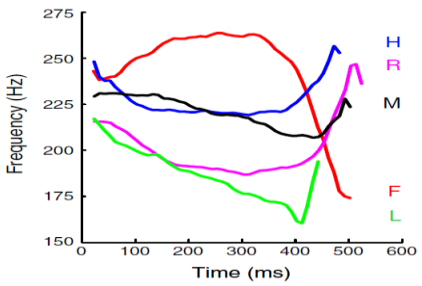
Probabilistic Modeling of Tone Perception: Autosegmental 'targets' are insufficient

Tone languages use pitch contrastively. For example, Thai has 5 contrastive tones, namely Mid, High, Low, Falling and Rising. The latter two ('contour' tones) show greater ranges of f0 values, while the first three ('level tones) show smaller ranges of f0 values.



Mid	na:	rice field
High	ná:	aunt
Low	nà:	custard apple
Falling	nâ:	face
Rising	nǎ:	thick

Fig.1: Examples of f0 pitch tracks on 5 Thai words with identical segments; single speaker in citation form

Extant models of tone perception appear too simplistic involving point-by-point matching of f0-value strings of a stimulus to category templates (e.g. Gauthier et al., 2007 for Mandarin) and are not motivated by phonological theory. Such models contain fine grained representation of tone; i.e. a string of f0 values for each tone category. Perception by these models therefore becomes a point-by-point matching of the stimulus with a set of category template f0 strings. The general problem for speech perception, namely variation, applies in the case of tone perception as well: naturally produced instances of tones rarely match 'canonical' f0 values and there is significant amount of variation even within a single speaker's utterances.

Importantly, another factor that such models do not take into consideration, is that tones interact in predictable ways like phonemes do, for which categorical, autosegmental phonology representations seem to be well-suited. The present study illustrates a model of tone perception informed by phonological theory, that aims to categorize stimuli correctly (i.e. matching human performance) despite the above mentioned variations. I shall briefly describe the phonological representation adopted, and how it is instantiated in the models tested here.

Autosegmental phonological representations of tone involve associating High(H) and Low(L) targets with syllables of a word. Combinations of these targets on a syllable result in either contour tones (multiple dissimilar targets) or level tones (single or same targets). For Thai, Morén and Zsiga (2006) present a further elaboration of the general approach mentioned above by proposing that the principal perceptual cues, i.e. tone targets, are associated with syllable moras as shown in Table 2.

Mid		High		Low		Falling		Rising	
		H		L		H	L	L	H
		|		|		|	|	|	|
μ	μ	μ	μ	μ	μ	μ	μ	μ	μ

Table 2: Morén and Zsiga (2006) representation of tone targets in Thai

In Thai, this moraic representation elegantly accounts for the fact that contour tones occur only on bimoraic syllables. Additionally, since bimoraicity supports at most two targets, and Thai syllables are maximally bimoraic, the non-existence of more complex tone trajectories in Thai is also explained.

In the present work, the Morén and Zsiga (2006) representation is adopted to build probabilistic models of tone perception. As illustrated in Table 2 above, the representation consists of tone targets associated with the right edge of moras; only the second mora in the case of the level tones, and both moras in the case of contour tones. In the models proposed here, this is translated as coding pitch values of the high and low extrema as equivalent to these theoretical H and L targets.

Certain additional assumptions in modeling this representation were necessary. While the current model requires information from all the tone targets and their (approximate) alignments to the syllable in order to categorize the tones, other experimental evidence (e.g. Lai and Zhang, 2008 with Mandarin), indicates that categorization can occur before the entire signal is perceived, and in fact makes use of initial f0 information. Hence, the first addition made was encoding each tone's initial point. Regarding the second addition, as per the representation in Table 2 above, contour and level tones have different 'profiles'; i.e. level tones do not have a middle target, while contour tones do. In the model, level tones are encoded with a 'virtual' middle target, implying that despite the absence of a target in the middle of the syllable for level tones, listeners have knowledge of the f0 halfway through the syllable, making early categorizations possible. Hence all tones are represented by sets of three target f0 values corresponding to initial, middle and final points.

Two probabilistic models were tested. These models add statistics to the representation described above, since the representations abstract away from variations. At present, the models are trained and tested on data drawn from a single speaker's productions (in Zsiga and Nitisaroj, 2008), but can be extended for multiple speakers. Also, syllable lengths are normalized.

