Computational modeling of tone in language documentation: citation tones vs. running speech in Chindwin Khamti

RIKKER DOCKUM
Yale University[*]

1 Introduction

Recent documentation of a four-tone system in Khamti (Dockum 2015), spoken along the Chindwin River in northwestern Myanmar, distinguishes it from all varieties of Khamti previously described, dating back to the 19th century (Robinson 1849, Needham 1894, Harris 1976, Morey 2005, etc). In addition to being tonally divergent from its closest known relatives, Khamti is also one of the more geographically distant members of the Southwestern Tai (SWTai) branch of the Kra-Dai language family, as compared to the majority of speakers of SWTai languages. This distance, both tonal and geographic, motivates closer investigation into the tones of the modern language.

This paper more closely examines tone in Chindwin Khamti along two tracks: first, it reports on a study to gather much more data on synchronic tones across social variables of age, gender, and village location; and, second, it uses a portion of this data to test methods for computational modeling of tone demonstrated by Shosted et al (2015) on Iu Mien. They advocate for the integration of computational methods early in and throughout the process of field documentation, to assist in achieving adequate description of understudied tonal languages while reducing impressionistic judgments in such descriptions. Computational models can serve as a potential check against human error in assigning tone categories, and can reveal interspeaker variation in tone systems that might go unnoticed using traditional methods. Both of these benefits then contribute toward the goal of greater replicability and falsifiability of linguistic research.

In the remainder of this paper, section 2 provides background on Khamti and the tones of the Chindwin River variety, section 3 describes audio data collection for a running speech corpus and a wordlist corpus, section 4 details the method for computational modeling of Khamti tones, section 5 presents the clustering results and discussion, and finally section 6 concludes the paper.

[*] My sincere thanks to Ryan Bennett for his tireless advising on this paper, and to Claire Bowern, Parker Brody, Martín Fuchs, Luke Lindemann, Josh Phillips and all others for their helpful feedback.

2 Background

Khamti [ISO 639-3: kht], also known as Tai Khamti and Khamti Shan, is a language of the Kra-Dai (formerly Tai-Kadai) family, from the Southwestern Tai branch. Khamti-speaking communities are found in Northeast India and northern Myanmar (Lewis et al 2016). Khamti is among the earliest Tai languages to be described in detail, including Robinson (1849), Needham (1894), and Grierson (1904).[1] In the intervening century, however, it has received relatively little attention, and was assumed to be homogeneous. Previous fieldwork (Dockum 2015) uncovered important differences in the lexical tones of Khamti spoken in Khamti District, Myanmar, in particular only four tonemes, whereas all previous literature on Khamti has described five. Further analysis revealed a very different history of historical tone splits and mergers than other Khamti varieties, but from the same common ancestor, termed Proto-Khamti.

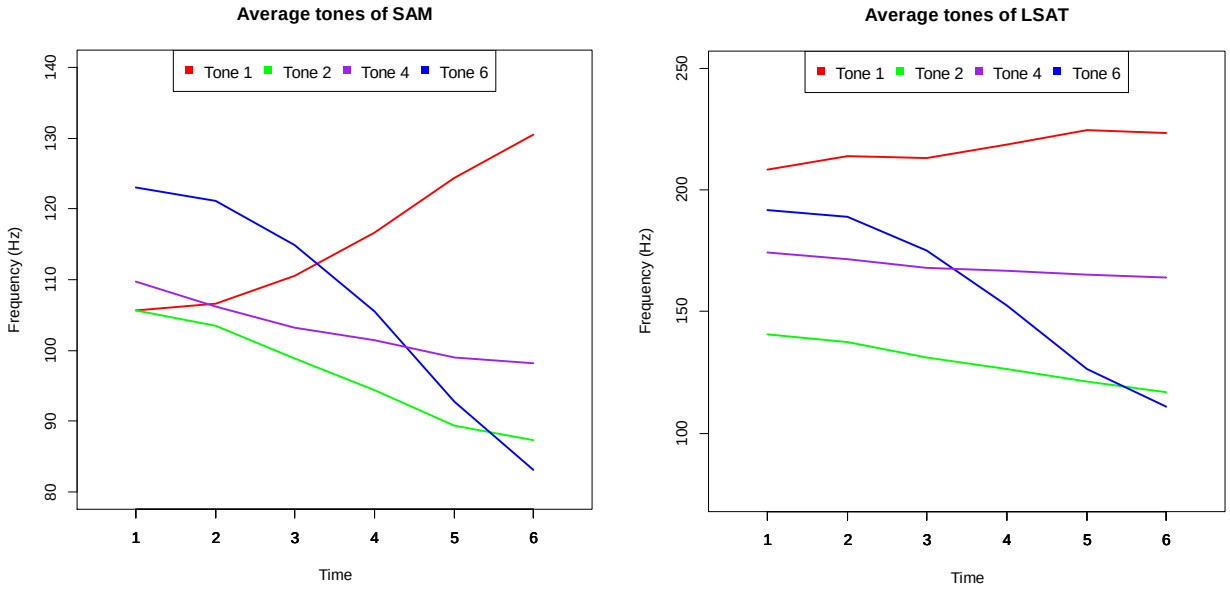
To help disambiguate the modern varieties, I use the name Chindwin Khamti to refer to the variety spoken along the upper Chindwin River in Khamti District, Sagaing Division, Myanmar. This distinguishes it from Khamti of Northeast India (see e.g. Needham 1894) and Khamti of Kachin State, Myanmar (see e.g. Inglis 2014), among other possible varieties.

2.1 Chindwin Khamti tones

Using traditional field methods of audition and instrumental comparison, Dockum (2015) reports that Chindwin Khamti has four lexical tones. These are referred to throughout this paper as tones 1, 2, 4, and 6, based on the subset of tone marks in Shan script that are used to write Chindwin Khamti locally. (Numbers 3 and 5 represent unused glyphs.) Examples of the mean tonal space of two speakers from that study are given in Figure 1. Both speakers are male natives of Khamti Township, and are aged 38 and 75, respectively.

[1] Preceded only by Low (1828), a sketch grammar of Thai.

Figure 1: Average toneme pitch tracks of SAM, age 38, and LSAT, age 75 (Dockum 2015)



Their approximate representation in the widely-used five level pitch notation system proposed by Chao (1930) is given in Table 1.

Table 1: Chindwin Khamti tones of two speakers in Chao (1930) tone numbers

Speaker	T1	T2	T4	T6
SAM	25	21	41	32
LSAT	45	21	41	33

The surface differences between these two speakers led me to gathering Khamti tone data on a larger scale, described in the next section.

3 Data collection

Two types of data were gathered, described further in §3.3 and §3.4: responses to question/answer stimuli and core vocabulary wordlist data.

3.1 Locations

Data were gathered from five locations in Khamti Township, Khamti District, Sagaing Division, Myanmar. All of the sites are located along the upper Chindwin River. The five locations include the main town of Khamti Township, abbreviated, KT, as well as four rural Tai villages, which are referred to herein as MP, MN, SN, and LP. All are accessible within a few hours by boat from the main town, a necessity as only day trips were allowed. A list of the village abbreviations and dates each was visited is given in Table 2, a map of approximate locations in Figure 2, and an area map in Figure 3.

Table 2: Recording sites and dates visited

Village	Date recorded
MP	8 June 2015
MN	11 June 2015
SN	13 June 2015
LP	15 June 2015
KT	17 June 2015

Figure 2: Map of recording sites along the Chindwin River, Khamti Township

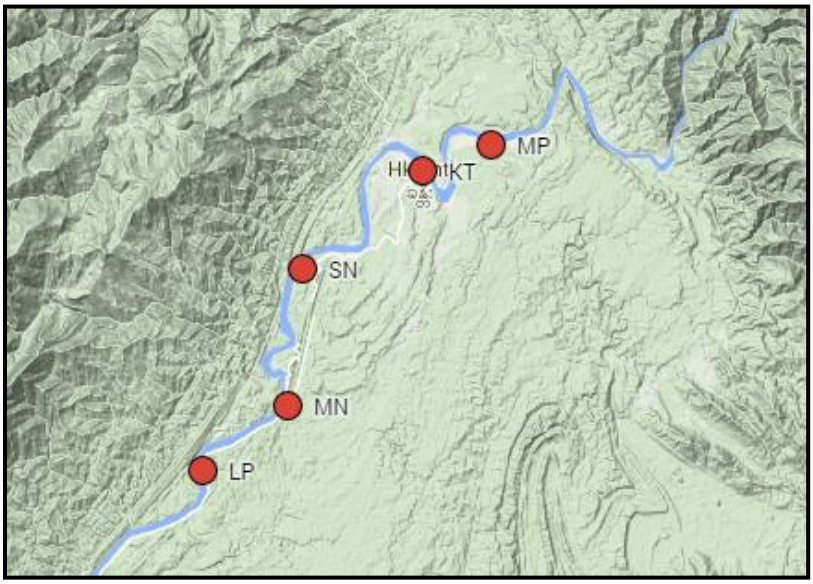


Figure 3: Area map with a black box to show the location of the area in Figure 2

3.2 Procedure

Upon arrival at each village recording site, my language informants recruited available speakers, with a goal of at least six speakers from four age groups: under 20, 20-40, 40-60, and over 60, with a mix of both genders. A full listing of the participants by age and gender is given in Table 3.

Table 3: Study participants by location, age and gender

Village	Male	Female
MP	28, 36, 47, 60	17, 18, 19, 19, 66, 76
MN	25, 35, 57	22, 42, 70
SN	20, 65, 74	18, 20, 40, 63
LP	24, 56, 78	24, 47, 63
KT	12, 18, 21, 62, 65	12, 53, 75

Explanation of the task ranged from approximately 10 to 20 minutes per speaker, and the recording of the questions answering itself took from 10 to 40 minutes, for a total participation time of between 20 minutes and 1 hour per speaker. All work was conducted at the monastery

building of each village, which is both the residence of local monks and the center of community activities.

