Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6458–6461
Marseille, 11–16 May 2020

Lexical Tone Recognition in Mizo using Acoustic-Prosodic Features

Parismita Gogoi[1,3], **Abhishek Dey**[4], **Wendy Lalhminghlui**[1], **Priyankoo Sarmah**[1], **S. R. M. Prasanna**[1,2]
[1]Indian Institute of Technology Guwahati, [2]Indian Institute of Technology Dharwad, [3]Dibrugarh University, [4]Kaliber.ai
Guwahati-781039, Dharwad-580011, Dibrugarh-786004, Guwahati-781039, India
{parismitagogoi, wendy, priyankoo, prasanna}@iitg.ac.in, {abhishekdey.gu}@gmail.com

Abstract

Mizo is an under-studied Tibeto-Burman tonal language of North-East India. Preliminary research findings have confirmed that four distinct tones of Mizo (High, Low, Rising, and Falling) appear in the language. In this work, an attempt is made to automatically recognize four phonological tones in Mizo distinctively using acoustic-prosodic parameters as features. Six features computed from Fundamental Frequency (F0) contours are considered, and two classifier models based on Support Vector Machine (SVM) & Deep Neural Network (DNN) are implemented for automatic tone recognition task respectively. The Mizo database consists of 31,950 iterations of syllables covering the four tones in Mizo, collected from 19 speakers using trisyllabic phrases. A four-way classification of tones is attempted with a balanced (equal number of iterations per tone category) dataset for each tone of Mizo. It is observed that the DNN-based classifier shows a performance in recognizing the four phonological tones in Mizo that is comparable to the SVM-based classifier.

Keywords: Tone recognition, Mizo, Tibeto-Burman, Pitch, Tone language

1. Introduction

The Mizo language belongs to the Kuki-Chin subgroup of Tibeto-Burman language family, and is spoken by 830, 846 (Census of India, 2011) people predominantly in Mizoram province in the Northeast of India. Mizo has four distinct lexical tones namely High (H), Low (L), Rising (R) and Falling (F) tone (Fanai, 1992; Chhangte, 1993; Sarmah and Wiltshire, 2010b). Two sets of tonal minimal pairs demonstrating the four-way contrast of Mizo tones obtained from (Lalhminghlui and Sarmah, 2018) are presented in Table 1. Previous acoustic studies on Mizo tones have shown that the four Mizo tones are distinct from each other in terms of $F0$ slope, averaged $F0$ and duration (Sarmah and Wiltshire, 2010b; Sarmah et al., 2015). In terms of duration, the canonical Mizo Falling tone is significantly shorter than the other three tones (Sarmah and Wiltshire, 2010b). In terms of average $F0$, the four tones in Mizo may not be distinct. However, in terms of $F0$ slope, all the four tones are significantly distinct from each other (Sarmah and Wiltshire, 2010b). As seen in Figure 1, the High and the Low tone in Mizo are comparatively static while the other tones have dynamic pitch contours. The Falling tone begins from nearly the same point of initiation of High tone and the Low tone has a falling contour. The Rising tone has an initial downward dip and then rises up from about 32% of the total duration. The presence of the effect of intrinsic *F0* on the four tones in Mizo is also reported where high vowels imposed higher *F0* and low vowels induced lower *F0* (Lalhminghlui et al., 2019).

Apart from the four distinct tones of Mizo, previous works on Mizo tones have reported the existence of a Rising Tone Sandhi (RTS) (Chhangte, 1993; Weidert, 1975; Sarmah et al., 2015; Lalhminghlui and Sarmah, 2018). RTS occurs in the environment where a Rising tone becomes a low tone when it is either followed by a High or Falling tone as clearly seen in Figure 1. However, at least phonetically, it is not identical to any of the other four phonological tones

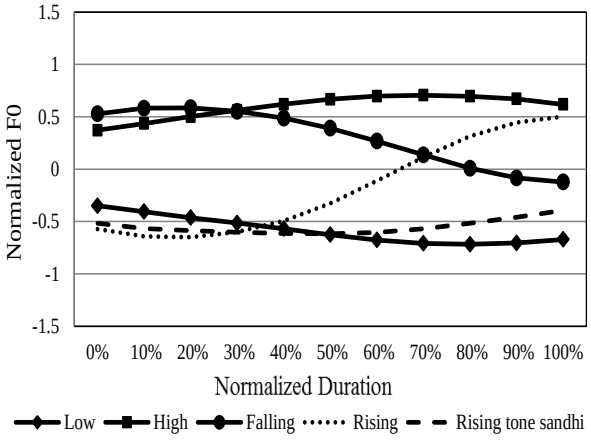

Figure 1: The normalized $F0$ contours of four Mizo tones and one sandhi tone averaged from 19 speakers.