The first model computes the mean for each target for each tone and a variance matrix over each set of three targets based on a training set of tones represented by three target values. The model can hence compute probability distributions over the targets. The variance matrix relates the three target distributions, allowing for the first target(s) to be informative about later ones, so that in principle, early categorizations are possible. In the first probabilistic model, the probability of the stimulus, given a particular tone category can be computed for each category. This can be done by multiplying the probabilities of each of the three targets, given a particular category. Hence for a stimulus 'x' and a category 'k', the probability of x given k can be written as:

(3) $\mathbf{p(x|category\ k) = p(x_{\alpha1}|\mu_{Catk\alpha1}, \sigma_{Catk\alpha1}) \cdot p(x_{\alpha2}|\mu_{Catk\alpha2}, \sigma_{Catk\alpha2}) \cdot p(x_{\alpha3}|\mu_{Catk\alpha3}, \sigma_{Catk\alpha3})}$
where $\alpha1$, $\alpha2$ and $\alpha3$ are the three target points.

From Bayes Theorem, we know that

(4) $\mathbf{posterior\ probability \propto likelihood \times prior}$
$\mathbf{p(category\ k|x) \propto p\ (x|category\ k) \times p(category\ k)}$

p(x|category k) is the conditional probability of stimulus x given category k, or the effect of the observed data, or the likelihood function. This is relativized (normalized) to the probability of the stimulus given all categories (Clayards et. al. 2006). For Thai tones, p(x|category k) is normalized by:

(5) $\mathbf{p(x|categoryF)+p(x|categoryR)+p(x|categoryH)+p(x|categoryM)+p(x|categoryL)}$.

Hence, the probability of each of the categories given stimulus x is computed (priors assumed to be equal) as shown in (6):

(6) $$p(\text{category k}|x) = \frac{p(x|\text{category k})}{p(x|\text{category F})+p(x|\text{category R})+ p(x|\text{category H})+ p(x|\text{category M})+ p(x|\text{category L})}$$

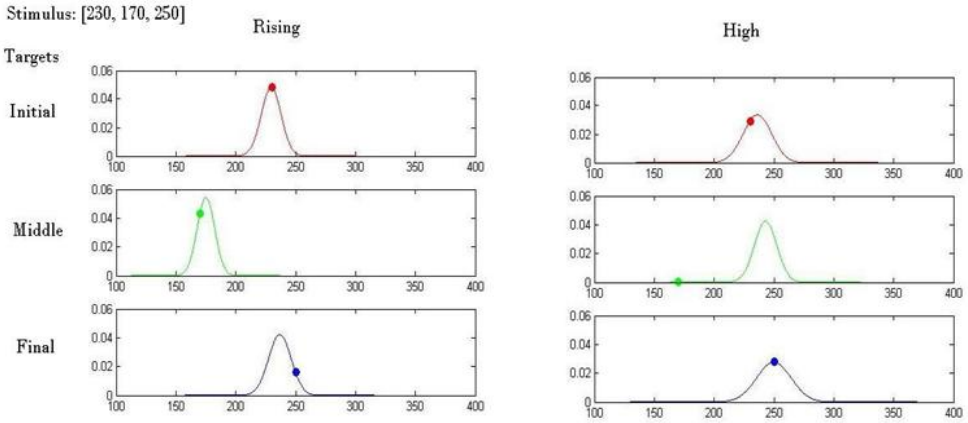


Fig.7: (i) Probability distributions for Rising (ii) Probability distributions for High
The red, green and blue dots represent where stimulus targets [230,170,250] lie on the probability distributions of the initial, middle and final targets for (i) Rising and (ii) High tones. For each category, probability of the stimulus given that category can be calculated by multiplying probabilities of each target given category target distribution.

The model was fit to the learning data, i.e. the model was trained on stimuli sets derived from naturally produced tones (i.e. the target values were gotten from naturally produced data of a single speaker), and the model was first tested on the same. Table 8 shows the model's performance.

