Leveraging Deep Learning to Shed Light on Tones of an Endangered Language: A Case Study of Moklen

Sireemas Maspong[1,2], Francesco Burroni[1,2], Teerawee Sukanchanon[2], Warunsiri Pornpottanamas[3], Pittayawat Pittayaporn[2]

[1]Spoken Language Processing Group, Institute for Phonetics and Speech Processing, LMU München
[2]Department of Linguistics & Center of Excellence in Southeast Asian Linguistics, Chulalongkorn University
[3]Department of English and Linguistics, Ramkhamhaeng University

Correspondence: s.maspong@phonetik.uni-muenchen.de

Abstract

Moklen, a tonal Austronesian language spoken in Thailand, exhibits two tones with unbalanced distributions. We employed machine learning techniques for time-series classification to investigate its acoustic properties. Our analysis reveals that a synergy between pitch and vowel quality is crucial for tone distinction, as the model trained with these features achieved the highest accuracy.

1 Introduction

Moklen, an endangered and understudied Austronesian language spoken along the western coast of southern Thailand (Larish, 2005), has sparked debate about its tonal status. While Austronesian languages are typically not tonal, Moklen exhibits a few minimal pairs suggesting the presence of two lexical tones (Larish, 1997; Pittayaporn et al., 2022).

The acoustic properties of Moklen tone were recently explored by Pornpottanamas et al. (2023). Their study revealed that Moklen tones are distinguished not only by pitch, but also by vowel quality and voice quality. Interestingly, these acoustic characteristics resemble those of register contrasts found in mainland Southeast Asian languages (Brunelle and Kirby, 2016). It is worth noting that the definition of tone in this paper refers to the suprasegmental contrast, which may be realized not only by pitch, but also by voice quality or vowel quality, similar to Vietnamese, Burmese, Shanghai Chinese, and other languages (See Abramson and Luangthongkum, 2009; Brunelle and Kirby, 2016).

What remains unclear is the relative weight of acoustic cues in Moklen tones and register systems. Phonetic contrasts often differ across multiple dimensions; for example, the English /b/ and /p/ differ in their voice onset time (VOT) as well as other dimensions, including the duration of stop closure and fundamental frequency (f0) after closure (Lisker, 1986). Furthermore, even though a contrast may involve several phonetic dimensions, they may not all be equally important. In other words, the phonetic cues may have different weights in production and/or perception. For instance, the English /b/ and /p/ are primarily distinguished by VOT, with f0 playing a secondary role (Abramson and Lisker, 1985). It is therefore possible that the acoustic cues in Moklen tones, including pitch, vowel quality, and voice quality, may have different relative weights.

In this paper, we investigate the contribution of individual acoustic features to Moklen tone distinction using an ablation study within a machine learning framework. We employed a Bidirectional Long Short-Term Memory (BiLSTM) Neural Network with self-attention for sequence classification. BiLSTM with self-attention has been used in tone recognition tasks in previous works (e.g., Yang et al., 2018). However, neural network classification has rarely been used with the tones of underrepresented languages such as Moklen.

This investigation confirms the presence of contrastive tones in Moklen. Furthermore, our analysis reveals that pitch and vowel quality features are crucial for distinguishing the two lexical tones. The model trained on this feature set achieved the highest accuracy in differentiating between Moklen tones.

1.1 Moklen and its lexical tones

Moklen is an indigenous language spoken by fewer than 4,000 people along the west coast of Phang Nga province in Thailand and on nearby islands (Arunotai, 2017). Currently, the language is facing endangerment, as its use is limited to older adults with low transmission to younger speakers (Pittayaporn et al., 2022).

Phonologically, Moklen shares similarities with mainland Southeast Asian (MSEA) languages, setting it apart from the broader insular Austronesian

Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024), pages 37–42

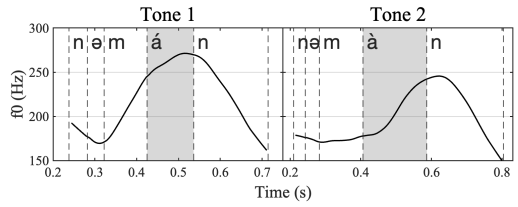

Figure 1: Difference in f0 of a minimal pair /nəmán/ ‘to fish’ *vs.* /nəmàn/ ‘to be glad’.