Table 1: Mizo lexical words with the four contrastive tones taken from (Lalhminghlui and Sarmah, 2018)

Mizo	Meaning	Category	Tone
t^haŋ	famous	adjective	High
	gone	verb	Falling
	greasy	adjective	Low
	trap	noun	Rising
vai	chaff	noun	High
	wave	verb	Falling
	dazzle	verb	Low
	search	verb	Rising

in Mizo and the Mizo native speakers could differentiate between the canonical low tone and the derived low tone (Lalhminghlui and Sarmah, 2018).

In a preliminary perceptual study conducted on Mizo speakers, three tone parameters, namely, average F0, slope and duration are reported to be crucial in identifying the

synthesized Mizo tones (Govind et al., 2012). Later, based on this study, a method was proposed to identify Mizo tones automatically (Sarma et al., 2015). Using a limited database of Mizo tones, the authors reported that pitch height and F0 slope can automatically classify Mizo tones into the four phonological categories with considerable accuracy of 70.28%. However, this work had several shortcomings, firstly, the Mizo database used for the work was considerably small; secondly, the approach for identifying tones was threshold based and no statistical method was incorporated. Considering this, the present work proposes automatic Mizo tone recognition system using larger database.

2. Related Works

Cross-linguistically, there are numerous works on tone recognition in languages such as, Mandarin and Cantonese. Mandarin tone recognition was done using Hidden Markov Modeling based on a delta modulation of pitch sequence, where the experiments were conducted for well-trained speakers (Chen et al., 1987). Another tone recognition was conducted in Mandarin using a combination of vector quantization (VQ) and Hidden Markov Model techniques where the recognition accuracy was 98.33% for the speaker-dependent case and 96.53% for the speaker-independent case (Yang et al., 1988). The recognition of lexical tones in Mandarin speech was also done based on VQ and Hidden Markov Model by extracting the fundamental frequency where the average recognition rate was 97.9% for isolated monosyllabic words, 92.9% for disyllabic words and 91.0% for trisyllabic words for speaker-independent case (Liu et al., 1989). Mandarin monosyllables using multi-layer perceptron (MLP) was used for tone recognition by extracting 10 features of fundamental frequency and energy contours. The best recognition rate for speaker-untrained test was 93.8% (Chang et al., 1990). Multiclass Support Vector Machines (SVM) was used as a discriminative classifier for Mandarin tone recognition, achieving an accuracy of 76.5% (Levow, 2005). Artificial neural network was used in recognizing the tone patterns of Mandarin by extracting the fundamental frequency of each monosyllabic words using auto-correlation method. This achieved an accuracy of 90% correct for speech samples from both adults and children (Li et al., 2006).

In case of Cantonese, tone recognition was done for isolated Cantonese syllables using suprasegmental features extracted from the voiced portion of a monosyllabic utterance. Three layers feed forward neural network was used to classify these features achieving an accuracy of 89.0% for single-speaker and 87.6% for multi-speaker respectively (Lee et al., 1995). Hidden Markov Model was used for tone recognition in Cantonese for continuous spoken speech. A tone recognition accuracy of 66.4% was achieved in the speaker-independent case (Lee et al., 2002). The work in (Lee et al., 2002) was further explored using SVM achieving an accuracy of 71.5% in the speaker-independent case which compares favorably with the 66.4% result (Peng and Wang, 2005).

The remainder of this paper is organized as follows. Section 3. describes the speech corpus used in this work. Section 4. discusses the tone classification model used for the recognition of Mizo Tones. In Section 5. we present the experimental results. Finally, the paper is concluded in Section 6.

