Yang and Castro's findings. Our results suggest that pitch contours are important for determining phonological distances, since the representations that do not represent contours (autosegmental and Chao tone letter representations) fared worse. Of the segmental distances, Hamming distances between articulatorily-based multivalued features fared best. However, a pure acoustic distance fared much worse than any of the phonological distances, suggesting that a balance between phonological abstraction and phonetic detail is needed.

Finally, separating onset, nucleus and coda distances were found to slightly improve WAIC for monosyllables, though not for disyllables. Onsets are found to be weighted much heavier than codas and tones, while nucleus weighting was similar to onset weighting for monosyllables and to coda weighting for disyllables. The results can only be partially explained by differences in entropy or functional load (Hocket, 1966).

References

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

Paul Christian Bürkner. 2017. brms: An R Package for Bayesian Multilevel Models using Stan. *Journal of Statistical Software*, 80(1):1-28.

Michael Broe. 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*. Association for Computational Linguistics, pages: 17-24.

Stefan Frisch, Michael Broe, and Janet Pierrehumbert. 1997. Similarity and phonotactics in Arabic. *Rutgers Optimality Archive*, 223.

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

Peter Ladefoged. 1975. A Course in Phonetics. Harcourt Brace Jovanovich, Inc., New York.

John Nerbonne & Wilbert Heeringa 1997. Measuring dialect distance phonetically. In *Computational Phonology: Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*.

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

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