Interspeech 2018
2-6 September 2018, Hyderabad

Analyzing Thai tone distribution through functional data analysis

Hong Zhang[1]

[1]University of Pennsylvania

zhangho@sas.upenn.edu

Abstract

This paper reports an analysis of tonal properties of Thai using a method based on functional data analysis (FDA) on a large collection of TIMIT-like corpus. Both density estimation pooled across time normalized syllable-wise F0 contours and Functional Principle Component Analysis (FPCA) were applied. The results suggest that the simple two dimensional representation of tones: pitch target height and contour slope, is not able to capture context dependent variations of tonal contour within and across tone categories. In addition, the shape and timing of pitch target are also crucial both in differentiating tonal categories and explaining variations associated with syllable structure. The third and fourth dimension of the functional basis have been shown to be able to represent these higher-order properties. Thus FPCA can provide a very simple yet interpretable low dimension representation for the tonal property of Thai. These findings are also potentially helpful for understanding tone distribution properties and tonal coarticulation more generally.

Index Terms: Thai tone, Functional Principle Component Analysis, tone space

1. Introduction

Thai has traditionally been analyzed as having a five-tone system, namely the three static tones: high, mid and low, as well as two dynamic ones: rising and falling[1]. Interestingly, contour shapes of Thai tones appear to vary depending on its position within a phrase [2, 3]. It has been reported that full contours are only realized at phrase final positions or in syllables articulated in isolation. The so-called static tones, on the other hand, in fact also bear somewhat dynamic contour shape [4, 5]. High tone is reported to be realized with a concave contour, while low and mid tones bear a falling contour. The distribution of tone categories is also restricted by syllable structure [1, 6, 2, 7]. Five-way tonal contrast is only present in open syllables[1] or syllables with a long vowel that are closed by a sonorant. Possible number of tonal contrasts is reduced in tones of syllables closed by an obstruent, where no rising or mid tone is allowed regardless of syllable nuclei, and only the syllables with a long vowel can have a falling tone.

Numerous studies [6, 8, 9, 2, 7, 10, 11, 12] have been conducted to understand tonal representations and the distributional properties of Thai tones. Proposals generally attempt to represent each tonal category as unified contour unit[12, 13] or as combinations of H-L autosegment sequence[6, 10], based on both empirical data and theoretical reasoning. However, most of studies are either constrained by the sample size or data generation process, where variability in tone production is highly limited, or both. Recent work has also seen some success in computationally modeling Thai tones through underlying pitch target[14, 15, 16]. Unfortunately, despite its success in extracting pitch targets from Thai tone contours, [14] did not explicitly address the questions regarding tone distribution listed above. Here, we show that the aforementioned properties of Thai tone can be visualized through 2-D density estimation and represented using Functional Principle Component Analysis (FPCA). We propose that F0 target of Thai tone can be better analyzed as a combination of F0 target point and contour shape conditioned on tonal category. This hybrid representation has direct bearings on theories in Thai tone distribution restriction.

[1]Open syllables with a short vowel are generally not included in the discussion, since they are not allowed to occur independently.

2. Method

2.1. The corpus

We used a TIMIT-like spoken corpus of Thai consisting of 6000 read sentences from 50 speakers of the Bangkok dialect. The speakers were selected with an effort to represent the demographic structure of speakers of the language. Each speaker read 120 sentences which may or may not be shared with other speakers. Sentences were selected to maximize the variety of allophonic contexts found in the texts. The total number of words in the corpus is about 120 thousand. Tones in the corpus are labeled according to *King Ross* Thai pronunciation dictionary, and forced-aligned using a Thai forced aligner trained with the same data.

2.2. F0 and density estimation

F0 of the recordings was measured using the auto-correlation method [17] with frame size of 0.01s and step size of 0.005s. F0 measurements in Hz were normalized to semitones for each speaker with the tenth percentile of the person's pitch range as the baseline, via the following formula:

$$semitone = 12 \cdot \log_2 \frac{F0}{\min F0 + 0.1(\max F0 - \min F0)} \quad (1)$$

F0 contours were first smoothed with linear interpolation to attenuate the effect of measurement errors. Each syllable was extracted and resampled to 50 samples for time normalization. Density estimation for each tone category was performed by pooling all examples belong to the same category and fitting a two-dimensional Gaussian kernel with bandwidth determined through Scott rule [18], with normalized time and F0 as the dimensions.