Tone 1		Tone 2	
Words	**Glosses**	**Words**	**Glosses**
nəmán	‘to fish’	nəmàn	‘to be glad’
bəlá:	‘to scold’	bəlà:	‘dehusked rice’
nəmá:ʔ	‘to enter’	dadà:ʔ	‘breast’
ʔá:k	‘to place’	ʔà:k	‘crow’
namát	‘wave, tide’	digàt	‘bedbug’
kɔlá:t	‘to be hot’	kɔlà:t	‘mushroom’

Table 1: Examples of stimuli.

family (Larish, 1999; Pittayaporn, 2024). Two features relevant to this study, shared by Moklen and other MSEA languages, are systematic word-final stress and tonal contrast.

Moklen follows a consistent iambic stress pattern, with stress assigned to the last syllable of the foot (Larish, 1999; Swastham, 1982). Moklen tones are consistently realized only on the ultimate syllable, which also bears stress (Pittayaporn et al., 2022; Pornpottanamas et al., 2023). The two tones are not predictable from any phonological environments despite an unbalanced distribution: the majority of words carry Tone 1, while only about 10-20% carry Tone 2 (Larish, 1997). A few minimal pairs have been identified, as shown in Table 1.

Acoustically, the two tones differ in several ways. Tone 1 is generally higher-pitched compared to Tone 2, which has a lower pitch and a steeper rise on the stressed vowel (Figure 1). Additionally, Tone 1 vowels tend to be lower and slightly more front compared to Tone 2 vowels. Finally, Tone 2 exhibits breathiness, while Tone 1 is more modal. These acoustic properties remain consistent regardless of vowel length, onset voicing, or coda categories (Pornpottanamas et al., 2023).

While previous research has identified acoustic correlates of Moklen tones, including pitch, vowel quality, and voice quality, one question remains unanswered: the relative importance of these features in distinguishing the two lexical tones. It is unclear whether all features contribute equally or if a specific combination proves most effective. Investigating this question can provide deeper insights into the acoustic realization of Moklen tones and potentially contribute to the development of more efficient automatic speech recognition systems for Moklen.

1.2 Research questions

This paper investigates two key questions regarding Moklen tone:

(i) Can pitch, voice quality, and vowel quality features be used to distinguish the two Moklen tones?

(ii) Which combination of these acoustic features leads to the most accurate classification of Moklen tones?

2 Methodology

2.1 Data collection and processing

Eight native Moklen speakers from Phang Nga Province participated in this study. Four participants (3 females, 1 male) resided in Bang Sak village, while the remaining four (3 females, 1 male) resided in Lam Pi village. Although the participants are from two different villages, previous research has not observed dialectal differences between them (Pornpottanamas et al., 2023).

The participants ranged in age from 46 to 70 years old at the time of recording. Notably, all participants were bilingual in Moklen and Southern Thai, with Moklen being their dominant language.

The participants were instructed to produce Moklen monosyllabic and disyllabic words in isolation. The stimuli were presented orally in Thai, and participants were asked to translate them into Moklen. Each target word was repeated three times.

The stimuli consist of 98 attested Moklen words with stressed final syllables containing /a/ or /a:/ vowels. These target words were systematically varied in terms of tone, onset voicing, vowel length, and coda classes to achieve a balanced representation. Examples of the stimuli are provided in Table 1. Notably, there are 74 words with Tone 1 and 24 words with Tone 2. This unequal distribution of stimuli roughly reflects the actual proportion of these two tones within the Moklen lexicon. We did not control for the semantic or syntactic categories of the target words.