Recordings were made in uncompressed WAV format using a Zoom H4N multitrack digital recorder, with the speaker wearing an Audio-Technica Pro 8HEx hypercardioid dynamic headset microphone connected via one XLR jack, and the second XLR jack recording with an Audio-Technica AT2005 cardioid dynamic tabletop microphone resting on a small tripod and pointing toward the speaker wearing the headset microphone. A backup recorder was also used, a Roland R-05, usually with the built-in stereo microphones for purposes of capturing the questions asked by the informant, should later review be needed. Though there was quite a bit variation in the specific questions asked, they were largely formulaic and thus entirely recoverable from the answers given by speakers, even in the event they did not get picked up by the tabletop microphone. On one occasion, the informant administering the stimuli also wore an Audio-Technica ATR-3350 omnidirectional condenser lavalier microphone on his shirt collar, in order to get a clearer recording of a set of questions.

3.3 Question/answer stimuli

Sixteen words were chosen, representing each of the four lexical tones with four words each. To guarantee speaker familiarity across all age groups, I selected concrete nouns, primarily nature and food items, with the assistance of my primary Khamti informants, two natives of Khamti Town, the small town that forms the administrative and economic center of Khamti Township. Originally a list of just 12 lexical items was used at the first village site, which I refer to as MP. Some of these differ from the final list, as they were changed after they turned out to be socially infelicitous to use.[2] The final list of target words is given in Table 4.

[2] Two words removed from the list after the first village trip were the words for 'father' and 'mother'. There was concern that asking elderly participants even in the abstract about their parents, who were likely deceased, could come across as inappropriate or insensitive.

Table 4: List of target words used in stimuli questions with order of presentation

#	Form[3]	Gloss	#	Form	Gloss	#	Form	Gloss	#	Form	Gloss
1	/maː^{1}/	dog	**2**	/kʰaw^{2}/	rice	**3**	/paː^{4}/	fish	**4**	/kai^{6}/	chicken
5	/mi^{1}/	bear	**6**	/maː^{2}/	horse	**7**	/kʰaːi^{4}/	buffalo	**8**	/kʰaː^{6}/	galangal
9	/pʰaː^{1}/	wall	**10**	/ɔi^{2}/	sugarcane	**11**	/naːw^{4}/	star	**12**	/taw^{6}/	turtle
13	/sʰɤ^{1}/	tiger	**14**	/saːŋ^{2}/	elephant	**15**	/nɤn^{4}/	moon	**16**	/tʰo^{6}/	bean/nut
Tone 1			**Tone 2**			**Tone 4**			**Tone 6**		

The stimuli were initially designed to showcase each of the tones in phrase initial, medial, and final positions, in order to capture variations in focus prosody, by using three carrier sentences for each word. The stimuli would need to be administered by one of the informants traveling with me. Based on their feedback, however, the stimuli were revised to avoid a task too confusing for potential participants.

Stanford (2008) describes similar difficulty with set carrier phrases, especially when the language to be documented is not a focus of formal study in school, and non-functional repetition of arbitrary sentences is a completely foreign task. Stanford instead settled on a 'flexible phrase list' (2008:16). In a similar vein, after some discussion with my informants we settled upon a 'flexible frame question' instead of trying to introduce the concept of carrier sentences.

The three frame questions followed the format in 1a-c:

(1) a. Have you ever seen / eaten / etc _____?[4]
b. What kind of _____ have you seen / eaten / etc?
c. Where is _____ found?

Participants were given some examples of the patterns by SAM, the informant who administered the stimuli, with an instruction to use the target word in their response, and to repeat their response

[3] All Khamti vowels are phonetically long in open syllables, but because there is a length contrast between /a/ and /aː/, the duration notation is used.

[4] As this is a binary question, it should be noted that Khamti, like many languages in Southeast Asia, has no simple 'yes' and 'no' words in the same sense as English. An affirmative response is minimally signaled by repeating the verb or verb complex of the question, and the negative is minimally signalled by negating it. Thus the expected answer for a binary question like, 'Have you ever seen a tiger?' would minimally be 'have seen' or 'have never seen', with pronouns and verb arguments recoverable from context. Thus while 'yes/no' answers to stimuli questions were not an issue, per se, participants still needed to be instructed to answer in a complete sentence, to ensure they would repeat the target word.

three times. With responses to three questions, and three repetitions of each response, for all sixteen words, this created 3 x 3 x 16 or 144 tokens per speaker. It was expected that each speaker would fail to use the target word in some responses, which did happen, but also other times they would also use it more than once (as in responses like, 'The rice that I've eaten is steamed rice'). As it turned out, the number of times the target word was repeated outnumbered the number of times it was omitted, for an average well above 144 tokens per speaker.

Some variation in prosody in responses was achieved, as desired, but this could not be controlled for closely. The majority of responses have focus prosody, and a large proportion were utterance initial, as Khamti is a topic-comment language with frequent topic fronting. While variation in focus prosody was not reliably achieved, ultimately the data gathering was still very successful. Each speaker produced several utterances of the target words, making for a substantial corpus to work with.

3.4 Wordlist data

Following a 2014 fieldwork trip, copies of the 436-item Mainland Southeast Asia (MSEA) wordlist (SIL 2002) were circulated to several neighboring Tai villages by my principal Khamti informants. This list contains glosses in English, Thai, and Burmese. A representative from each village that received the list filled out the list with the corresponding lexical items used in that village. Upon my return to the area in 2015, if time permitted after completing the stimuli response recordings at each village, one or more speakers were recruited to record a reading of the full wordlist, using the same equipment setup described in §3.2, by repeating each entry three times. The list of speakers who made wordlist recordings is given in Table 5. The wordlist recorded by the speaker from LP village was used to create one of the corpora in this study.

Table 5: Speakers who recorded a wordlist reading

Village	Speaker age, sex
MN	49, M
SN	57, M
LP	24, F
KT	36, F; 40, M

3.5 Corpora

Two corpora were prepared for use in computational modeling. A corpus of citation tones spoken in isolation from a wordlist, with a total size of 173 tokens, was prepared to serve as a baseline for evaluating modeling performance on the larger, more complex corpus: 750 tokens segmented from the stimuli responses of five speakers. Details by speaker with token counts for both corpora are given in Table 6.

Table 6: Speaker demographics and toneme distribution in each corpus

				Tone tokens				
ID	**Sex**	**Age**	**Data type**	**Tone 1**	**Tone 2**	**Tone 4**	**Tone 6**	**Total**
LP5	F	24	Wordlist reading	54 (31%)	43 (25%)	47 (27%)	29 (17%)	173
								173
MN2	F	42	Stimuli responses	35 (28%)	18 (14%)	39 (31%)	32 (26%)	124
SN2	M	74	Stimuli responses	49 (31%)	37 (23%)	39 (25%)	34 (21%)	159
SN3	F	40	Stimuli responses	47 (31%)	34 (22%)	37 (24%)	34 (22%)	152
SN5	F	20	Stimuli responses	40 (24%)	43 (26%)	43 (26%)	40 (24%)	166
KT5	M	65	Stimuli responses	33 (22%)	34 (23%)	41 (28%)	41 (28%)	149
				204	**166**	**199**	**181**	**750**

3.5.1 Controls

Both corpora were controlled for tone type, in order to achieve representative samples from each toneme. The wordlist corpus was also controlled for syllable shape, with only CV syllables selected. While Khamti has no contrastive vowel length in open syllables, all vowels are phonetically long in this position. Both monophthongs and diphthongs were used.

The target words in the stimuli response are described in §3.3. Syllable shape was also restricted to CV syllables in 14 of the 16 target words. The other two have shape CVN.[5] The stimuli corpus was also controlled for lexical category, targeting exclusively common nouns due to their ease of elicitation, whereas the wordlist corpus contains items from all lexical categories.

[5] Though open syllable vowels in Khamti are all phonetically long, the average duration of the sonorant period in CVN syllables in the running speech corpus was 28% longer (149ms) than the average duration of CV syllables (117ms).

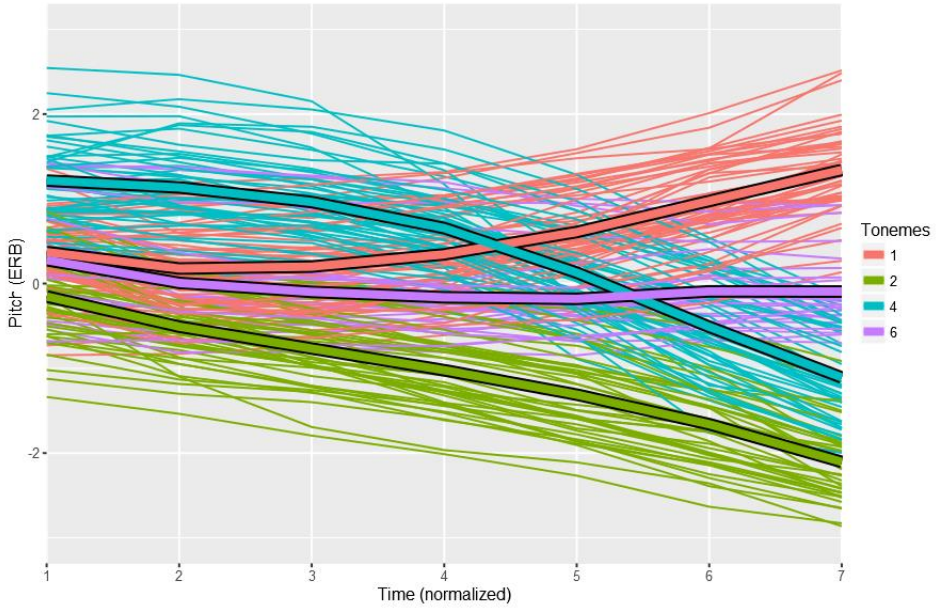
3.5.2 Segmentation and measurement extraction

For both corpora, each utterance of the target words was segmented and annotated using TextGrids in Praat (Boersma and Weenink 2016). The annotated portion for each token begins from the zero crossing immediately following the first full period after the onset of the second and third vowel formants, and ends in one of two places: (i) for open syllables, at the zero crossing after the last full period at the end of the vowel, or (ii) for sonorant final syllables, at the end of the sonorant coda. In some cases where the open syllable of a target word was immediately followed by the onset of the next syllable, then the end of the higher formants of the vowel were used to segment the end of the target word. Disfluent utterances of target words were not annotated.