2.3. FPCA basis functions and re-synthesis

The basis functions used for parameter extraction and re-synthesis were approximated by the first 5 principle components (PCs) of F0 observations pooled across all syllables. The first 5 PCs were able to explain 99.74% of the variance of the original covariance matrix. Figure 1 shows the functional shape of the first 4 PCs. This method is a simplified version of the procedure described in [19, 20], in which time invariant effect

10.21437/Interspeech.2018-2115

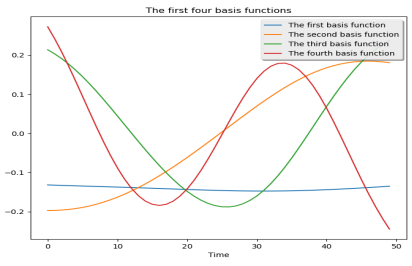


Figure 1: *The first 4 basis functions for tone re-synthesis.*

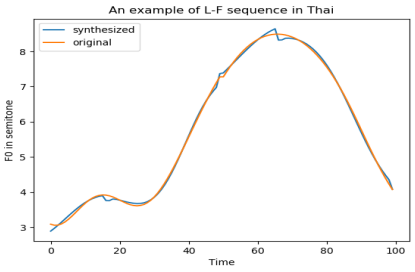


Figure 2: *An example of re-synthesized two-tone sequence compared to the original F0 contour. The horizontal axis represents time as in sample numbers. The first 50 samples are from the first syllable (low tone), and the second 50 are from the second syllable (falling tone).*

such as speaker idiosyncrasy and random noise are explicitly modeled.

It can be seen from Figure 1 that the basis functions have the shapes similar to the first four degree orthogonal polynomials, which have been proposed to effectively model F0 variation [21]. Thus the coefficients attached to each PC dimension for a given F0 contour example can get linguistically meaningful interpretations. Specifically, the first two PCs correspond to F0 height and contour slope, and the third and fourth PCs can get interpretation in terms of the curvature structure and peak location of the pitch contour.

Coefficients attached to basis functions (PC scores) for each example were estimated by projecting the original F0 contours onto the new space spanned by first five PCs via equation 2:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{y} \tag{2}$$

where $\boldsymbol{X}$ is $n \times d$ matrix of basis functions, represented as the eigenvectors of covariance matrix of F0 contours. n is the number of F0 observations in the F0 contour, and d the first d PCs being used. $\boldsymbol{y}$ is the original n dimension F0 contour, and $\hat{\boldsymbol{\theta}}$ is $d \times 1$ estimated coefficient vector.

Equation 2 can be easily generalized to extract coefficient vectors for syllable sequence after syllable-wise time normalization by manipulating the structure of $\boldsymbol{X}$ matrix in the same manner as constructing System Ordinary Least Squares (SOLS) estimators in a Seemingly Unrelated Regression (SUR) fashion [22]. Define the $kn \times kd$ block-diagonal parameter matrix $\boldsymbol{X}$, where k is the number of segments to be analyzed, with the following structure:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{B_1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{B_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{B_k} \end{pmatrix} \tag{3}$$

where $\boldsymbol{B_i}, i = 1, \cdots, k$ are the matrices of basis functions associated with the i^{th} syllable. Figure 2 is an example of re-synthesized two-tone (low-falling tone) sequence using equations (2) and (3). The reconstructed F0 contour is highly similar to the original. Therefore the basis functions generated from single syllables can be conveniently used to investigate more complicated phenomena such as tone coarticulation.