3. Mizo Speech Corpus

The Mizo speech corpus used in this work consists of 19 (10 male and 9 female) native speakers. The speech data was collected in a sound proof recording booth using a Tascam DR100 MKII linear PCM recorder connected to a Shure SM10A unidirectional head-worn, close-talk microphone. All the speakers were born and brought up in Mizoram and their average age was 22. The speakers were given printed sets of meaningful trisyllabic Mizo phrases, comprising of three monosyllabic words with all the possible combinations of the four Mizo lexical tones resulting in 64 distinct tone combinations. Each tone combination consists of five unique phrases which were recorded three times by each speaker which outcome is $17,280$ phrases resulted in $54,720$ total tokens (19 speakers x 64 tonal combinations x 5 trisyllabic phrases x 3 monosyllables x 3 repetitions). However, $22,770$ tokens are not considered as these are the low tones derived from RTS which is not considered in the present work. The Tone Bearing Unit (TBU) in Mizo is the syllable rime, made up of vowel nucleus or vowel with a sonorant coda. Accordingly, tone boundary was segmented and annotated by native Mizo speakers by listening and visually examining the pitch track of the TBU using Praat (Boersma, 2001).

4. MIZO TONE CLASSIFICATION MODEL

4.1. F0 estimation and normalization

The F0 values were automatically extracted at every 2% interval of the total duration and the values are collected in a spreadsheet. For speaker independent tone recognition, the speaker effect needs to be removed from the F0 contour. The z-score normalization is considered to be the best method for gender normalization (Rose, 1991). Z-score normalization is achieved with the equation: $x^* = \frac{x-\mu}{\sigma}$, where μ is the mean F0, and σ is the standard deviation of the F0 values considered for mean F0. After F0 normalization is done, speaker-dependent F0 values are non-existent while the shape of the original pitch contour and its relative height is maintained.

4.2. Front-End Acoustic Prosodic Feature Extraction

In the present study, a comprehensive set of acoustic-prosodic features is extracted for Mizo lexical tones. The lexical tones in Mizo are distinct in terms of F0 height and slope. While the Low and High tones in Mizo have level pitch contours, their F0 variability is less. On the other hand, the Rising and the Falling tones in Mizo have high F0 variability owing to their rising and falling F0 contours. Additionally, the High and the Low tones in Mizo differ in terms of their relative height. Similarly, the Rising and the Falling tones differ in terms of the direction of the slope. Considering these prosodic characteristics of the

Table 2: Balanced data set

Tone	Training Examples	Testing Examples
Falling	6,100	1100
High	6,100	1100
Rising	6,100	1100
Low	6,100	1100
Total	24,400	4,400

Mizo tones, the following parameters are derived from the F0 contour of the four Mizo tones.

- F0_slope
- F0_height
- F0_variance
- Initial_F0
- Final_F0
- F0_difference

The pitch level and the movement of the four Mizo tones are characterized by these sets of features, assisting in correct identification of the lexical tones. At various stages of the tone model, different features are investigated to capture the variation in tone contour patterns of the Mizo tones. Here, the pitch contour profile is time-aligned by a technique described in (Lee et al., May 1995). This is performed to characterize each tone recorded with different duration. Pitch profile of every monosyllable is segmented evenly into 16 equal portions and corresponding F0 values are noted. Given the pitch profile a syllable as $P(1)$, $P(2)$, $P(3)$, ..., $P(16)$, the mean pitch levels at the starting and the ending of that syllable are estimated as Initial_F0 = $\frac{P(3)+P(4)}{2}$ and Final_F0= $\frac{P(13)+P(14)}{2}$. First-two and last-two values are not considered in order to reduce consonantal effects and to enhance the stability of the mean estimates (Sarmah and Wiltshire, 2010a).