A Praat script by Bennett (2015) was then used to extract acoustic measurements including pitch in hertz and ERB at seven time-normalized intervals across the space of each annotated tone token. In all, the annotated corpus used for the computational modeling below consists of the responses of the five speakers to the stimuli questions. With an expected 144 tokens per speaker (3 x 3 x 16 question responses), the actual raw count after was 832 tokens, though this was reduced to 759 tokens after omitting items for which F0 could not be extracted; still well above the goal of 144 per speaker.

Figure 4 shows average F0 tracks for the average of each of the four lexical tones for each of these five participants, as well as a normalized average for the whole tone corpus.

Figure 4: Normalized pitch tracks and toneme averages (speaker = LP5, corpus = wordlist)

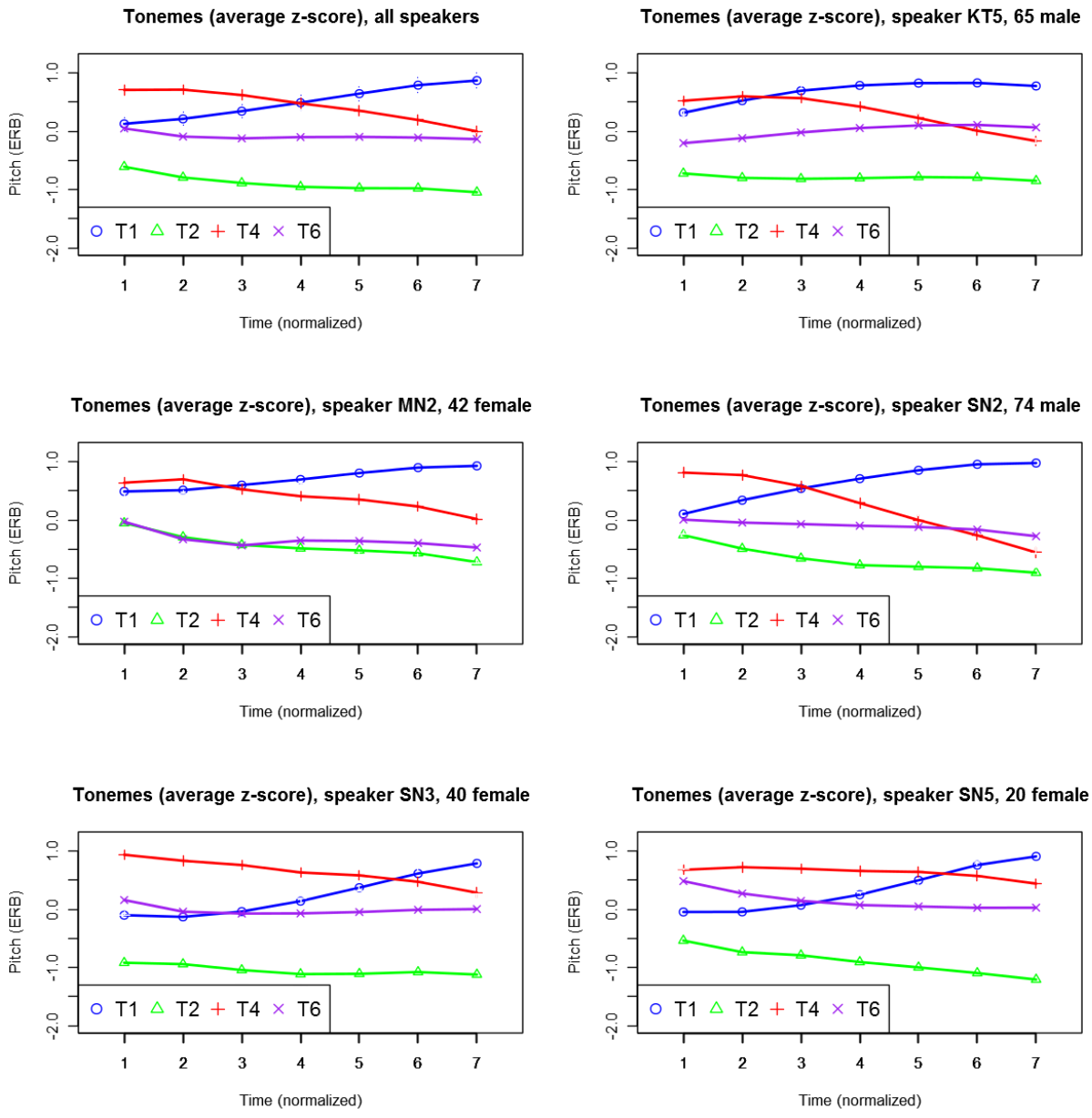


3.5.3. Tonemes

Pitch tracks from the wordlist corpus are given in Figure 3 along with the averages for each toneme category. These are in ERB and z-score normalized using the *scale* function of R. The toneme averages more closely resemble the younger of the two speakers whose tonal spaces were presented in Figure 1.

For comparison, the averages for speakers in the larger corpus are given in Figure 5.

Figure 5: Tonemes of each speaker (stimuli corpus)



Though these tonemes are averages, and intraspeaker variation by prosodic context would need to be examined in more detail before drawing strong conclusions, there are many notable points of variation between these speakers. For instance, speaker MN2 has tones 2 and 6 virtually on top of one another, which may indicate the use of non-modal phonation as a more salient cue for this speaker. Also, the falling tone of tone 4 appears nearly level in speaker SN5.

4 Computational modeling

Though much sophisticated hardware and software exists to assist with field language documentation, the core of the activity frequently still consists of linguists of varying experience making informed judgments to the best of their ability. This ranges across everything from deciding what symbols to use to represent sounds in transcription—even when not sure exactly which sounds they may be yet—to deciding how to boil down an often large number of allotones into a precise count of underlying tonemes in the language under study. There is inherently an impressionistic element in the task, and interactions with phenomena like phonation type or tone sandhi make the task very complex. Snider (2014) criticizes the state of tonal descriptions in the literature generally, arguing that there is a tendency to emphasize minimal pairs, while often failing to notice or control for copious possible confounds, both phonological and grammatical, that can invalidate minimal pair evidence.

One potential way to address the issue of impressionistic judgments or other weaknesses in the state of the art of tonal analysis is with the development and refinement of computational methods. Shosted et al (2015) outline a method for computationally modeling lexical tone. They advocate for incorporating such methods directly into the fieldwork process, as early as possible, rather than leaving it to post-analysis performed after leaving the field site, if performed at all. This adds replicability to the typically highly individualized task of language documentation fieldwork, and acts as a check against human error and impressionistic judgments. Principal Components Analysis is one possible method.

4.1 Principal Components Analysis

Principal Components Analysis (PCA) is a dimensionality reduction that abstracts observations with possible correlations into a set of uncorrelated variables (Jolliffe 1986). The output of a principal components analysis is a matrix of ‘scores’ and a matrix of ‘loadings’. Each column in

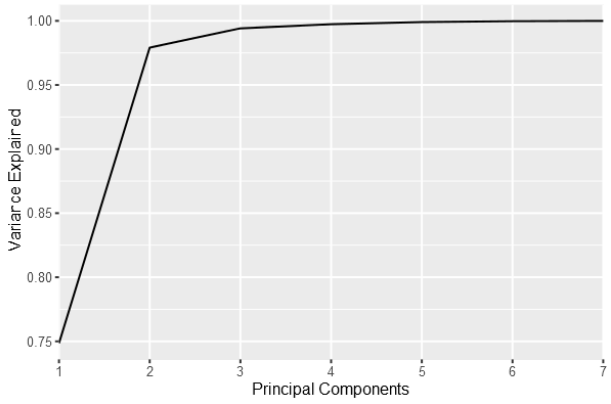
the loadings is one 'principal component', and represents a correlation present in the data. (Johnson 2008:99). The principal components are ordered by how much of the variance in the data they account for, so the first principal component (PC1) explains the most, followed by PC2, and so on. Each observation in the input data is assigned a score with respect to each PC.

The observations used for PCA in this paper are phonetic measures of pitch and phonation across several time steps in each tone token. The *prcomp* method of the default R package *stats* was used to perform the analysis (R Core Team 2015).

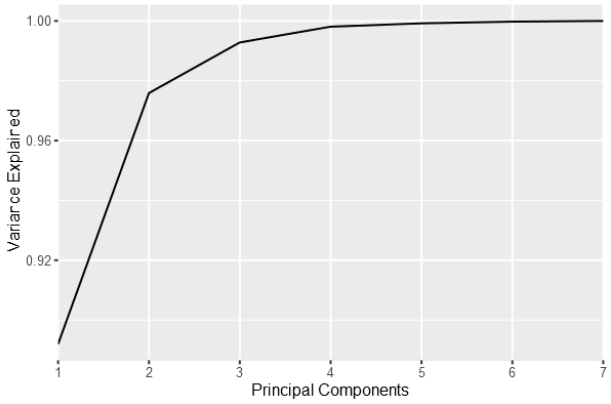
4.1.1 Variance explained

To determine how many of the principal components to make use of in analyzing a dataset, we look at the variance explained. This is calculated from the output of PCA and graphed in an elbow plot. While there is no hard rule for how many principal components to use, a rule of thumb is to use as many as are needed to explain 95% of the data, or else stop when the next principal component would explain less than 5% more of the variance (Baayen 2008:121). Two examples are in Figure 6.