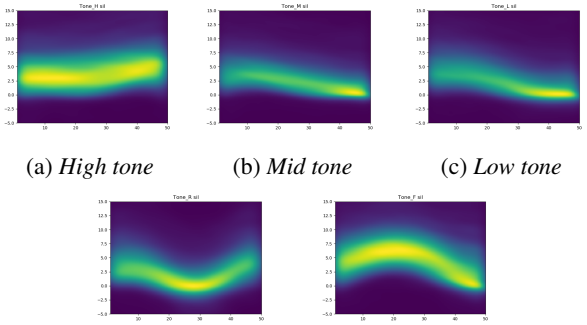
3. Results

3.1. Fully realized tone contours

Since Thai tone contours are only fully realized at phrase final positions [2], we first show the density estimation for F0 contours when they are expected to fully realize.

3.1.1. Contour density estimates

Figure 3 shows the density plots for five tone categories at phrase final position. Higher density region in these plots indicate smaller F0 variation across individual observations. Thus the high density regions in each plot can be interpreted as the common F0 targets across speakers and tokens. With this interpretation, two F0 targets can be identified for high and falling tone: one that occurs earlier in the contour and another that occurs late. Both targets for high tone are bar-shaped, suggesting the target is consistently achieved through tone articulation. A single unit of contour shape can thus better model this target property. On the other hand, the second F0 target in falling tone approximates a high density point, hence the falling tone can be decomposed into an early target of slightly convex contour shape and a late low target as single point.



(a) *High tone* (b) *Mid tone* (c) *Low tone*

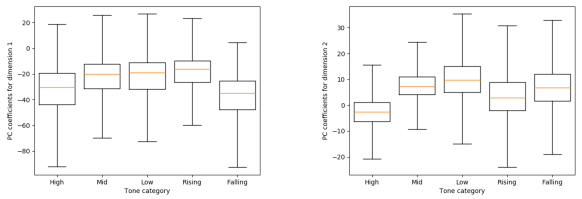
(d) *Rising tone* (e) *Falling tone*

Figure 3: *Density plots of tone shapes at phrase final position. Vertical axis represents F0 in semitone, and horizontal axis represents normalized time in terms of sample number. Each plot was constructed with a sample size of about 3000.*

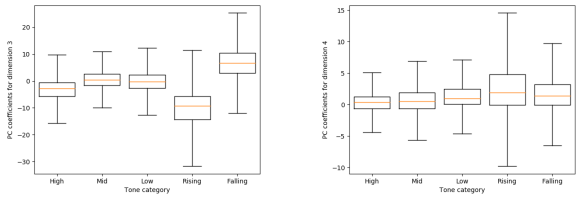
The other three tone categories, however, only show single targets which are better modeled as high density points. Mid and low tone not only show similar F0 range distribution across time, but also have similar F0 target points at the end of the contour. Therefore it can be expected that these two categories are likely to be confused by listener. Rising tone shows a clear low F0 target point in the middle of the contour, while more variability is expected in the initial falling and later rising, although both the rise and fall are clearly seen in the density plot.

3.1.2. *Functional data analysis*

PC coefficients were extracted and plotted in Figure 4. Each subplot shows the distribution of coefficients attached to five tones corresponding to one of the first 4 PCs. In the first PC dimension, little variation is found among mid, low and rising tone, while high and falling tone have smaller values. This pattern suggests that mid, low and rising tone do not differ in their F0 height, which are lower than that for high and falling tone. In the second dimension, high tone has slightly negative coefficients, indicating slight rising contour. Both mid and low tone have positive coefficients at roughly same magnitude, indicating they share a similar monotonic falling contour, although greater variability is seen in the low tone. The slope estimates for rising and falling tone, however, may be compromised with their presence of more complicated contour structure. Nevertheless, both of them still maintain a general falling trend.