Stimuli / Categorizations	Falling (12)	Rising (8)	High (12)	Mid (8)	Low (14)
Falling	**12**				
Rising		**8**			
High			**10**	1	
Mid			2	**6**	2
Low				1	**12**

Table 8: Model's fit to the learning data

The model was next tested on synthetic stimuli used in behavioral experiments in Zsiga and Nitisaroj (2008), in which human subjects categorized synthetic stimuli that consisted of linear interpolations between fixed sets of idealized pitch values at initial, middle and final points. A comparison of the human subjects' performance and the model's performance is illustrated in Figure 9.

The model's classification performance matches human subjects' classification performance 36.36% of the time. This indicates that only 'target' information is insufficient to account for the human subject's performance.

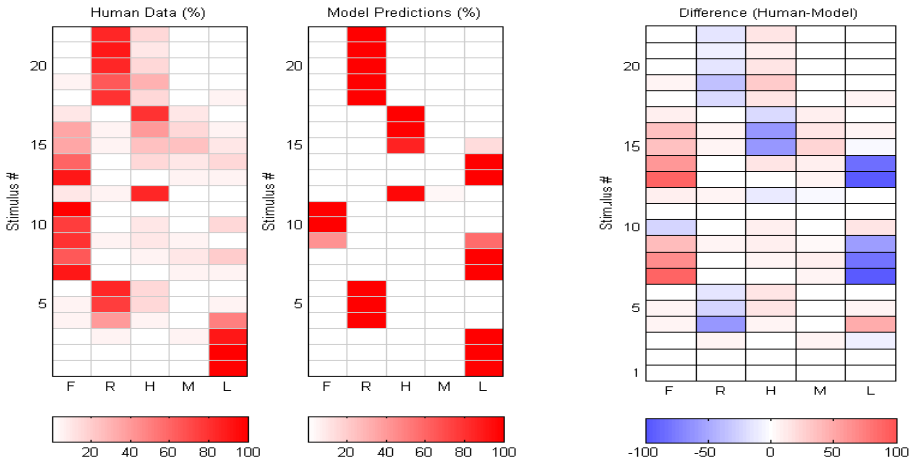


Fig.9: Human and model categorizations on the various stimuli, and their difference in performance.

Since these are dynamic stimuli, the rate of change of pitch is an important cue for perception. Hence, another model was trained on probability distributions over the slopes between adjacent tone targets (hence, 2 distributions), instead of distributions over the targets themselves, and categorization in this model is based on the slopes between the target points (p.c. Liberman). When this model was tested on the training data, it was found to be 90.7% correct. When tested on the steeply sloped synthesized stimuli used in Zsiga and Nitisaroj (2008), it was found to be 63.6% the same as human responses. Though an improvement, slope information alone does not accurately characterize human performance.

The main conclusions from this study are highlighted here. While autosegmental representations are valuable (good model performance), they do not fully characterize human perception. Information about o*nly target* or *only slope* is insufficient to model human tone perception. Thus generative properties of tone representation used in online perception are not limited to one form of cue information or the other (i.e. target or slope).

Further investigation of properties of cue distributions may provide important evidence of the use of statistical regularities in human perception. The effects of certain types of cue manipulations on perception are currently being investigated using experimental and computational approaches (Ramadoss, in progress).

Acknowledgments: I would like to thank Colin Wilson, Luigi Burzio, Elizabeth Zsiga, Rattima Nitisaroj, Paul Smolensky and Srivatsun Sadagopan for their valuable input.

References:

Clayards M., Tanenhaus M.K., Aslin R., Jacobs R.A. (2008). "Perception of speech reflects optimal use of probabilistic speech cues." Cognition.

Gauthier B., Shi. R., and Xu Y. (2007). "Learning phonetic categories by tracking movements." Cognition 103: 80-106.

Lai, Y. and Zhang, J (2008). "Mandarin Lexical Tone Recognition: The Gating Paradigm." Kansas Working Papers in Linguistics 30: 183-194.

Moren B, a. Zsiga. E. (2006). "The lexical and post-lexical phonology of Thai tones." Natural Language & Linguistic Theory 24: 113-178.

Ramadoss, D. (in progress). PhD Dissertation. Johns Hopkins University

Zsiga, E., and Nitisaroj, R. (2007). "Tone Features, Tone Perception, and Peak Alignment in Thai." Language and Speech 50(3): 343-383.