Figure 6: Proportion of variance explained by principal components



Speaker = LP5
Corpus = wordlist
Measures = pitch



Speakers = KT5, MN2, SN2, SN3, SN5
Corpus = stimuli
Measures = pitch

Here, for the Khamti wordlist corpus, using only pitch measures, two principle components explain above 95% of the variance. Similarly, for the stimuli response corpus using only pitch measures, the first two principal components explain above 97.9% of the variance.

4.1.2 Loadings

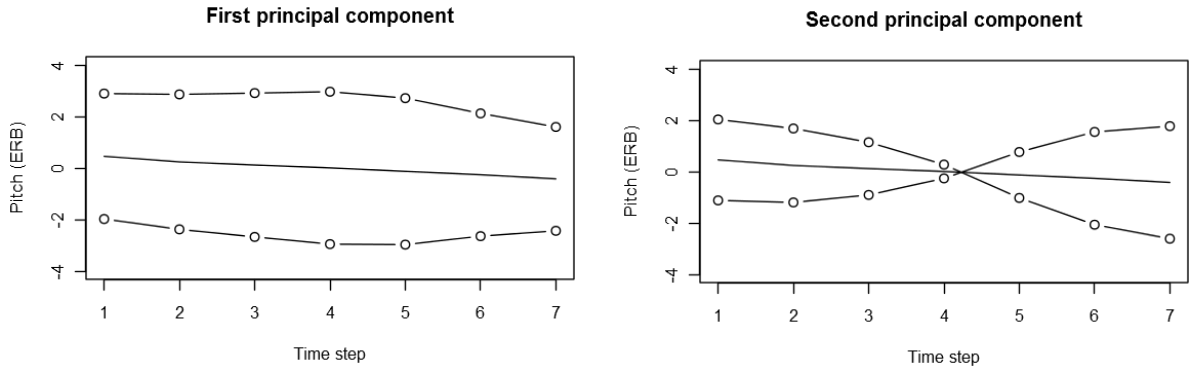
Having determined that for both corpora, the first two principal components explain above 95% of the data, we can make our cutoff for use in *k*-means clustering at PC2. To understand what we are clustering on, we need to look at the PCA loadings, given in Table 7.

Table 7: Loadings of the principal components (wordlist corpus)

Pitch step	PC1	PC2
Step 1	-0.3546	0.4132
Step 2	-0.3815	0.3774
Step 3	-0.4065	0.2689
Step 4	-0.4307	0.0702
Step 5	-0.4140	-0.2345
Step 6	-0.3470	-0.4737
Step 7	-0.2935	-0.5746

We can examine PC1 and PC2 to determine what their likely phonetic correlates are. Figure 7 plots the deviations from the corpus mean based on the loadings and standard deviations of the first two principal components.

Figure 7: First and second principal components, reflected around their centers



Given their shape, and what we know about the input tonal data, it is likely that PC1 corresponds to pitch height, and PC2 corresponds to pitch slope.

4.2 *k*-means Clustering

The output of PCA, with two PCs per tone token determined as optimum, can then be used as input to *k*-means clustering. This was done using the *kmeans* method of the default R package *stats* (R Core Team 2015). This method initiates by randomly selecting a specified number of centers after an initial burn in, here set to 1000 iterations. The distance between each initial center and datapoint is calculated, and each observation is assigned to its nearest center. New centers are then calculated from the mean of the currently assigned clusters, after which each observation is again assessed and reassigned to a new center if a closer one now exists. This process repeats until it converges (Hartigan and Wong 1979).

In order to narrow the space of possible models, I chose to begin by specifying four centers, so that the number of fitted clusters would be the same as the number of expected clusters. These could then be scored as a baseline and compared against models with different number of centers as needed.

4.2.1 Scoring *k*-means output

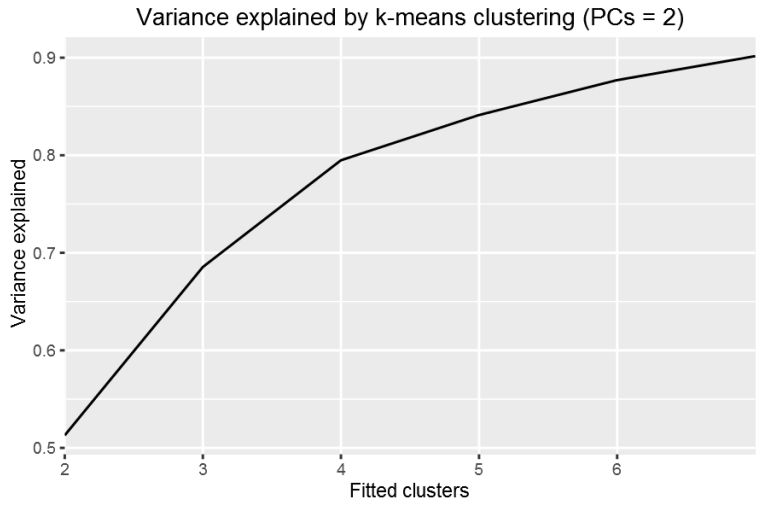
Although multiple runs of *k*-means will converge on the same clusters, given the same input and a sufficient number of iterations, since the starting clusters are randomly selected, which cluster corresponds to which toneme is random with each run. In order to allow repeated runs of *k*-means on different data sets and testing inclusion of different pitch and phonation measures, I created a simple method in R that determines the optimal assignment of output cluster to toneme categories. It operates by first identifying for each expected tone cluster, which inferred cluster matches the highest proportion of its tokens. If two tonemes are matched with the same inferred cluster, the second-best performing inferred clusters are compared, and the one with higher performance in that comparison is assigned to its toneme. If more than two tonemes are matched to a single inferred cluster, the process iterates until each inferred cluster is uniquely assigned to an expected toneme cluster. This also enables automated scoring of the output of *k*-means with maximized scores.

4.2.2 Determining optimal number of clusters

One of the potential benefits to computational modeling of tones is as a check against impressionistic evidence in language documentation, by providing falsifiable quantitative results (Shosted et al 2015). While computational models are a step in this direction, interpreting the output of the models may introduce a different vector for impressionistic judgments. In this case, determining the optimal number of clusters in a dataset. Numerous methods exist, with 30 methods included in the R package *NbClust* (Charrad et al 2015) alone.

One is the same method used above to identify the optimal number of principal components: with an elbow plot of the variance explained. This is given for a *k*-means analysis of pitch data from speaker LP5, using two principal components, in Figure 8.

Figure 8: Variance explained by number of clusters (speaker = LP5, corpus = wordlist, measures = pitch)



With this method, the location of the 'elbow'—past which an increase in the number of clusters does not gain substantial ground in variance explained—is not always unambiguous (Ketchen and Shook 1996). In Figure 8, with $k = 4$, we have accounted for 80% of the variance, and all successive clusters add gains of less than 5% each. But since the fifth cluster explains an additional 4.6% of the variance, one might also decide to make the cutoff there instead. The potential for influence from the expected number of clusters is obvious.

To address this problem, Charrad et al authored the R package *NbClust* (2015), which aggregates 30 different methods for determining optimal cluster performance. These include methods proposed by Scott and Symons (1971), Krzanowski and Lai (1988), Halkidi et al (2000), to select a few. *NbClust* returns counts which number of clusters had the most 'votes' by the different methods and returns a consensus.

Using the same data used to make the elbow plot in Figure 7, *NbClust* returns a consensus of $k = 3$ as optimal, with votes from 11 of the methods. The distribution of optimal scores returned by *NbClust* (configured to allow a maximum of 15 clusters) is given in Table 9.

Table 9: Number of clusters proposed as optimal by different methods in *NbClust*

No. of clusters	2	3	4	6	7	11	13	14	15
No. of methods	3	**11**	3	1	3	2	1	1	2

This would seem to be at odds with the results of the elbow plot method, however, as three clusters leaves just 68.5% of the variance explained. The discrepancies between these methods highlights the difficulty of this question. In this paper, I will follow the lead of Shosted et al (2015), however, in focusing on models where the *k* value matches the number of categories assigned by the researcher, that is, the four Chindwin Khamti tonemes.

5 Results and discussion

5.1 Wordlist corpus

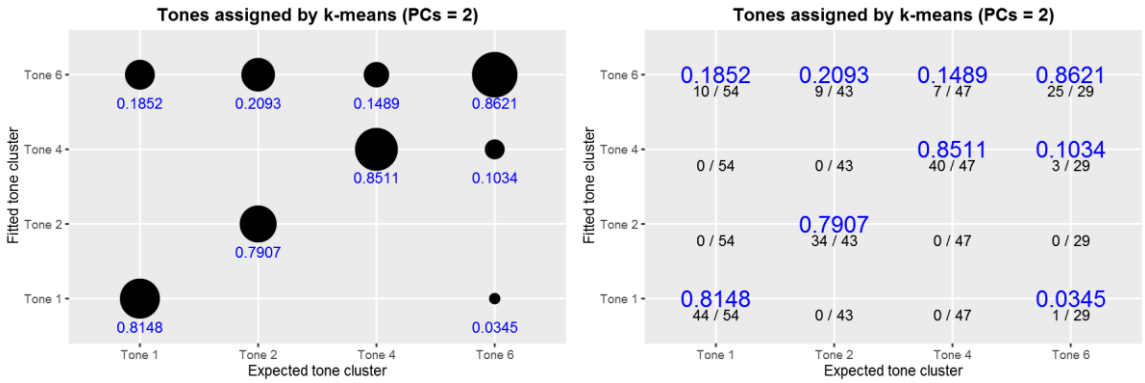
With input of 173 wordlist tokens from speaker LP5, a 25-year-old female, using two principal components, *k*-means produced the clusters summarized in Table 10. These have been subdivided by toneme for ease of comparison against expected categories.

