10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Automatic Tone Recognition of Ao Language

Parismita Gogoi[1,3], *Moakala Tzudir*[1], *Priyankoo Sarmah*[1], *S. R. M. Prasanna*[1,2]

[1]Indian Institute of Technology Guwahati, Guwahati-781039, India
[2]Indian Institute of Technology Dharwad, Dharwad-580011, India
[3]DUIET, Dibrugarh University, Dibrugarh-786004, India

{parismitagogoi, moakala, priyankoo}@iitg.ac.in, {prasanna}@iitdh.ac.in

Abstract

Ao is an under-resource Tibeto-Burman tonal language spoken in the North-Eastern state of India, Nagaland. Preliminary research findings have confirmed that Ao has three lexical tones, namely, High, Mid and Low. There are three distinct dialects of Ao, namely, Chungli, Mongsen, and Changki, differing in tone assignment in lexical words. In this work, tone distributions in trisyllabic words are considered with 4320 iterations consisting of 4176 high, 5473 mid and 3311 low tones, collected from 36 speakers, for the three dialects of Ao. An attempt is made to automatically recognize the phonological tones in Ao using SVM with zero-frequency filtering (ZFF) derived F_0 profile as the preliminary feature.

Index Terms: Tone recognition, Ao, ZFF.

1. Introduction

Ao is a Tibeto-Burman [1] language spoken in the Northern part of Nagaland, India. According to the Census of India 2001, Ao is the highest populated community among the Nagas with a total of 227,000. There are three known distinct dialects of Ao, namely, Chungli, Mongsen and Changki [2]. The language is reported to have three lexical tones, High (H), Mid (M) and Low (L) [3]. It is also reported that tone assignment in the Changki dialect of Ao is different from Mongsen and Chungli dialects which is noticed even in the same words as listed in Table 1 [4]. From Table 1, [4] stated that a High tone of Changki dialect corresponds to a Low tone of Mongsen and Chungli dialects and vice-versa as most of the words belong in that category. Ao being an under-described language, there are a few

Table 1: *Disyllabic words with different tone sequences for the three Ao dialects. (From [4])*

Changki	Mongsen	Chungli	Gloss
lata - HH	lata - LL	ita - LH	'moon'
alik - LH	alik - HL	azek - HL	'necklace'
akung - HL	akung - MM	akung - HH	'shrimp'

works on Chungli dialect which is the standard dialect of Ao [5, 6, 7, 8] and a description of tones in the Chungli dialect [9]. Similarly, there are a few works available for the Mongsen dialect [2, 10, 11, 12] and also on acoustic and perceptual features of tone in Mongsen dialect [3]. However, Changki dialect is the least documented. [4] describes the tonal correspondence based on perception basis. An automatic discrimination of Ao dialects is performed by [13] and finally, dialect identification in two dialects of Ao [14] is reported.

To derive the canonical F_0 patterns for the tones in Changki, Mongsen and Chungli dialects, a set of 40 trisyllabic words were considered where each word was produced by 12 speakers for each dialect. The normalized F_0 values were extracted at every 10% of the total duration of the tone. The canonical z-score normalized F_0 contour for the three dialects of Ao are plotted as shown in Figure 1, where all the three tones are static level tones and can be categorized based on the height of the tones [13]. Statistical analysis was done using the *lme4* package on R, where three Linear Mixed Effects (lme) models were built, each with average F_0, initial_F_0 and final_F_0 as dependent variables, tone types as fixed effect and speaker, gender, dialect, syllable number (first, second or third), context (sentence, isolation or carrier phrase) as random effects [15, 16]. The three models were subjected to an analysis of deviance using Type II Wald chisquare tests using the *car* package on R[17]. The results of the test are summarized in Table 2. As it is seen in Table 2, there is a significant effect of tone types on average F_0, initial_F_0 and final_F_0, confirming that F_0 does change significantly with the change of tones in all three varieties of Ao.