The recordings were manually segmented in Praat (Boersma and Weenink, 2020). From the stressed vowel intervals, five acoustic measurements were extracted to serve as time-series fea-

tures in the classification process: fundamental frequency (f0) for pitch, first and second formant frequencies (F1, F2) characterizing vowel quality, the difference between corrected first harmonics and corrected spectral amplitude of F3 (H1*-A3*) (using the correction method from Iseli and Alwan, 2004), and Cepstral Peak Prominence (CPP) as measures of voice quality. These measurements are commonly reported as acoustic correlates of tone in Southeast Asian languages (Brunelle and Kirby, 2016). Measurements during the vowel interval were chosen over the rime (vowel and coda) interval because our target words include those with final voiceless stops. Many of these measurements, especially f0, F1, and F2, cannot be tracked during the voiceless stop coda interval. Therefore, measurements during the vowel interval provide the only fair comparison across all syllable structures.

PraatSauce (Kirby, 2018) was used to extract these acoustic measurements. A consistent window size of 30 milliseconds (ms) with a 5 ms time step was applied to all measurements. f0 tracking was performed in two steps to account for individual variations in f0 range across participants, following the method described in De Looze (2010).

To standardize the acoustic measurements, each participant's data were z-scored based on participant-specific mean and standard deviation.

2.2 Data preparation

To prepare the data for classification analysis, we first addressed missing values due to tracking errors using the fillmissing function in MATLAB (MathWorks, 2024), employing linear interpolation of neighboring, non-missing values. Trajectories with too few existing values that could not be adequately filled were removed. The remaining number of tokens for classification is 1,684 for Tone 1 and 567 for Tone 2.

We randomly partitioned the data into an 80:10:10 split for training, validation, and testing sets, respectively, using the cvpartition function in MATLAB. The training set contained 1,801 tokens (1,353 tokens of Tone 1 and 448 tokens of Tone 2), the validation set included 225 tokens (157 tokens of Tone 1 and 68 tokens of Tone 2), and the testing set comprised 225 tokens (174 tokens of Tone 1 and 51 tokens of Tone 2).

Due to the imbalanced class distribution, we upsampled Tone 2 tokens in the training set to match the number of Tone 1 tokens. To achieve a more robust classification, we augmented the training data using two methods adapted from Flores et al. (2021): time-warping and adding random Gaussian noise. We time-warped each token to a length randomly drawn from a Poisson distribution with a lambda parameter corresponding to the mean length of all tokens. Then, we added Gaussian noise with a standard deviation of 0.05 to all measurements of all tokens. Finally, we combined the permuted data with the original data to enlarge the training set. In total, our training set included 2,706 tokens for each tonal category.

Hyperparameters	Ranges	Optimized Values
# Hidden Layers	[1, 4]	1
# Hidden Units	[16, 64]	52
Batch Size	[16, 64]	23
Initial Learning Rate	[10^{-6}, 0.005]	0.0032

Table 2: Search ranges for Bayesian Optimization and the optimized values.

2.3 Sequence classification using bidirectional LSTM with Self-Attention

To classify Moklen tone, we trained a Bidirectional Recurrent Neural Network with Long Short-Term Memory units (BiLSTM). BiLSTM is well-suited for sequential tasks like speech recognition (Graves and Schmidhuber, 2005). Additionally, we enhanced the model by incorporating a self-attention mechanism to focus the network on the most relevant parts of the input sequence for tone classification.

The BiLSTM architecture consisted of an input layer with five units (one for each acoustic measurement), hidden layers using a sigmoid activation function, and an output layer with two units (one for each tone class), followed by a softmax layer for probability estimation. Additionally, recurrent dropout was applied to the hidden layer for regularization.

Other hyperparameters, including the number of hidden layers, number of hidden units, batch size, and initial learning rate, were optimized using Bayesian Optimization. The search ranges for Bayesian Optimization and the optimized values were summarized in Table 2.