(a) *Dimension 1* (b) *Dimension 2*



(c) *Dimension 3* (d) *Dimension 4*

Figure 4: *Distribution of PC coefficients in the first 4 dimensions for the five tonal categories in Thai: high, mid, low, rising and falling.*

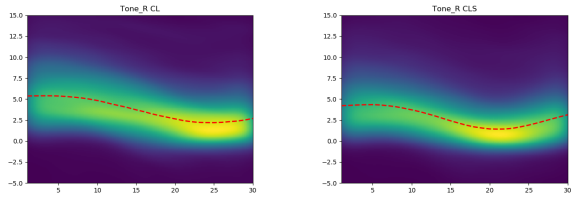
Greatest distinction between rising and falling tone is found in the third PC dimension, where negative coefficients are attached to the rising tone, while falling tone generally has positive coefficients. This dimension can be interpreted as distinguishing between second degree contour structures, that is, whether the contour shape is concave or convex. High tone also shows slight negative coefficients in this dimension, confirming the observation that it has some concave contour structure. On the other hand, coefficients of mid and low tone are centered around 0, indicating lack of second degree polynomial structure. The fourth dimension can be interpreted as distinguishing peak location [5]. Here, both rising and falling tone have large variation in the distribution of coefficients, which can be explained by the interaction between F0 contour and syllable structure that will be elaborated in the following section.

3.2. Tone contour shapes conditioned on syllable structure

Two of the interesting distributional properties of Thai tone are the lack of rising tone in syllables closed by an obstruent, and the prohibition of falling tone in syllables with short vowels and closed by an obstruent[2]. Here we examine these properties with the methodology outlined above.

3.2.1. *The rising tone*

Figure 5 plots the density distribution of F0 contours in open syllables with long vowel (CVV syllable) and syllables with long vowel closed by a sonorant coda (CVVS syllable). The plots were produced by pooling all the syllables without controlling position in phrase.

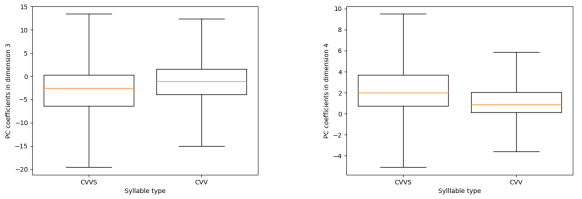


(a) *Rising tone in CVV syllables* (b) *Rising tone in CVVS syllables*

Figure 5: *Density plots of the rising tone in two syllable structures: CVV and CVVS. The red dotted line indicates median F0. Horizontal axis represents normalized time (30 time points per syllable) as sample units, and vertical axis is F0 in semitone. The sample size for CVV syllable is around 2,000, and for CVVS syllables is around 3,300.*

The density plots clearly indicate the different locations of F0 targets in two syllable structures. In syllables not closed by a sonorant coda, F0 target is located at the end of the contour, and the rising contour is not realized. On the other hand, the low F0 target is realized earlier in syllables with sonorant coda, where slight rising contour is also present.

On average 31.2% ($SD = 7.84\%$) of total syllable duration is taken by the coda, which corresponds to the horizontal location of high density region in 5(b). Therefore low F0 target is realized roughly at the boundary between syllable nucleus and coda.



(a) *Rising tone in CVV syllables* (b) *Rising tone in CVVS syllables*

Figure 6: *Distribution of PC coefficients in the third and fourth dimension for rising tone in CVV and CVVS syllables.*

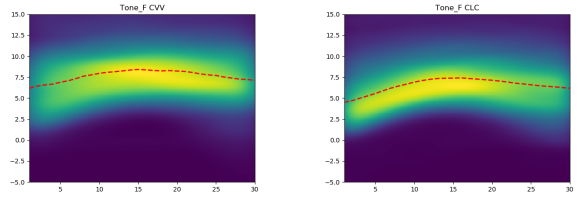
The distribution of PC coefficients in the third and fourth dimensions, as shown in Figure 6, suggests that rising tone in CVVS syllables has clearer concave-shaped contour, since there are stronger negative coefficients in the third PC dimension. Greater coefficients in the fourth dimension is also found for CVVS syllables, which can be associated with earlier realization of the peak.