4.3. SVM-based Mizo tone recognition model

We propose an SVM-based Mizo tone recognition system for a larger tone database, which includes iterations of four tones in the language. Six possible $F0$ cues are investigated, that resulted in modest, but significant improvement in recognition accuracy. For the training and testing of the SVM-based tone recognition system, the database is divided into train and test sets. Out of total 19 Mizo speakers, 15 speakers' (7 female, 8 male) data is used for training. The remaining 4 speakers' (2 female, 2 male) data is used for the evaluation of the tone recognition system. This way of separation is assured to be speaker independent by excluding the same speaker data in the training and testing set at a time. For training the SVM model, a total of $24,400$ tokens are used by taking uniform numbers of samples from each tone category. For the testing part, $1,100$ tokens from each classes are considered. The distribution of training and testing tokens are shown in the Table 2. The optimum parameters (c, γ) of SVM are experimentally determined using the grid-search method. The best accuracy obtained in the considered range of c and γ is reported and taken as the final accuracy value.

Table 3: Confusion matrix of SVM-based Mizo tone recognition for one of the five folds. The average tone recognition accuracy over 4 tone classes turns out to be 73.39%.

	Falling	High	Rising	Low
Falling	**66.0**	15.45	2.81	15.72
High	14.00	**70.54**	8.45	7.00
Rising	3.27	7.54	**79.81**	9.36
Low	16.36	2.90	3.54	**77.18**

4.4. DNN-based Mizo tone recognition model

A Deep Neural Network (DNN) (Chen et al., 2014) with 3 hidden layers is trained for Mizo tone recognition, using Keras toolkit (Chollet and others, 2015). The distribution of training and testing samples are same as the SVM-based model as shown in Table 2. The input layer consists of acoustic-prosodic features as described in section 4.2. The output layer is a softmax of 4 dimension, one output for each of the four Mizo tones. The network is trained with random initialization of weights and biases, and optimized using Adam optimizer to minimize the categorical cross entropy loss between the target label and network output. The values for the parameters of Adam optimizer are β_1 = 0.9 and β_2 = 0.999 with an initial learning rate of 0.001. The network is trained for 100 epochs with a batch size of 64. ReLU is used as the activation function in the hidden layers.

Table 4: Confusion matrix of DNN-based Mizo tone recognition. The average tone recognition accuracy over 4 tone classes turns out to be 74.11%.

	Falling	High	Rising	Low
Falling	**64.72**	15.81	3.36	16.09
High	8.55	**74.54**	9.72	7.18
Rising	2.54	8.18	**82.09**	7.18
Low	14.63	4.09	4.18	**77.09**

5. Results and Discussion

In this section, the detailed results of the SVM- and DNN-based tone recognition are discussed. The SVM-based classifier with the six F0 features provides an accuracy of 73.39% for c = 100 and γ = 0.001. The confusion matrix for recognizing four classes of Mizo tones based on the SVM classifier is given in Table 3. From the confusion matrix, it can be seen that the recognition of the Falling tone is significantly low, and it is highly confused with the High and Low tones. The recognition rate of Rising and Low tones are high as compared to the Falling and High tones. The DNN-based tone recognition provides an accuracy of 74.11%, which provides an improvement of 0.72% over the SVM-based model. Table 4. provides the confusion matrix of DNN-based Mizo tone recognition system. The optimum number of the hidden layer is found as 3, where each layer contains 64 neurons. The confusion matrix obtained

from the DNN-based tone recognition model provides similar observations found from the SVM-based model. Since the number of tone tokens are very less for each class, SVM provides comparable results as of DNN.

6. Conclusions

The present work proposes a method for Mizo tone recognition using SVM- and DNN-based classifiers. The database reported in this work is prepared using trisyllabic utterances with tonal contrasts for 19 Mizo native speakers. The results of the current study have several practical and theoretical implications. The phonological tones in Mizo can be classified with considerable accuracy by using acoustic-prosodic features. The present work validates that $F0$ slope, $F0$ height, $F0$ variance, etc. can be the effective features for recognition of tones in a language. In this work, the experiment is performed with data obtained from 19 speakers having balanced training and testing samples of all four lexical tones where an accuracy of 73.39% is obtained with the SVM model. Furthermore, recognition of the four Mizo tones is attempted with acoustic-prosodic features and a DNN-based model which gives an accuracy of 74.11%. The results demonstrate that both the classification techniques provide comparable results in classifying the Mizo lexical tones.

The present work uses hand-crafted feature representation derived from the F0 contour for the classifier input. Learning of the tone specific speech signal and the tone contours using the deep learning techniques will be carried out in future. And further, those learned features may be used for the classification of tone category.