Table 10: Output of *k*-means clustering (*k* = 4 speaker = LP5, measures = pitch)

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Total**
Total	**34**	**51**	**43**	**45**	**173**
Tone 1	0	10	0	44	54
Tone 2	34	9	0	0	43
Tone 4	0	7	40	0	47
Tone 6	0	25	3	1	29

Using the method described in §4.2.1 to match each fitted cluster with the expected toneme that it most overlaps with, we can visualize these results with the graphs in Figure 9.

Figure 9: Clusters assigned by *k*-means (*k* = 4, speaker = LP5, measures = pitch)



These graphs reveal very good performance in all clusters except the cluster corresponding to tone 6, which is assigned a sizable portion of tokens from all four expected categories. The precision, recall, and f-scores of the *k*-means output are given in Table 11.[6]

Table 11: Performance of *k*-means clustering (*k* = 4, speaker = LP5, measures = pitch)

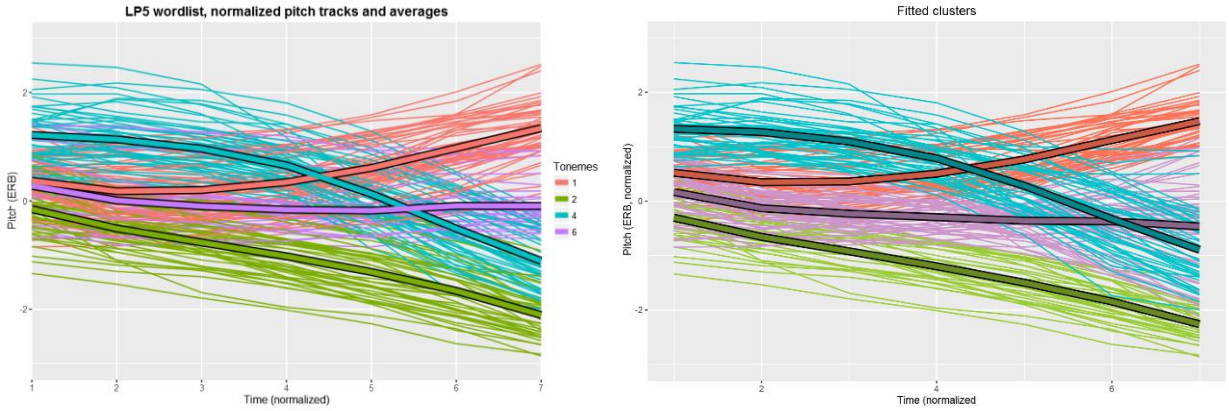
	Precision	**Recall**	**F-score**
Tone 1	0.97778	0.81481	0.88889
Tone 2	1.00000	0.79070	0.88312
Tone 4	0.93023	0.85106	0.88889
Tone 6	0.49020	0.86207	0.62500
Overall	**0.82659**	**0.82659**	**0.82659**

Matching the results seen in Figure 9, we get excellent precision for expected tones 1, 2, and 4, all well over 90%, and tone 2 even getting perfect precision. The precision for tone 6 is at 49%, but its recall is very high, meaning that nearly all of the actual tone 6 tokens were assigned to it, but that there were again as many false positives also assigned to that cluster, hence the poor

[6] Precision for a given toneme is calculated as the number of true positives (correctly clustered tokens) divided by the sum of true positives and false positives (tokens incorrectly clustered with that toneme). Recall is the number of true positives divided by the sum of true positives and false negatives (tokens that should have been clustered with that toneme but were not). F-score is the harmonic mean of precision and recall, calculated as 2 * (precision * recall) / (precision + recall).

precision. The normalized pitch tracks of the input are compared side by side with the fitted clusters in Figure 10.

Figure 10: Expected clusters (left) vs. fitted clusters (speaker = LP5, measures = pitch).

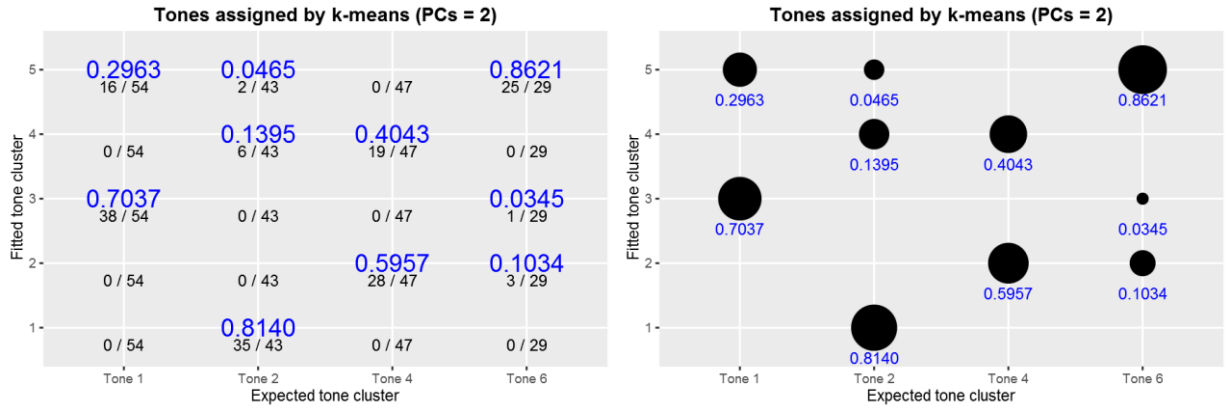


The cluster corresponding to tone 6 is slightly lower in the fitted model, but otherwise the average fitted tones are virtually identical to the expected clusters.

5.1.1 Testing other *k* values

To compare briefly with other possible models, if we take the consensus of *NbClust* for the wordlist corpus and model $k = 3$, we get the results shown in Figure 11.

Figure 11: Tone assignments (PCs = 2, *k* = 3, speaker = LP5, measures = pitch)



Although we have only three clusters, tones 1, 2, and 4 are all clustered by *k*-means as expected at rates above 90%. Tone 6, meanwhile, splits its mass predominantly between the two clusters corresponding to tones 1 and 2. The scores for this model are given in Table 12.

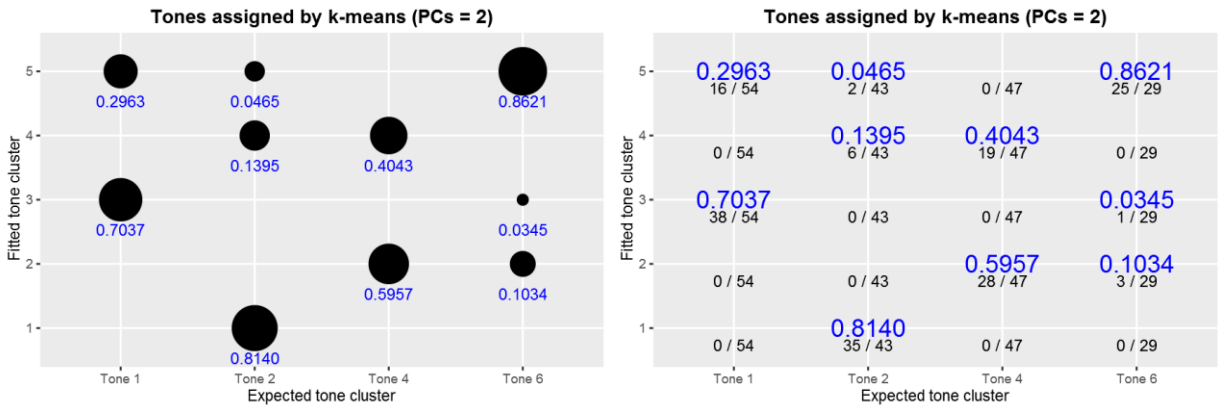
Table 12: Performance of *k*-means (PCs = 2, *k* = 3, speaker = LP5, measures = pitch)

	Precision	**Recall**	**F-score**
Tone 1	0.75362	0.96296	0.84552
Tone 2	0.73684	0.97674	0.83999
Tone 4	0.91489	0.91489	0.91489
Tone 6	--	--	--
Overall	**0.79191**	**0.79191**	**0.79191**

Recall improves for tones 1, 2 and 4, because there is no fourth cluster to split up the remaining three, but naturally precision suffers as a result. This seems to reflect the fact that tones 1, 2, and 4 are at the extremes of the tonal space, and are all contour tones, while tone 6, the mid-level tone, is in the center of the tonal space and overlaps to some degree with each of the other categories. Thus, when *k* = 4, the tokens from the other three categories that are closest to the center of the tonal space get misclassified by the *k*-means algorithm. Conversely, when *k* = 3, all of the tokens that should be assigned to tone 6 simply blend into the other three categories, but especially tones 1 and 2.

Turning to *k* = 5, on the other hand, we see the results shown in Figure 12.