2.4 Feature ablation

To assess the contribution of different acoustic feature sets to tone classification, we conducted a feature ablation study. We trained separate classification models with seven feature combination inputs:

	Features	Overall Acc.	Tone 1 Acc.	Tone 2 Acc.
(iii)	**Pitch+Vowel**	**84.0%**	**86.2%**	**76.5%**
(i)	Pitch+Voice+Vowel	81.3%	85.6%	66.7%
(v)	Pitch	79.6%	83.9%	64.7%
(ii)	Pitch+Voice	79.1%	83.3%	64.7%
(iv)	Voice+Vowel	78.2%	83.3%	60.8%
(vi)	Voice	73.3%	72.4%	62.7%
(vii)	Vowel	70.2%	77.6%	58.8%

Table 3: Accuracy of models with different feature combinations sorted based on the total accuracy.

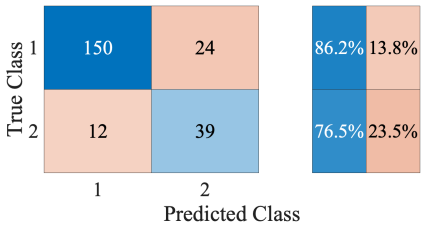

Figure 2: Confusion matrix of the model with pitch and vowel quality features.

(i) Pitch (f0), voice quality (CPP and H1*-A3*), and vowel quality (F1 and F2) features.
(ii) Pitch and voice quality features.
(iii) Pitch and vowel quality features.
(iv) Voice quality and vowel quality features.
(v) pitch features only.
(vi) Voice quality features only.
(vii) Vowel quality features only.

For a fair comparison across models, we applied the hyperparameters optimized using the model with all five feature inputs, as listed in (i), to all ablation models.

3 Results

3.1 Ablation study

We found that the performance of all models significantly exceeded the chance level (50% overall accuracy). Specifically, all models achieved an overall classification accuracy of over 70%, as shown in Table 3. The model using pitch (f0) and vowel quality (F1 and F2) features achieved the highest overall accuracy (84%) and F1-score (0.89). The confusion matrix of the model is illustrated in Figure 2. We also observed the importance of pitch information, as models excluding the pitch features exhibited lower performance, achieving the lowest accuracy among all models (Table 3).

An interesting observation is that the model using only pitch (f0) and vowel quality (F1 and F2) features exhibited significantly better performance

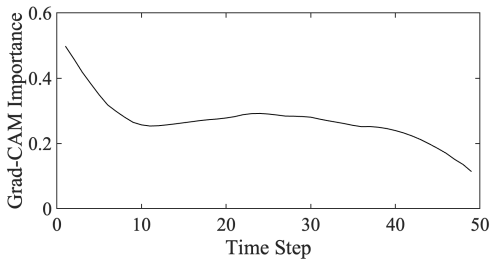

Figure 3: Grad-CAM importance map for a representative token classified by the best performing model.

in classifying Tone 2 tokens (76.5% accuracy) compared to other models (all below 70% accuracy for Tone 2). This behavior contrasts with the classification of Tone 1 tokens, where all models with pitch features performed similarly.

To understand which parts of the vowel trajectory contribute most to tone classification, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2019). Figure 3 illustrates the Grad-CAM importance map for a representative token classified by our best-performing model. As evident from the map, the model focuses heavily on the vowel's onset, with the importance decreasing gradually towards the end. This observation aligns with the f0 trajectories presented in Figure 1, suggesting that the initial portion of the vowel plays a crucial role in distinguishing the tones.

3.2 Error analysis

We also conducted an error analysis on the best-performing model, (iii) Pitch + Vowel. We examined whether the following four features had an effect on the model's classification: onset voicing (voiced, voiceless), vowel length (short, long), coda manners of articulation (stop, nasal, fricative, glide, open syllable), and coda places of articulation (bilabial, alveolar, palatal, velar, glottal, open syllable).

To determine if any of these features affected the model's classification, we performed Chi-squared tests. Each feature was tested separately against the correct and incorrect classifications of the model as one of the variables.