Table 2: *Results of analysis of deviance for three LME models (N=12960)*

Variable	df	χ^2	p-value
Average F_0	2	199.6	< 0.001
Initial_F_0	2	193.8	< 0.001
Final_F_0	2	119.6	< 0.001

In our paper, an attempt is made to automatically recognize the phonological tones in Ao using tonal features computed from zero-frequency filtering (ZFF) derived F_0 profile. In Ao, three different tone recognition systems based on SVM are developed for the three dialects individually.

The remainder of this paper is organized as follows. In section 2, related works on tone recognition is presented. Section 3 describes the speech corpus. Section 4 describes the feature extraction techniques using ZFF. Section 5.1 discusses the SVM based model. In Section 6, we present the experimental results and finally conclude the paper in section 7.

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 HMM based on a delta modulation of pitch sequence [18]. Tone recognition was conducted in Mandarin using a combination of vector quantization (VQ) and HMM techniques with recognition accuracy 98.33% for speaker-dependent and 96.53% for speaker-independent case [19]. The recognition of lexical tones in Mandarin speech was done based on VQ and Hidden Markov Model by extracting the fundamental frequency. The average recognition rate was

10.21437/SpeechProsody.2020-205

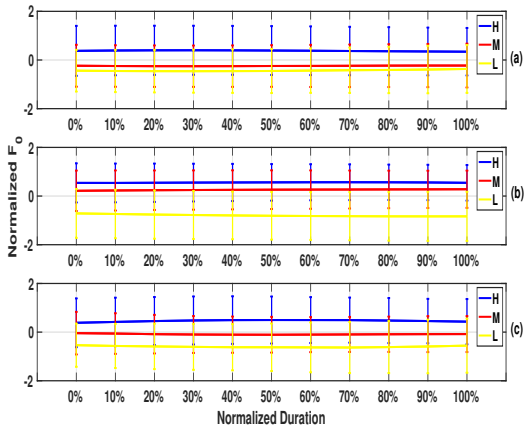

Figure 1: *The normalized F_0 contours of three Ao tones averaged from* 12 *speakers for (a) Changki, (b) Mongsen and (c) Chungli dialects*

97.9% for isolated monosyllabic words, 92.9% for disyllabic words and 91.0% for trisyllabic words, for speaker-independent case [20]. 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% [21]. Multiclass SVM was used for Mandarin tone recognition, achieving an accuracy of 76.5% [22]. 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 [23].

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 layer 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 [24]. 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 [25]. The work in [25] was further explored using SVM achieving an accuracy of 71.5% in the speaker-independent case which compares favorably with the 66.4% result [26].

3. Preparation of Ao Speech Corpus

For this work, Changki dialect spoken in Changki village in the Western Changkikong range, Mongsen dialect spoken in Khensa village in the Southern Onpangkong range and Chungli dialect spoken in Mopungchuket village in the Central Asetkong range were taken into consideration. For each of the three dialects, data from 12 native speakers, 6 males and 6 females, were recorded reading a set of trisyllabic words. A set of 40 target words was considered for the three dialects, resulting in a total of 1440 trisyllabic utterances in each dialect, as in the examples shown in Table 4. The speakers were asked to utter each target word in a sentence, in isolation and in a carrier phrase. For example, if the target word is "water", the speakers uttered as (a) there is water in the bottle, (b) water and (c) I said water. The target words being trisyllabic, a total of 12960 tokens are considered for the three dialects as shown in Table 3. Data was recorded with TASCAM DR-100 MKII 2-channel portable digital recorder with 44.1 KHz sampling rate connected to a head-mounted Shure SM10A microphone for high-quality recordings. After the recording, data was annotated and the tone boundaries were marked manually, using Praat 6.0.35 [27]. However, for this work, the speech signal is re-sampled to 8KHz for pitch estimation using ZFF.