Figure 12: Tone assignments (PCs = 2, *k* = 5, speaker = LP5, measures = pitch)



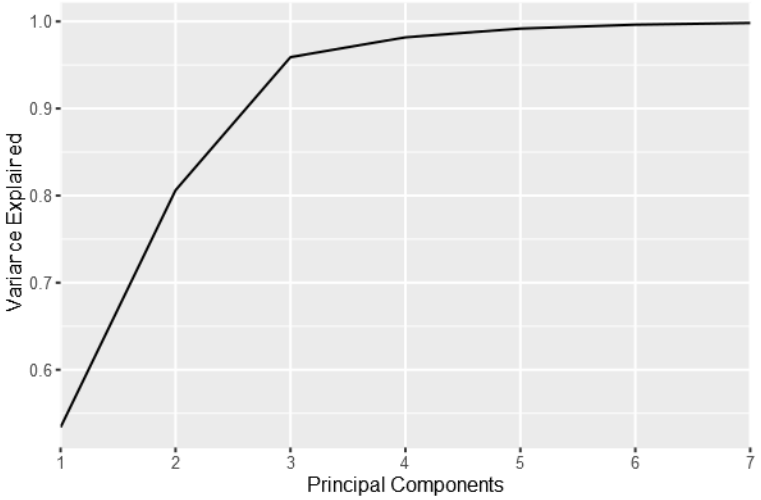
These figures indicate that tones 1, 2, and 6 correspond relatively well to clusters 3, 1, and 5, respectively. However, tokens from tone 4 are split in half between clusters 2 and 4. While this presents an interesting avenue to investigate what might cause that split, it is clear that the five-cluster model does not give us any obvious improvements over the model with four clusters. This does leave us with another option to see if we can improve our model: introducing additional phonetic measures besides pitch.

5.1.2 Introducing phonation measures

The only data provided to the principal components analysis so far was the 7-step pitch data. However, we can introduce additional measures to PCA in an attempt to improve performance: spectral tilt (H1-H2) measured over 3 intervals within each tone token. Spectral tilt is indicative of non-modal phonation, with creaky voice exhibiting shallower spectral tilt (Kirk et al 1993). Dockum (2015) claims that glottal constriction (or possible creaky phonation) is a historical conditioning factor for the emergence of tone 2 in modern Chindwin Khamti, so these additional data might be expected to improve performance.

Additional dimensions in our data require us to reevaluate the number of principal components we should optimally be feeding to *k*-means. The elbow plot for the variance explained in Figure 13 indicates that three PCs is optimal. The first two PCs explain 80.6% of the variance, while the third brings the total up to 95.9%.

Figure 13: Variance explained by PCs (speaker = LP5, measures = pitch + spectral tilt)



Since we have introduced a third principal component, we should examine the loadings again to determine what its likely phonetic correlate is, and see how if we can expect much to be gained from introducing phonation measures.

Table 13: Loadings for PC 1-3 (corpus = wordlist, measures = pitch + spectral tilt)

	PC1	PC2	PC3
Pitch step			
Step 1	-0.33973532	0.211421802	-0.36242291
Step 2	-0.37129187	0.155498357	-0.35394212
Step 3	-0.39807764	0.120794049	-0.25358796
Step 4	-0.42458736	0.065918618	-0.06373062
Step 5	-0.41076796	-0.006377505	0.23556202
Step 6	-0.34675064	-0.064300256	0.47182595
Step 7	-0.29507263	-0.095411381	0.56941716
H1-H2 step			
Step 1	-0.05706411	-0.564178097	-0.13072954
Step 2	-0.07821412	-0.572019417	-0.16238065
Step 3	-0.15024674	-0.503391448	-0.17088997

In PIC1 there is little to no movement at all in the loadings corresponding to spectral tilt (the bottom three rows), while in PC3 we see what appears to be another kind of contour change in the variables derived from the pitch steps. We might interpret this to mean that there is not significant ground to be gained in this particular corpus by adding spectral tilt. The clusters assigned by *k*-means with a three-PC model are given in Figure 14. Scores are given in Table 14.

Figure 14: Tone assignments (PCs = 3, *k* = 4, speaker = LP5, measures = pitch + spectral tilt)

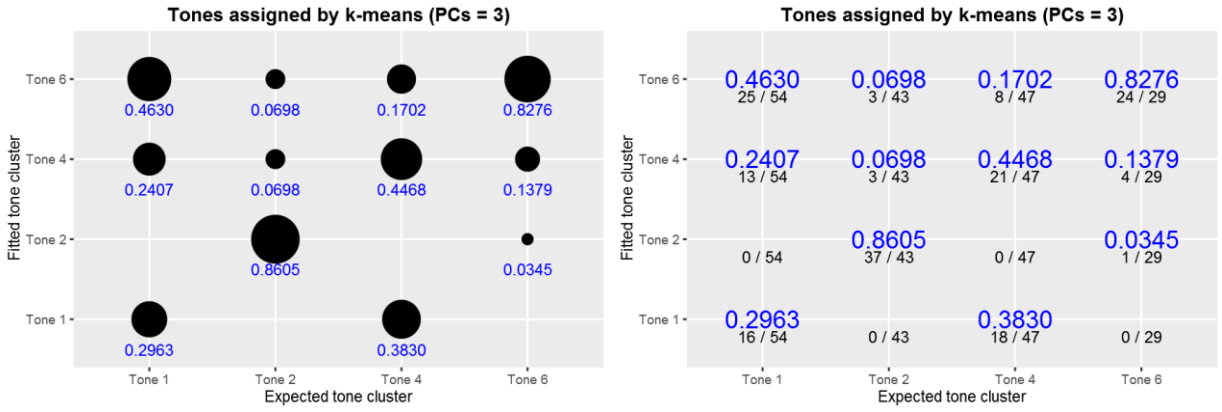


Table 14: Performance of *k*-means (PCs = 3, speaker = LP5, measures = pitch + spectral tilt)

	Precision	Recall	F-score
Tone 1	0.47059	0.2963	0.36364
Tone 2	0.97368	0.86047	0.91358
Tone 4	0.5122	0.44681	0.47727
Tone 6	0.4	0.82759	0.53933
Overall	0.56647	0.56647	0.56647

Indeed, the performance of most clusters suffers greatly as a result of introducing phonation measures, with the exception of tone 2. In this category, recall and thus f-score improve with the introduction of phonation measures and a third principal component, as shown in Table 15.

Table 15: Comparison of scores for tone 2

	Precision	Recall	F-score	PCs	Measures
Tone 2	1.00000	0.79070	0.88312	2	pitch
Tone 2	0.97368	0.86047	0.91358	3	pitch + spectral tilt

Though the effect is not very large, this likely is the result of tone 2 having non-modal voice in some portion of tokens, something easily confirmed by audition of Khamti recordings. This creaky phonation may act as a secondary phonetic cue to the tone category for Khamti speakers. This observation also supports the proposal in Dockum (2015) that this phonation conditioned the merger of multiple historical tonal categories into the modern tone 2 surface tone. Those historical tonemes would have had different tone shapes but shared creaky phonation, and eventually collapsed into a single tone, with remnant creakiness.

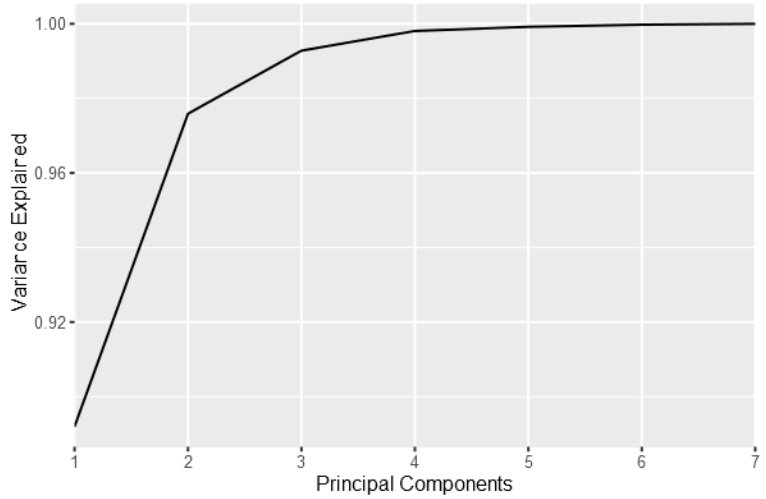
Ultimately, the two-PC model with only pitch data is the best performing model. This serves as a nice baseline for comparison against the more complex corpus of stimuli responses.[7]

5.2 Stimuli response corpus

As described in §3.3, the corpus of stimuli responses is the larger of the two corpora used, consisting of 750 tokens from five speakers. Starting with only the pitch measures, we can determine the number of principal components we need from the elbow plot in Figure 15.

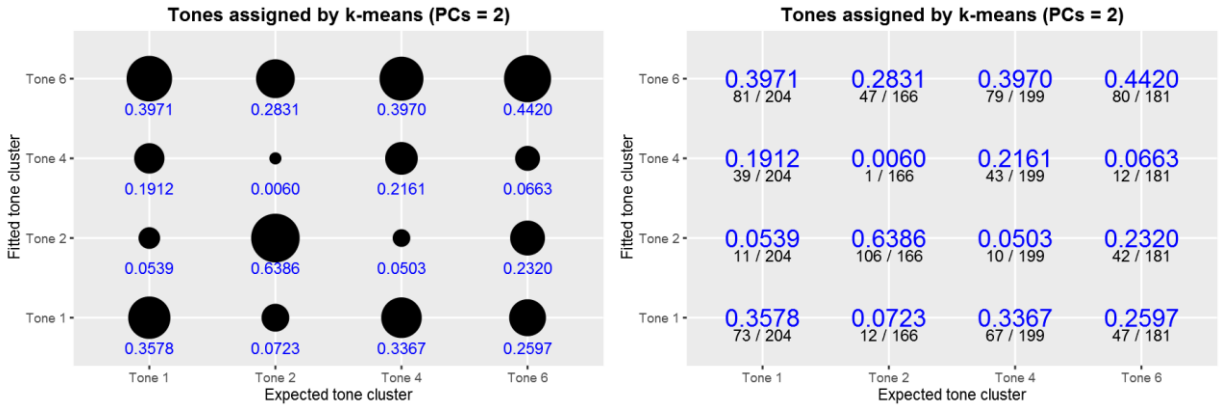
[7] Additional models were tested include measures of duration, jitter, and shimmer. None improved clustering.