Significant effects were observed for two features: vowel length ($\chi^2(1, 225) = 7.02, p = .008$) and coda places of articulation ($\chi^2(5, 225) = 14.82, p = .011$). The other two features did not show significant effects: onset voicing ($\chi^2(1, 225) = .33, p = .56$) and coda manners of

articulation ($\chi^2(4, 225) = 8.46, p = .08$). These results suggest that vowel length and coda place of articulation significantly impacted the model's classification performance.

Regarding vowel length, it was found that words with long vowels were more likely to be misclassified than those with short vowels, with approximately 72% of the misclassified tokens being words with long vowels.

In terms of coda places of articulation, tokens with a velar coda were more frequently misclassified than those with other types of final consonants. Specifically, about 30% of tokens with velar codas were incorrectly classified, compared to only 12% of tokens with alveolar codas. Notably, none of the tokens with bilabial codas were misclassified.

We also examined whether unique words influenced the model's classification. However, no patterns were observed, leading us to conclude that unique words were not a direct factor in the model's errors.

4 Discussion and conclusion

Our investigation into Moklen tone classification using acoustic features sheds light on the nature of tones in this unique Austronesian language. The ablation study confirmed that all features (pitch, voice quality, and vowel quality) contribute to Moklen tone classification. This is evidenced by the findings that models utilizing only a single feature set representing each acoustic aspect achieved relatively good performance (> 70% accuracy). However, the model combining pitch and vowel quality achieved the highest overall accuracy and F1-score. This result suggests that a synergy between pitch and vowel quality information plays a crucial role in distinguishing the two Moklen tones.

One potential explanation for the importance of pitch and vowel quality in distinguishing Moklen tones lies in their historical development. As mentioned, tonal contrast in Moklen is an innovation absent in its ancestral language. Moklen tones may have developed from reanalyzing different contrasts, such as stress, that utilize pitch and vowel quality (Gordon and Roettger, 2017)

Furthermore, we observed that the models excluding the pitch features achieved the lowest accuracy. This finding confirms that pitch emerges as the primary cue for Moklen tone. On the other hand, although other acoustic cues can be used to distinguish the two Moklen tones, they appear to play a more secondary role.

Our analysis of the features' relative importance across time steps within the vowel interval revealed that the most important features cluster around the vowel onset. This suggests that the distinction between Tone 1 and Tone 2 is most salient at the onset of the vowel. This pattern closely aligns with register contrast, where the distinction between registers is most prominent at the vowel onset (Brunelle and Tạ, 2021).

We also conducted an error analysis on the best-performing model, examining four features: onset voicing, vowel length, coda manners, and coda places of articulation. Chi-squared tests revealed that vowel length and coda places of articulation significantly impacted the model's classification, with words having long vowels and velar codas being more frequently misclassified. Further investigation is needed to understand why vowel length and coda place of articulation affected the model's performance.

One potential aspect for future work is to investigate Moklen tones from the perspective of acoustic features within a larger time interval, such as the entire syllable rather than just the vowel interval used in this paper. In other words, there may be more aspects of the tones that we have not yet explored. This broader analysis could include features like the f0 peak location on the final open syllable or final syllable with sonorant coda, as shown in Figure 1, where Tone 1 generally exhibits an earlier peak compared to Tone 2.

Further investigation into the perception of the two tones by Moklen speakers could provide deeper insights into the nature of this unique tonal system.

This study demonstrates the potential of machine learning approaches for analyzing acoustic features in endangered languages like Moklen. By leveraging deep learning for tone classification, we can gain valuable insights into the sound system of a language, even with limited documentation or speaker availability. One limitation of Moklen tone documentation is that tones are not predictable from the phonological environment or comparative studies, making it challenging for language fieldworkers to identify tones in Moklen words. Classification models trained on words with identified tones can assist fieldworkers in identifying the tones of undocumented words. Furthermore, these models can aid in creating a dictionary of Moklen, which is an important step in language revitalization.