Table 3: *Ao dataset distribution*

Dialect/ tones	**Low**	**Mid**	**High**
Changki (D1)	756	1584	1980
Mongsen (D2)	1295	2305	720
Chungli (D3)	1260	1584	1476
Total samples	3311	5473	4176

Table 4: *Three examples of trisyllabic words with different tone sequences for the three dialects.*

Trisyllabic Words	**Changki Tones (D1)**	**Mongsen Tones (D2)**	**Chungli Tones (D3)**	**Gloss**
Temesen	HMM	HHL	MHL	'liver'
Wamaba	LHM	LHL	HHL	'slice into pieces'
Watangba	HLH	MLL	MHL	'to saw/cut into two pieces'

4. Front-End Acoustic Prosodic Feature Extraction

The classification of tone is performed in a systematic procedure for both the languages. Speech data is first pre-processed to sample at 8 KHz, and F_0 values of each monosyllabic tone utterances are extracted using Zero Frequency Filter (ZFF) technique [28]. Here, the pitch is estimated using the zero frequency resonator. In this case, the speech signal is passed through a cascade of two ideal resonators followed by a local mean subtraction process. The resultant signal is sinusoidal in nature which is known as ZFF signal (ZFFS), whose positive-to-negative zero crossings are used to estimate the glottal closing instants (GCIs). The inverse of the interval between two successive GCIs are taken as instantaneous pitch of the signal.

For speaker independent tone recognition, the gender effect needs to be removed from the F_0 contour. The z-score normalization is considered to be the best method for gender normalization [29]. Z-score normalization is achieved with the equation,

$$x^* = \frac{x - \mu}{\sigma} \tag{1}$$

where, μ is the mean F_0, and σ is the standard deviation of F_0 values considered for mean F_0. After z-score normalization, speaker-dependent F_0 values are non-existent while the shape of the original pitch contour and its relative height is maintained. In the present study, a comprehensive set of acoustic-prosodic features are extracted for Ao phonological tones respectively. It has been discussed in section 1 that the characteristics of existing lexical tones are unique for the language. Hence, several F_0 features are investigated for Ao. For Ao, four F_0 features are extracted from the three level tones, namely,

- F_0_height

- Initial_F_0
- Final_F_0
- Average F_0

To calculate the F_0height, the pitch contour of every syllable is fitted with a line. Fitting is done using linear regression which estimates the coefficients (a, b) of a polynomial of degree one that fits the values best in a least-squares sense. The estimated line is

$$y = ax + b \tag{2}$$

where, x is a vector $= (1, 2..., N)$ and N is the length of the segment, b is the F_0height parameter [30]. Here, the pitch contour profile is time-aligned by a technique described in [31]. This is performed to characterize each tone recorded with different duration. ZFF derived pitch profile of every monosyllable is segmented evenly into 16 equal portions and corresponding F_0 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,

$$Initial\_F_0 = \frac{P(3) + P(4)}{2} \tag{3}$$

$$Final\_F_0 = \frac{P(13) + P(14)}{2} \tag{4}$$

Note, first-two and last-two values are not considered in order to reduce consonantal effects and to enhance the stability of the mean estimates [32].

5. Experimental Details

5.1. Ao Tone Recognition Model

In the literature till now, no work has been reported on tone recognition of Ao language and not many acoustic analysis on the varieties of Ao that describe the assignment of tones. However, the canonical F_0 patterns for Changki and Mongsen tones have been described in [13]. Using the tonal features, one with praat derived values and the other extracted using ZFF, dialect discrimination was performed on the two dialects of Ao, namely, Changki and Mongsen which showed that ZFF extracted values gave a more robust accuracy [13]. It was further explored using tonal and spectral features in the two varieties of Ao for dialect identification. Based only on spectral features yielded satisfactory results 85.1%, while, with the addition of tone information in automatic dialect identification improved the results to 86.2% [14]. In present work, as a preliminary study, SVM based tone recognition system for the three dialects of Ao, namely, Changki, Mongsen and Chungli is proposed using a limited database of Ao tones (Table 5).