Figure 15: Variance explained by PCs (5 speakers, corpus = stimuli, measures = pitch)



The first two principal components explain 97.6% of the variance, so that is the number to be used with *k*-means for the pitch data in this corpus. The results of *k*-means clustering are given in Figure 16.

Figure 16. Tone assignments (PCs = 2, *k* = 4, 5 speakers, corpus = stimuli, measures = pitch)



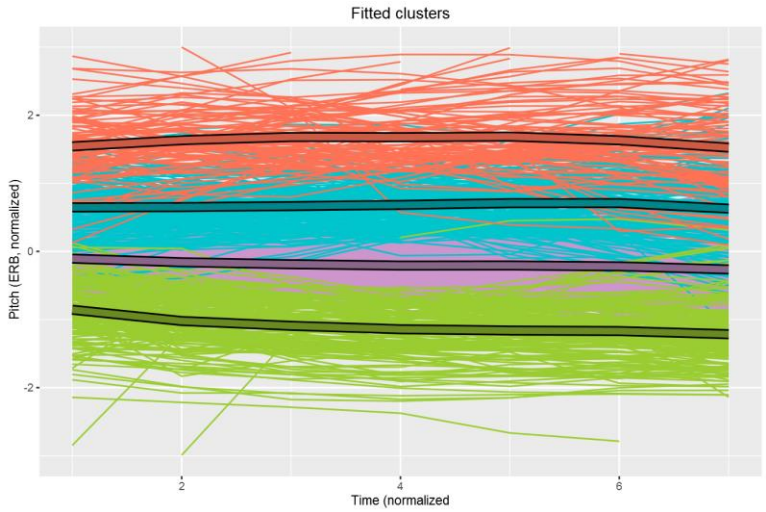
The single largest overlap is 63% of tokens from tone 2, but the poor fit of this model to the data is driven home by the precision, recall, and f-scores, given in Table 16.

Table 16: Performance of *k*-means (PCs = 2, 5 speakers, corpus = stimuli, measures = pitch)

	Precision	Recall	F-score
Tone 1	0.36683	0.35784	0.36228
Tone 2	0.62722	0.63855	0.63284
Tone 4	0.45263	0.21608	0.29252
Tone 6	0.27875	0.44199	0.34188
Overall	0.40267	0.40267	0.40267

Further, the average normalized pitches for the assigned clusters bear little resemblance to the tonemes of Khamti. This is seen in Figure 17.

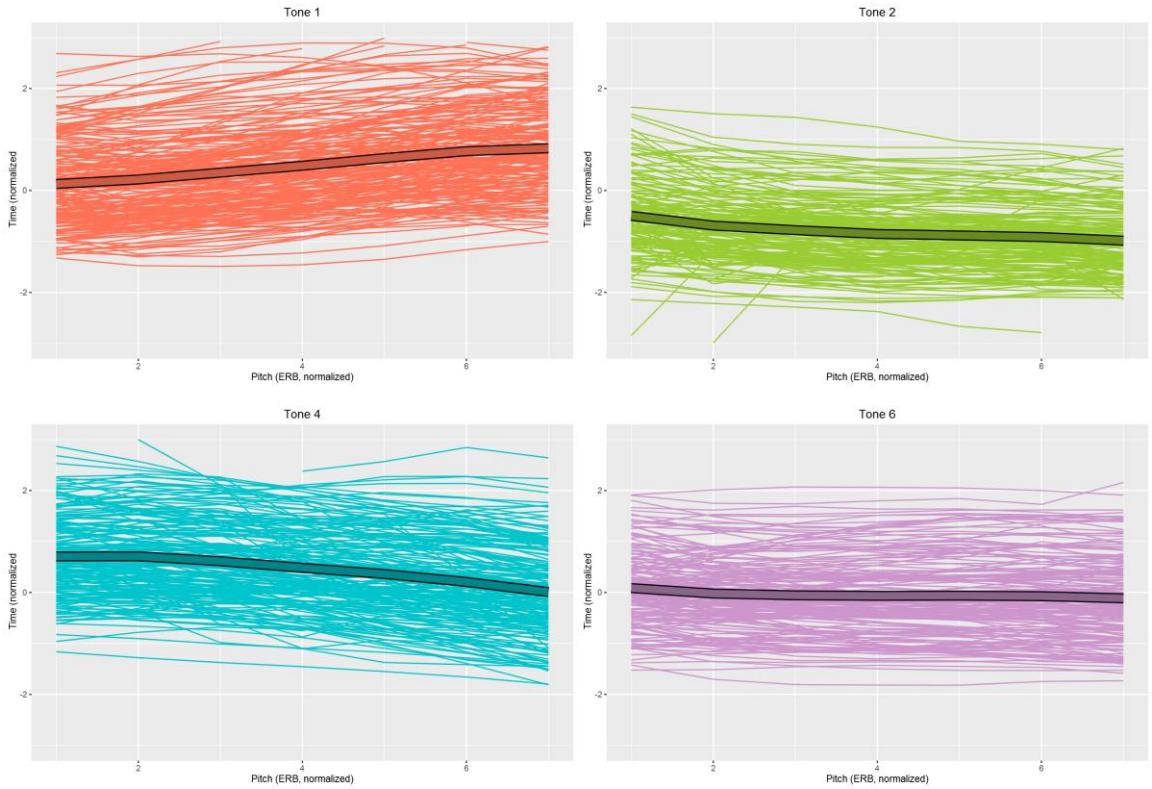
Figure 17: Fitted tone clusters (5 speakers, corpus = stimuli, measures = pitch)



There appears to be so much variation in the data that tone contours do not emerge from the clustering at all, and we are left with what is essentially four level tones of different heights. To understand this result, it is useful to step back and examine both the input and output more closely.

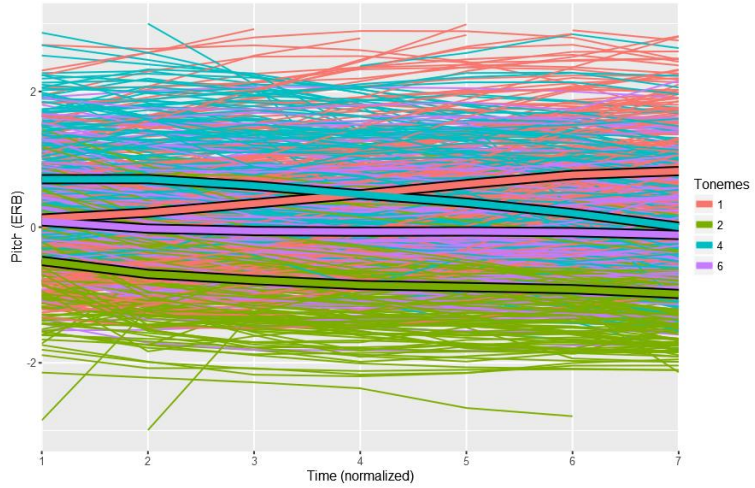
The pitch tracks of the expected clusters from the stimuli corpus are given in Figure 18. Even though this data has been normalized, the spread of possible pitch tracks for each tone makes it perhaps unsurprising that the clustering algorithm has performed so poorly.

Figure 18: Pitch tracks of expected tone clusters (5 speakers, corpus = stimuli)



The messiness of the stimuli corpus comes into sharp focus when these are overlaid one on top of another in Figure 19. Though, notably, these look like the same tonemes seen in Figure 4, just with a compressed pitch space.

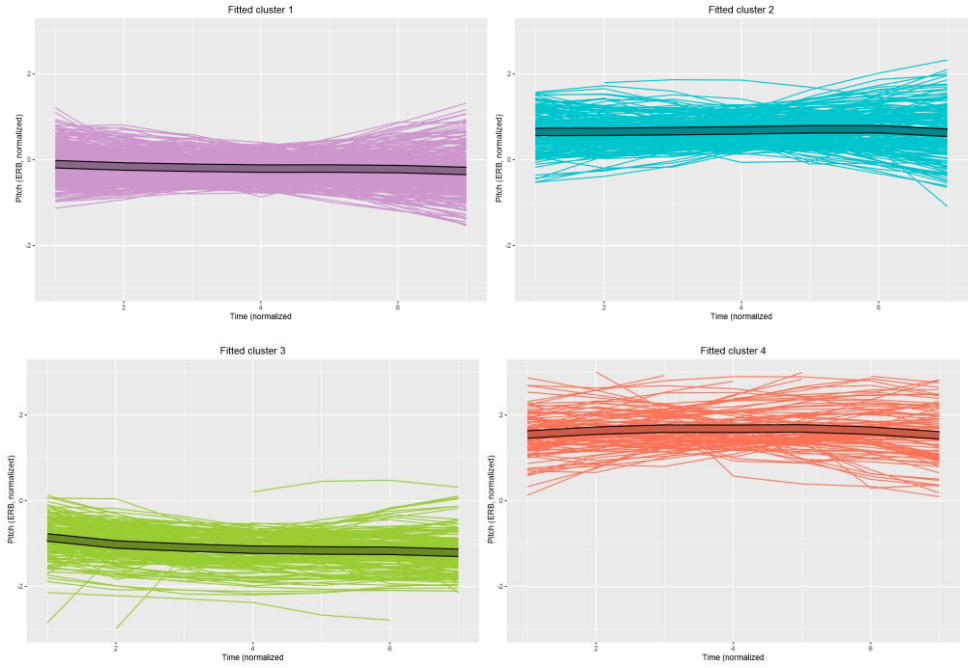
Figure 19: Expected tone clusters (5 speakers, corpus = stimuli)



Tone sandhi is typologically unexpected for this language family, so the reason for this is likely to be the result of general intraspeaker variation in producing the pitch targets, combined with tone boundary effects and other prosodic effects like focus.

