Distributional Learning Across Contexts: Learning Cantonese Tones in Naturalistic Speech

Keywords: Distributional learning, Cantonese, Tones, Variability, Naturalistic speech
Motivation: Infants initially discriminate most sound contrasts but quickly attune to those of their native language. This raises the question: how do infants identify the relevant acoustic dimensions for learning phonetic categories? The **distributional learning** account proposes that infants track the distribution of sounds, and identify acoustic dimensions as contrastive if their distribution has two or more distinct peaks (i.e. multimodal distributions) [1]. However, while multimodality appear in controlled experiments, they are rarely found in naturalistic, highly variable speech, suggesting that multimodality is not a reliable way to identify contrastive dimensions [2]. Recent work comparing languages with/without vowel length contrasts suggests that even without multimodality, contrastive dimensions show more **contextual variability**: when a dimension is contrastive, the shape of its distribution will vary more across contexts [3]. The **distributional learning across contexts** hypothesis proposes that infants utilize this contextual variability to distinguish phonetic categories. This study tests this hypothesis by examining Hong Kong Cantonese tones, exploring whether ease of acquiring different tonal contrasts is linked to their contextual variability in distribution shape. Cantonese serves as a valuable test case due to the overlapping acoustic distributions between its six tones: high-level (T1), high-rising (T2), mid-level (T3), low-falling (T4), low-rising (T5), and low-level (T6).
Methods: We analyzed the Multi-ethnic Hong Kong Cantonese Corpus (MeHKCC) [4], which consists naturalistic speech recordings from 24 native Cantonese female speakers. 65,106 monosyllabic and disyllabic content words were extracted. Pairwise F0 contour comparisons showed varying acoustic overlap among tones, except for the phonetically distinct T1 (e.g., distinct pair: T1T4, overlapping pairs: T3T5, T2T5; see Fig.1). Based on acoustic overlap and documented acquisition difficulty [5], tone pairs were categorized into: (1) *Easy* pairs, which are phonetically distinct and easy to learn (e.g., T1T4); (2) *Hard* pairs, which are acoustically overlapping but learnable (e.g., T3T5); and (3) *Merger* pairs, which are acoustically overlapping and challenging to learn (e.g., T2T5). We predict that contextual variability in distribution shapes aligns with developmental acquisition patterns, with *Easy* contrasts showing the most separation and variability, followed by *Hard* and *Merger* pairs. Although *Hard* and *Merger* pairs both show acoustic overlap, we predict that *Hard* pairs are more learnable due to the greater contextual variability in their distributional shapes.

To test this, nine F0 landmarks (mean, median, variance, max-min, onset, 25%, 75%, offset, duration) were extracted, and t-distributed stochastic neighbor embedding (t-SNE) was used to reduce these dimensions to a 2D space. Distributional differences were quantified using Earth Mover's Distance (EMD) for pairwise tone comparisons across contexts. Contexts were defined as combinations of (1) neighboring sounds (e.g., stops, fricatives, nasals), (2) syllable position in a word (i.e., first or second syllable in a word), and (3) prosodic position (i.e., utterance-initial, -medial or -final).
Results: Analyses were conducted for all tone pairs, with T1T4 (*Easy*), T3T5 (*Hard*), and T2T5 (*Merger*) selected for illustration. Fig. 2 shows the frequency distribution of the *Hard* tone pair T3T5 after dimensionality reduction. While tone pairs show unimodal distribution when pooled across contexts (Panel A), they show different distribution shapes across specific contexts (Panel B shows two illustrative contexts). Figure 3 presents a boxplot of EMD for the three tone pairs, where each data point represents the pairwise EMD of two tones within a single context. Higher mean EMD values indicate greater distributional separation in general, while higher variance across contexts reflects greater contextual variability. Across four EMD metrics—mean, median, variance, and maximum—*Easy* pairs consistently show the highest values, followed by *Hard* pairs, and then *Merger* pairs (values provided in Fig. 3). This hierarchy aligns with developmental acquisition patterns: tones with greater separation and contextual variability are learned more readily than tones with lower values. Analyses of all 15 tone pairs reveal similar trends, with more nuanced interactions between distributional learning across contexts and acoustic realizations.
Discussions: This study explored the learning of multiple tone contrasts, a relatively unexplored area in distributional learning. Findings suggest that infants may rely on distributional shapes across contexts to learn contrasts, offering a plausible mechanism for learning in the absence of invariance in speech signals. Future direction will expand to additional corpora with additional contexts, and develop computational learning models to quantitatively capture the learning trajectories of all tone pairs.

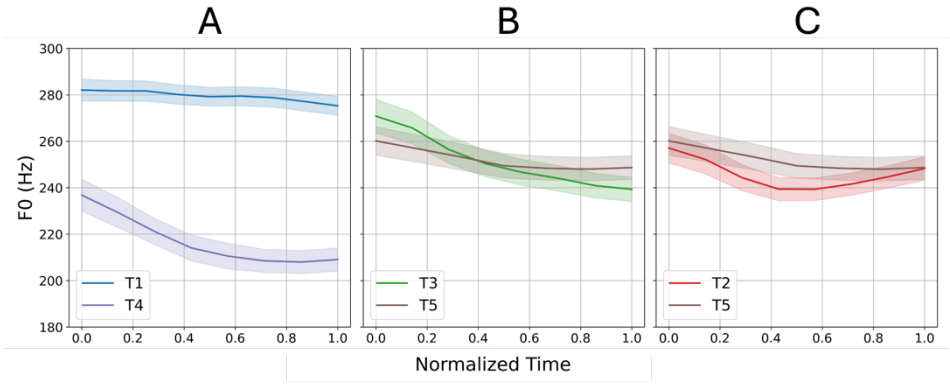

Figure 1. Pairwise F0 contours for three example tone pairs from a female speaker, with mean and 95% confidence intervals, time-normalized. (A) T1T4 shows clear phonetic distinction, while (B) T3T5 and (C) T2T5 exhibit varying degrees of overlap.