The results from density estimate and FPCA analysis on the distributional property of the rising tone have direct implications on the debate over the underlying reason for the distributional restriction. One possible simple explanation for this

question could just be that the rising contour is only realized after the low F0 target, which happened to be at the end of vowel articulation. That is, the rising contour is only born on the sonorant coda in general.

3.2.2. The falling tone

We combine syllables closed by sonorant and obstruent as a single category for the falling tone, due to the imbalanced token distribution of the two syllable structures. Thus open (CVV) and closed (CVVC) syllables with long vowels are considered. Figure 7 shows the density plots for the two syllable structures. The plots were produced without controlling for position in phrase.

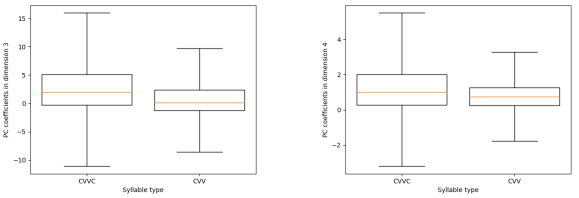


(a) *Falling tone in CVV syllables* (b) *Falling tone in CVVC syllables*

Figure 7: *Density plots of the falling tone in two syllable structures: CVV and CVVC. The red dotted line indicates median F0. Horizontal axis represents normalized time (30 time points per syllable) in sample units, and vertical axis is F0 in semitone. The sample size for CVV syllable is around 6,600, and for CVVC syllables is around 8,000.*

Similar to the case in rising tone, falling tone in CVVC syllables shows earlier F0 target realization, and more variable, though not always falling, contour shape after the target. Unlike rising tone where a rising trend is more apparent in closed syllables, falling tone in general doesn't show a realized falling contour regardless of syllable structure, which is consistent with previous consensus.

The time difference between target realization in two syllable structures also largely corresponds to the boundary between the vowel and consonant coda, where on average the coda takes 26.3% of the total duration ($SD = 7.84\%$).



(a) *Falling tone in CVV syllables* (b) *falling tone in CVVC syllables*

Figure 8: *Distribution of PC coefficients in the third and fourth dimension for falling tone in CVV and CVVC syllables.*

Distributions of PC coefficients in the third and fourth PC dimensions, as shown in Figure 8, also suggest a similar relation between closed and open syllables as observed in rising tone. Small positive coefficients in dimension 3 in CVVC syllables indicate slight convex-shaped contour. However, CVV syllables do not seem to clearly bear such contour structure. Coefficient distribution in the fourth dimension suggests that F0 peak is reached earlier in CVVC syllables. Greater variation is also present in both dimensions for CVVC syllables.

The distribution of PC coefficients in the third and fourth dimensions agrees with the observation from density plots. The trends of F0 distribution reported above, for both tones, seem to suggest that coda consonants have the effect of enforcing an earlier F0 target realization. For falling tone specifically, this means the entire contour unit is realized earlier in closed syllables compared to the open ones. In addition, since the second F0 target is often omitted, the shape of the first contour target may also be slightly different across the two types of syllables, as seen from the stronger curvature in CVVC syllables.

4. Discussion

This paper reports a reanalysis of Thai tone space using density estimation and FPCA using a TIMIT-like corpus of spoken Thai. Density estimation offered a robust and more intuitive way to visualize the distribution of F0 estimates throughout the segmental unit of interest. This method could offer detailed information about the distributional properties of F0 values throughout the entire F0 contour. High density estimates can also be conveniently related to the common F0 targets shared across examples in the dataset. PC dimensions found through FPCA can serve as a simple yet interpretable parameterization of the tonal properties observed from density plots. We demonstrated how such simple techniques can be used to address theoretically driven linguistic questions.

Two types of F0 target were identified through F0 density estimation from our Thai corpus: the target with particular contour shapes and the one as a particular F0 value in speaker's tone space. Thus a robust theory on the tonal representation of Thai is desirably reflect the dynamics of tonal target reported here, which has not been incorporated in previous theories on Thai tone representation [2]. Density estimates have also identified the correspondence between F0 target and vowel-coda boundary in both rising and falling tone, which points to a better grounded explanation for the distribution restriction. The observations made from density plots are confirmed quantitatively via the distribution of PC scores.