References

Arthur Abramson and Theraphan Luangthongkum. 2009. A fuzzy boundary between tone languages and voice-register languages. In G. Fant, H. Fujisaki, and J. Shen, editors, *Frontiers in phonetics and speech science*, pages 149–155. The Commercial Press, Beijing.

Arthur S. Abramson and Leigh Lisker. 1985. Relative power of cues: F0 shift versus voice timing. In Victoria A. Fromkin, editor, *Phonetic linguistics: Essays in honor of Peter Ladefoged*, pages 25–33. Academic Press, Orlando.

Narumon Arunotai. 2017. "Hopeless at sea, landless on shore": contextualising the sea nomads' dilemma in Thailand. *AAS working papers in social anthropology*, 31:1–27.

Paul Boersma and David Weenink. 2020. Praat: doing phonetics by computer.

Marc Brunelle and James Kirby. 2016. Tone and Phonation in Southeast Asian Languages. *Language and Linguistics Compass*, 10(4):191–207.

Marc Brunelle and Thành Tấn Tạ. 2021. Register in languages of Mainland Southeast Asia: the state of the art. In Paul Sidwell and Mathias Jenny, editors, *The languages and linguistics of Mainland Southeast Asia: A comprehensive guide*, pages 683–706. De Gruyter Mouton, Berlin/Boston.

Céline De Looze. 2010. *Analyse et interprétation de l'empan temporel des variations prosodiques en français et en Anglais*. Ph.D. Thesis, Université de Provence-Aix-Marseille I.

Anibal Flores, Hugo Tito-Chura, and Honorio Apaza-Alanoca. 2021. Data Augmentation for Short-Term Time Series Prediction with Deep Learning. In *Intelligent Computing*, pages 492–506, Cham. Springer International Publishing.

Matthew Gordon and Timo Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, 3(1):20170007.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.

Markus Iseli and Abeer Alwan. 2004. An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 669–672.

James P. Kirby. 2018. PraatSauce: Praat-based tools for spectral analysis.

Michael David Larish. 1999. *The position of Moken and Moklen within the Austronesian language family*. Ph.D. dissertation, University of Hawai'i at Manoa.

Micheal David Larish. 1997. Moklen-Moken Phonology: Mainland or Insular Southeast Asian Typology? In *Proceedings of the Seventh International Conference on Austronesian Linguistics*, pages 125–149, Leiden. Rodopi.

Micheal David Larish. 2005. Moken and Moklen. In K.A. Adelaar and N. Himmelmann, editors, *The Austronesian Languages of Asia and Madagascar*. Routledge.

Leigh Lisker. 1986. "Voicing" in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees. *Language and Speech*, 29(1):3–11. _eprint: https://doi.org/10.1177/002383098602900102.

MathWorks. 2024. MATLAB version: 24.1.0 (R2024a).

Pittayawat Pittayaporn. 2024. On Becoming Mainland: Unravelling Malay Influence on Moklenic Languages. *SOJOURN: Journal of Social Issues in Southeast Asia*, 3(1):62–89.

Pittayawat Pittayaporn, Warunsiri Pornpottanamas, and Daniel Loss, editors. 2022. *Moklen-Thai-English dictionary: a pilot version*. Academic Work Dissemination Project, Faculty of Arts, Chulalongkorn University, Bangkok.

Warunsiri Pornpottanamas, Sireemas Maspong, and Pittayawat Pittayaporn. 2023. A Preliminary Investigation of the Phonetic Characteristics of Moklen Tones. In *The Second International Conference on Tone and Intonation*, pages 59–63. ISCA.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359. Publisher: Springer Science and Business Media LLC.

Pensiri Swastham. 1982. A description of Moklen: A Malayo-Polynesian language in Thailand. Master's thesis, Mahidol University.

Longfei Yang, Yanlu Xie, and Jinsong Zhang. 2018. Improving Mandarin Tone Recognition Using Convolutional Bidirectional Long Short-Term Memory with Attention. In *Proc. Interspeech 2018*, pages 352–356. ISSN: 2958-1796.