7. Acknowledgment

The speech corpus used in this work was developed for the project titled "Acoustic and Tonal Features based Analysis of Mizo", funded by the Department of Electronics & Information Technology (DeitY), Ministry of Communication & Information Technology (MC&IT), Government of India.

8. Bibliographical References

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Chang, P.-C., Sun, S.-W., and Chen, S.-H. (1990). Mandarin tone recognition by multi-layer perceptron. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520. IEEE.

Chen, X.-X., Cai, C.-N., Guo, P., and Sun, Y. (1987). A hidden markov model applied to chinese four-tone recognition. In *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 797–800. IEEE.

Chen, M., Yang, Z., and Liu, W. (2014). Deep neural networks for Mandarin tone recognition. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1154–1158, July.

Chhangte, L. (1993). *Mizo Syntax*. Ph.D. thesis, University of Oregon.

Chollet, F. et al. (2015). Keras link url: https://keras.io/.

Fanai, L. (1992). *Some Aspects of the Lexical Phonology of Mizo and English: An Autosegmental Approach*. Ph.D. thesis, CIEFL, Hyderabad, India.

Govind, D., Sarmah, P., and Prasanna, S. R. M. (2012). Role of Pitch Slope and Duration in Synthesized Mizo Tones. In *Speech Prosody*.

Lalhminghlui, W. and Sarmah, P. (2018). Production and perception of rising tone sandhi in mizo. *Proc. TAL*.

Lalhminghlui, W., Terhiija, V., and Sarmah, P. (2019). Vowel-tone interaction in two tibeto-burman languages. *Proc. Interspeech 2019*, pages 3970–3974.

Lee, T., Ching, P., Chan, L.-W., Cheng, Y., and Mak, B. (1995). Tone recognition of isolated cantonese syllables. *IEEE Transactions on speech and audio processing*, 3(3):204–209.

Lee, T., Lau, W., Wong, Y. W., and Ching, P. (2002). Using tone information in cantonese continuous speech recognition. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):83–102.

Lee, T., Ching, P. C., Chan, Y. H., and Mak, B. (May 1995). Tone recognition of isolated cantonese syllables. *IEEE Transactions on Speech and Audio Processing*, 3(3):204–209.

Levow, G.-A. (2005). Context in multi-lingual tone and pitch accent recognition. In *Ninth European Conference on Speech Communication and Technology*.

Li, X., Wenle, Z., Ning, Z., Chaoyang, L., Yongxin, L., Xiuwu, C., and Xiaoyan, Z. (2006). Mandarin chinese tone recognition with an artificial neural network. *Journal of Otology*, 1(1):30–34.

Liu, L.-C., Yang, W.-J., Wang, H.-C., and Chang, Y.-C. (1989). Tone recognition of polysyllabic words in mandarin speech. *Computer Speech & Language*, 3(3):253–264.

Peng, G. and Wang, W. S.-Y. (2005). Tone recognition of continuous cantonese speech based on support vector machines. *Speech Communication*, 45(1):49–62.

Rose, P. J. (1991). Considerations on the normalization of the fundamental frequency of linguistic tone. *Speech Communication*, 10(3):229–247.

Sarma, B. D., Sarmah, P., Lalhminghlui, W., and Prasanna, S. R. M. (2015). Detection of Mizo Tones. In *Interspeech*, pages 934–937.

Sarmah, P. and Wiltshire, C. (2010a). An Acoustic Study of Dimasa Tones. *North East Indian Linguistics*, 2:25–44.

Sarmah, P. and Wiltshire, C. R. (2010b). A Preliminary Acoustic Study of Mizo Vowels and Tones. *Journal of the Acoustical Society of India*, 37(3):121–129.

Sarmah, P., Dihingia, L., and Lalhminghlui, W. (2015). Contextual Variation of Tones in Mizo. In *Interspeech*, pages 983–986.

Weidert, A. (1975). *Componential Analysis of Lushai Phonology*, volume 2. John Benjamins Publishing.

Yang, W.-J., Lee, J.-C., Chang, Y.-C., and Wang, H.-C. (1988). Hidden markov model for mandarin lexical tone recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):988–992.