Table 5: *Ao dataset for SVM training- testing experiment*

Ao Dialect	No. of Speakers	Training	Testing
Changki (D1)	9-Training 3-Testing	3244	1076
Mongsen (D2)	9-Training 3-Testing	3240	1078
Chungli (D3)	9-Training 3-Testing	3197	1076

Table 3 shows the token for each of the three Ao dialects recorded from 12 native speakers, 6 males and 6 females each in a single dialect. The classification task is determining which of the three tones each Ao syllable has within single dialect. The tone classes are not balanced and some tokens are removed as the pitch was not estimated; as seen in Table 5. In the current work, SVM classifier is used for the classification of Low, Mid and High tones using an RBF kernel. The kernel function maps the data features in higher dimensional space, which makes it linearly separable. Separate SVM models are built for each dialect for optimum parameters (c, γ). For each SVM, the optimum values of the parameters c and γ are experimentally determined using the grid-search method. Each set in three dialects contains randomly selected 9 speaker data used for training and remaining 3 speaker data are used for testing. Each of the set is assured to be speaker independent by excluding the same speaker data in the training and testing set. The three-class SVM model is trained and tested for each of the Ao dialect. All combinations of the RBF kernel parameters c and γ are considered in the range of $c = [10^{-1}, 10^0, .., 10^{+2}]$ and $\gamma = [10^{-3}, 10^{-2}, .., 10^0]$ during classification. The best accuracy obtained in the considered range of c and γ is reported as a classification result for the specific set of each dialect.

6. Results and Discussion

The present work describes a method for Ao tone recognition using SVM based classifier consisting of trisyllabic utterances recorded with 36 Ao. From Ao dataset, for each dialect class, 9 speakers data is used for training and 3 speakers for testing. Therefore, no overlap in training and testing speech samples. The recognition accuracy for Changki, Mongsen and Chungli dialect is found to be 53.20%, 62.26% and 52.69%, respectively. From the recognition accuracies, it can be observed that Mongsen shows more discrimination in tone, compared to other two dialects. The reason could be smaller number of samples for the low-tone category in Changki, and high-tone in Mongsen. Also there are large number of Mid tone samples in Mongsen dialect unlike the other two dialects. Accuracies can be later improved by taking equal number of tones in all the three dialects.

The results of the current study have several practical and theoretical implications. The results should not be compared with other dialect identification works on Chinese or in case English varieties as phonetic salience is much less in our case. Ao being a preliminary study, it is evident from the results that the level tones in Ao are classified with considerable accuracy by using acoustic-prosodic features. Tone related works have reported that there is always complexity in recognizing level tones. When contour tones are present, $\delta F0$ may provide additional discriminatory feature, which is not possible in Ao dialects.

7. Summary and Future works

This work presents an SVM-based dialect dependent tone detection system for Ao. The motivation for this work is to explore the existing acoustic features in Ao tone recognition system, and to study its effectiveness for three different dialects. The SVM-based system is designed to discriminate three lexical tones of Ao, namely, High, Mid and Low. We have considered tone distributions in trisyllabic words, which are collected from 36 speakers. F_0 contour is derived using ZFF-based algorithm, and four acoustic features are computed from the F_0 contour. For each dialect, separate SVM-based tone recognition model is developed. Results show that the recognition accuracies are 53.20%, 62.26% and 52.69% for Changki, Mongsen

and Chungli dialects, respectively which are above chance level. Further works need to be done for more robust recognition accuracy. Also, new acoustic features which are discriminable amongst the dialects need to be derived in the future works.

In future, we plan to exploit the spectral features further for more robust tone recognition in Ao. Ao corpus will be strengthened by incorporating more speakers data and increased number of tonal words. This study shows the tone recognition in dialect specific way. Therefore, a future work is planned to develop a dialect independent tone recognition system for Ao. Also, there are tonal coarticulation in the trisyllabic word, such that the tone from previous or the following syllable can affect the realisation of the target tone. Such studies are not present in Ao language. Future work will be framed in this direction to study whether this affects the classification accuracy of the model too.

8. Acknowledgment

The authors would like to thank the speakers from Changki dialect spoken in Changki village in the Western Changkikong range, Mongsen dialect spoken in Khensa village in the Southern Onpangkong range and Chungli dialect spoken in Mopungchuket village in the Central Asetkong range who voluntarily took part in this work.

9. References

[1] G. Grierson, "Linguistic survey of India vol iii part iii," 1904.

[2] A. R. Coupe *et al.*, "A phonetic and phonological description of Ao: A Tibeto-Burman language of Nagaland, north-east india," 2003.

[3] A. R. Coupe, "The Acoustic and Perceptual Features of Tone in the Tibeto-Burman Language Ao Naga." in *ICSLP*, 1998.