The fitted clusters from this model, especially 1 and 2, show a bowtie shape in the normalized pitch tracks, as seen in Figure 20. The bowtie shapes indicate that both rising and falling contours are being clustered together, which supports the idea that the only thing this model is really distinguishing is overall pitch height, as shown in Figure 16.

Figure 20: Fitted clusters (PCs = 2, *k* = 4, 5 speakers, corpus = stimuli)

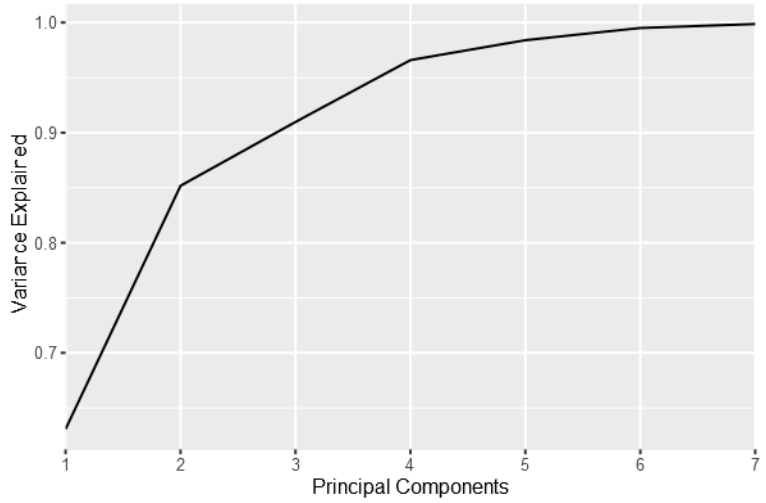


Once again, we can turn to additional measures to attempt to achieve a better fit.

5.2.1 Introducing phonation measures

The elbow plot in Figure 21 indicates that the number of principal components we will need with the expanded dataset is four.

Figure 21: Variance explained by PCs (5 speakers, corpus = stimuli, measures = pitch + spectral tilt)



And once again, running our *k*-means clustering with a four-PC model using pitch and spectral tilt measures, we get the clusters in Figure 22 and the scores in Table 17.

Figure 22: Tone assignments (PCs = 4, *k* = 4, 5 speakers, corpus = stimuli, measures = pitch + spectral tilt)

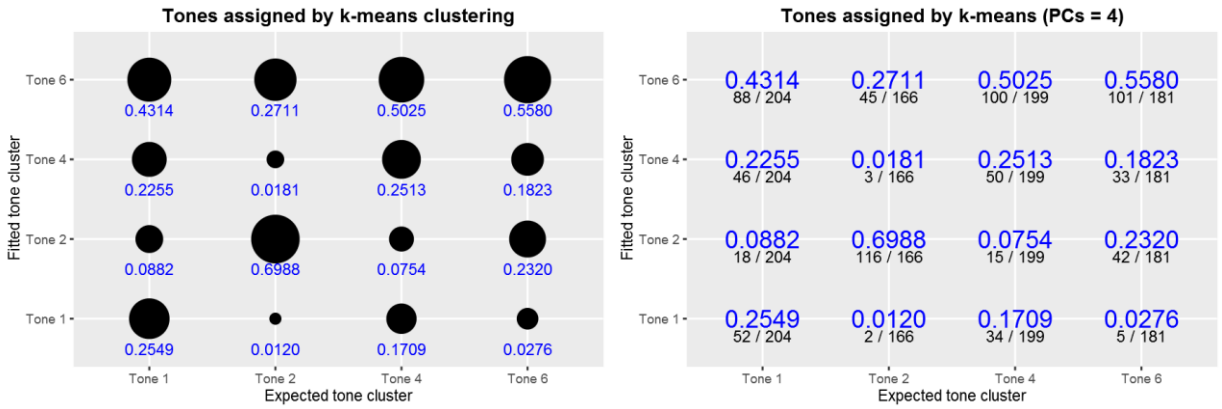


Table 17: Performance of *k*-means (PCs = 4, 5 speakers, corpus = stimuli, measures = pitch + spectral tilt)

	Precision	Recall	F-score
Tone 1	0.55914	0.2549	0.35017
Tone 2	0.60733	0.6988	0.64986
Tone 4	0.37879	0.25126	0.30211
Tone 6	0.3024	0.55801	0.39223
Overall	0.42533	0.42533	0.42533

Whatever marginal improvement there is in some categories is offset by others that perform worse, and overall once again this is a very poor fit for the data, even though we have expanded the dataset to include spectral tilt information.

6 Conclusion

Computational modeling of lexical tone by combining principal components analysis with *k*-means clustering produced very promising results on the small single-speaker wordlist corpus. These results motivate the markup of additional data to create a larger database, and to expand across all syllable shapes in the lexicon. This method also uncovered non-modal phonation in tone 2, which might be a phonetic cue used by speakers, though this requires further study.

The performance on the larger corpus of stimuli responses was poor, apparently due to the considerable intraspeaker variation in tone shapes in running speech. Further tagging of the corpus to account for prosodic focus and position within the utterance may help to make this dataset more tractable. As is, however, the low performance serves to highlight the limitations of the technique. Given a larger corpus of properly annotated tone tokens, this method has good potential for relatively easy comparison between speakers or speaker groups. For instance, coding several parameters such as age, sex, and village location, along with toneme, one could quite readily compare each tone category pairwise with each other tone category across all of these dimensions, and identify vectors of sociolinguistic variation in the tonal space. In the pursuit of individual tonal variation, it may also be useful in identifying tonal splits or mergers in progress. This analytical toolset does come at the cost of the time investment in marking up the tone corpus and processing the input and output data. Development of a method that reduces or removes the need for manual markup for running speech would greatly increase the utility of these methods.

Tonal categorization is a promising area to apply these computational techniques to. This paper highlights some of their successes and weaknesses, which helps to advance their development and refinement. Improvement of these methods has the potential to greatly benefit linguists conducting language documentation on tonal languages.

REFERENCES

BAAYEN, R. H. 2008. *Analyzing linguistic data: A practical guide introduction to statistics using R*. Cambridge University Press.

BENNETT, RYAN. 2015. All measures extractor. Script for the Praat software program. Provided by personal communication.

BOERSMA, PAUL, and DAVID WEENINK. 2016. Praat: doing phonetics by computer [Computer program]. Version 5.3.7.1. <http://www.praat.org>

CHAO, YUEN-REN. 1930. A system of tone-letters. *Le maître phonétique* 45.24–27.

CHARRAD, MALIKA; NADIA GHAZZALI; VERONIQUE BOITEAU; and AZAM NIKNAFS. *NbClust*. package for R <https://cran.r-project.org/web/packages/NbClust/NbClust.pdf>

DOCKUM, RIKKER. 2015. Tonal evidence and language classification: The place of Khamti in Southwestern Tai. New Haven: Yale University, Unpublished MS.

HALKIDI, M.; M. VAZIRGIANNIS; and I. BATISTAKIS. 2000. Quality scheme assessment in the clustering process. *Proceedings of PKDD*, Lyon, France.

HARRIS, JIMMY G. 1976. Notes on Khamti Shan. *Tai linguistics in honor of Fang-Kuei Li*. 113–141.

HARTIGAN, J. A., and M. A. WONG. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, Series C 28.1.100–108.

INGLIS, DOUGLAS. 2014. This here thing: Specifying morphemes an3, nai1, and mai2 in Tai Khamti reference-point constructions. PhD Dissertation. University of Alberta.

JOHNSON, KEITH. 2008. *Quantitative methods in linguistics*. Wiley-Blackwell.

JOLLIFFE, I. T. 1986. Principal component analysis. Springer-Verlag. <doi:10.1007/b98835>

KETCHEN, DAVID J., and CHRISTOPHER L. SHOOK. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal* 17.6.441–458.

KIRK, P. L.; LADEFOGED, J.; and LADEFOGED, P. 1993. Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. *American Indian linguistics and ethnography in honor of Laurence C. Thompson*, ed. by A. Mattina and T. Montler, 435–450. Missoula, MT: University of Montana Press.

KRZANOWSKI, W. J., and Y. T. LAI. 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44.1.23–34.

LEWIS, M. PAUL; GARY F. SIMONS; and CHARLES D. FENNIG (eds.) 2016. *Ethnologue: Languages of the World, Nineteenth edition.* Dallas, Texas: SIL International. <http://www.ethnologue.com.>

LOW, JAMES. 1828. *A grammar of the Thai, or Siamese language*. Calcutta: Baptist Mission Press.

MOREY, STEPHEN. 2005. Tonal change in the Tai languages of Northeast India. *Linguistics of the Tibeto-Burman Area* 28.2.139–202.

NEEDHAM, J. F. 1894. *Outline Grammar of the Tai (Khamti) Language*. Rangoon: Superintendent, Government Printing, Burma.

R CORE TEAM. 2015. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>

ROBINSON, WILLIAM. 1849. Notes on the languages spoken by the various tribes inhabiting the valley of Assam and its mountain confines [Section: The Khamti]. *Journal of the Royal Asiatic Society of Bengal* 18.1.183–237, 310–349.

SCOTT, A. J., and M. J. SYMONS. 1971. Clustering methods based on likelihood ratio criteria. *Biometrics* 27.2.387–397.

SHOSTED, RYAN; MARISSA BARLAZ; and DI WU. 2015. Modeling Iu Mien tone with eigenpitch representations. *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0728.

SIL INTERNATIONAL. 2002. Mainlaind Southeast Asia comparative wordlist.

SNIDER, KEITH. 2014. On establishing underlying tonal contrast. *Language Documentation & Conservation* 8.707–737.

STANFORD, JAMES N. 2008. A sociotonetic analysis of Sui dialect contact. *Language Variation and Change* 20.3.409–50.