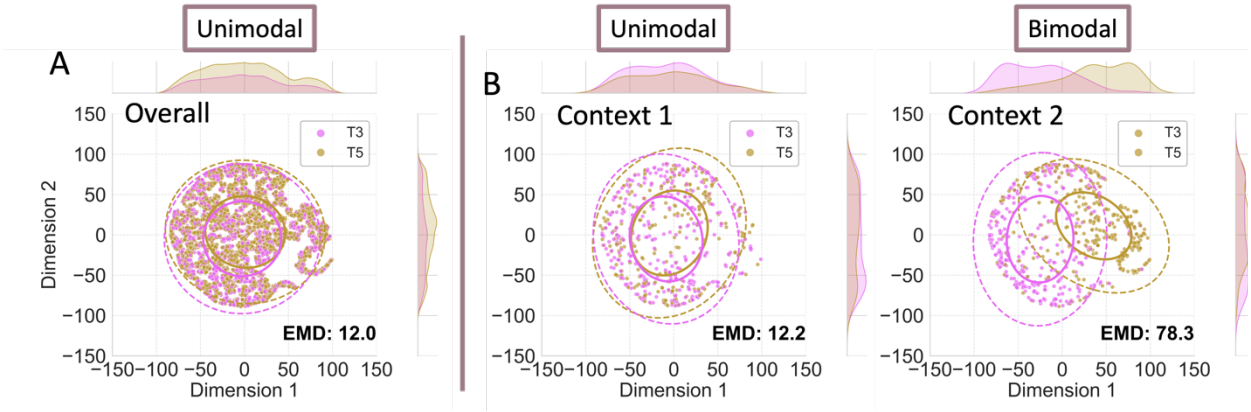

Figure 2. (A) The overall frequency distribution of the *Hard* tone pair T3T5 along a reduced 2D space shows a unimodal distribution, despite T3 and T5 being contrastive tones. (B) Frequency distributions for the same tone pair across two specific contexts reveal different distributional shapes. Context 2 (e.g., nasal onsets without codas, second syllable, utterance-medial) exhibits greater separability (EMD = 78.3) compared to Context 1 (EMD = 12.2). Even though the overall frequency distribution is unimodal, we can see differently shaped distributions across context.

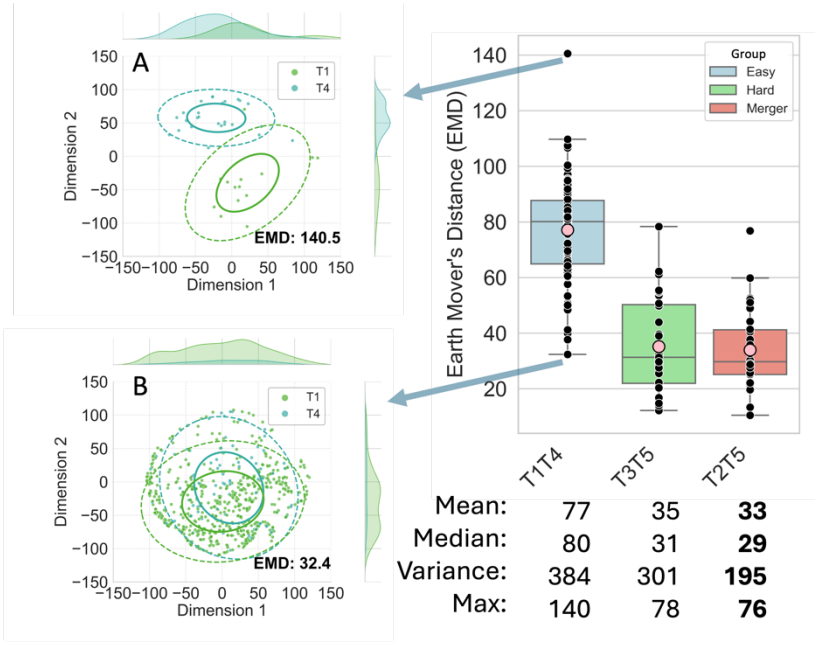

Figure 3. Earth Mover's Distance (EMD) distributions for three tone pair categories: Easy (T1T4), Hard (T3T5), and Merger (T2T5). Each data point represents the pairwise EMD between two tones within a specific context. Panel A illustrates a context with the greatest separability, while Panel B shows a context with the lowest separability. Lowest values for four EMD metrics (mean, median, variance, and maximum) were bolded.

References: [1]Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82(3), B101-111. https://doi.org/10.1016/s0010-0277(01)00157-3 [2]Bion, R. A. H., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning Phonemic Vowel Length from Naturalistic Recordings of Japanese Infant-Directed Speech. PLOS ONE, 8(2), e51594. https://doi.org/10.1371/journal.pone.0051594 [3]Hitczenko, K., & Feldman, N. (2022). Naturalistic speech supports distributional learning across contexts. https://doi.org/10.1073/pnas.2123230119 [4] Yu, A., Delisle, N., Martin, N., Zhang, V., Yao, Y., & To, C. (2024). The Multi-ethnic Hong Kong Cantonese Corpus. CorpusPhon satellite workshop at LabPhon19. https://ccds.edu.hku.hk [5] Mok, P. P. K., Fung, H. S. H., & Li, V. G. (2019). Assessing the Link Between Perception and Production in Cantonese Tone Acquisition. Journal of Speech, Language, and Hearing Research, 62(5), 1243–1257. https://doi.org/10.1044/2018_JSLHR-S-17-0430