[4] T. Temsunungsang, "Tonal correspondences in Ao languages of Nagaland," in 22 *Himalayan Languages symposium, IIT Guwahati*, 8-10 june 2016.

[5] Clark and M. E. Winter, *The Ao Naga Grammar*. Delhi: Gian Publications, 1893.

[6] K. G. Gowda, *Ao-Naga phonetic reader*. Central Institute of Indian Languages, 1972, vol. 7.

[7] K. Gurubasave-Gowda, "Ao grammar," *Mysore: Central Institute of Indian Languages*, 1975.

[8] E. Clark, *Ao-Naga dictionary*. Updated in 2013, 1911.

[9] D. Bruhn, "The tonal classification of Chungli Ao verbs," 2009.

[10] A. R. Coupe, "A phonetic and phonological description of Ao: A language of Nagaland, North-east India," Ph.D. dissertation, Australian National University, 1999.

[11] A. Coupe, *A Grammar of Mongsen Ao*. Walter de Gruyter, 2007, vol. 39.

[12] T. Temsunungsang, "The structure of Mongsen: Phonology and morphology," 2003.

[13] M. Tzudir, P. Sarmah, and S. M. Prasanna, "Tonal feature based dialect discrimination in two dialects in Ao," in *Region 10 Conference, TENCON 2017-2017 IEEE*. IEEE, 2017, pp. 1795–1799.

[14] M. Tzudir, P. Sarmah, and S. R. M. Prasanna, "Dialect identification using tonal and spectral features in two dialects of Ao," in *Proc. SLTU*, 2018.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[17] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[18] X.-X. Chen, C.-N. Cai, P. Guo, and Y. Sun, "A hidden markov model applied to chinese four-tone recognition," in *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12. IEEE, 1987, pp. 797–800.

[19] W.-J. Yang, J.-C. Lee, Y.-C. Chang, and H.-C. Wang, "Hidden markov model for mandarin lexical tone recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 988–992, 1988.

[20] L.-C. Liu, W.-J. Yang, H.-C. Wang, and Y.-C. Chang, "Tone recognition of polysyllabic words in mandarin speech," *Computer Speech & Language*, vol. 3, no. 3, pp. 253–264, 1989.

[21] P.-C. Chang, S.-W. Sun, and S.-H. Chen, "Mandarin tone recognition by multi-layer perceptron," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990, pp. 517–520.

[22] G.-A. Levow, "Context in multi-lingual tone and pitch accent recognition," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[23] X. Li, Z. Wenle, Z. Ning, L. Chaoyang, L. Yongxin, C. Xiuwu, and Z. Xiaoyan, "Mandarin chinese tone recognition with an artificial neural network," *Journal of Otology*, vol. 1, no. 1, pp. 30–34, 2006.

[24] T. Lee, P. Ching, L.-W. Chan, Y. Cheng, and B. Mak, "Tone recognition of isolated cantonese syllables," *IEEE Transactions on speech and audio processing*, vol. 3, no. 3, pp. 204–209, 1995.

[25] T. Lee, W. Lau, Y. W. Wong, and P. Ching, "Using tone information in cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1, no. 1, pp. 83–102, 2002.

[26] G. Peng and W. S.-Y. Wang, "Tone recognition of continuous cantonese speech based on support vector machines," *Speech Communication*, vol. 45, no. 1, pp. 49–62, 2005.

[27] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[28] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, May 2009.

[29] P. J. Rose, "Considerations on the normalization of the fundamental frequency of linguistic tone," *Speech Communication*, vol. 10, no. 3, pp. 229–247, 1991.

[30] B. D. Sarma, P. Sarmah, W. Lalhminghlui, and S. M. Prasanna, "Detection of mizo tones," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[31] T. Lee, P. C. Ching, Y. H. Chan, and B. Mak, "Tone recognition of isolated cantonese syllables," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 3, pp. 204–209, May 1995.

[32] P. Sarmah and C. Wiltshire, "An Acoustic Study of Dimasa Tones," *North East Indian Linguistics*, vol. 2, pp. 25–44, 2010.