5. Conclusions

It has been shown in this paper that density estimation can offer a more intuitive perspective on properties of lexical tones in a tonal language. FDA methods, such as FPCA, can effectively capture the variations in tone distribution and offer a quantitative yet interpretable representation of the tonal property of interest.

6. Acknowledgements

I would like to thank Nattanun Chanchaochai for making the corpus available for this analysis, and Mark Liberman and Jianjing Kuang for their insights in functional data analysis.

7. References

[1] A. S. Abramson, *The vowels and tones of standard Thai: Acoustical measurements and experiments*. Indiana Univ., 1962, vol. 20.

[2] B. Morén and E. Zsiga, "The lexical and post-lexical phonology

of Thai tones," *Natural Language & Linguistic Theory*, vol. 24, no. 1, pp. 113–178, 2006.

[3] S. Potisuk, J. Gandour, and M. P. Harper, "Contextual variations in trisyllabic sequences of Thai tones," *Phonetica*, vol. 54, no. 1, pp. 22–42, 1997.

[4] A. S. Abramson, "The tones of Central Thai: Some perceptual experiments," *Studies in Thai linguistics in honor of William J. Gedney*, pp. 1–16, 1975.

[5] J. Gandour, S. Potisuk, S. Ponglorpisit, and S. Dechongkit, "Inter- and intraspeaker variability in fundamental frequency of Thai tones," *Speech Communication*, vol. 10, no. 4, pp. 355–372, 1991.

[6] J. Gandour, "On the representation of tone in Siamese," *UCLA Working Papers in Phonetics*, vol. 27, pp. 118–146, 1974.

[7] E. Zsiga and R. Nitisaroj, "Tone features, tone perception, and peak alignment in Thai," *Language and Speech*, vol. 50, no. 3, pp. 343–383, 2007.

[8] J. Gandour, "On the interaction between tone and vowel length: Evidence from Thai dialects," *Phonetica*, vol. 34, no. 1, pp. 54–65, 1977.

[9] J. Zhang, *The effects of duration and sonority on countour tone distribution: A typological survey and formal analysis*. Routledge, 2013.

[10] M. Yip, "Against a segmental analysis of Zahao and Thai: A laryngeal tier proposal," *Linguistic Analysis Seattle, Wash.*, vol. 9, no. 1, pp. 79–94, 1982.

[11] ——, *Tone*. Cambridge University Press, 2002.

[12] Y. Xu, "Consistency of tone-syllable alignment across different syllable structures and speaking rates," *Phonetica*, vol. 55, no. 4, pp. 179–203, 1998.

[13] A. S. Abramson, "The coarticulation of tones: An acoustic study of Thai," *Studies in Tai and Mon-Khmer phonetics and phonology in honour of Eugenie JA Henderson*, pp. 1–9, 1979.

[14] S. Prom-On and Y. Xu, "Pitch target representation of Thai tones," in *Tonal Aspects of Languages-Third International Symposium*, 2012.

[15] S. Potisuk, M. P. Harper, and J. Gandour, "Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 95–102, 1999.

[16] P. Seresangtakul and T. Takara, "A generative model of fundamental frequency contours for polysyllabic words of Thai tones," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–I.

[17] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[18] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[19] J. A. Aston, J.-M. Chiou, and J. P. Evans, "Linguistic pitch analysis using functional principal component mixed effect models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 59, no. 2, pp. 297–317, 2010.

[20] P. Z. Hadjipantelis, J. A. Aston, H.-G. Müller, and J. P. Evans, "Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin Chinese," *Journal of the American Statistical Association*, vol. 110, no. 510, pp. 545–559, 2015.

[21] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A statistics-based pitch contour model for Mandarin speech," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 908–925, 2005.

[22] A. Zellner, "Estimators for seemingly unrelated regression equations: Some exact finite sample results," *Journal of the American Statistical Association*, vol. 58, no. 304, pp. 977–992, 1963.