



Prosodic Information in Dialect Identification of a Tonal Language: The case of Ao

Moakala Tzudir¹, Priyankoo Sarmah¹, S. R. Mahadeva Prasanna²

¹Indian Institute of Technology Guwahati, Guwahati - 781039, India

²Indian Institute of Technology Dharwad, Dharwad - 580011, India

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

Abstract

Dialect identification has been explored profusely in major languages, such as Arabic, Chinese and Spanish. This paper presents an automatic dialect identification system in the Ao language using prosodic features. Ao is a low-resource Tibeto-Burman tonal language spoken in Nagaland in the North-Eastern part of India. It consists of three distinct dialects: Chungli, Mongsen and Changki. Prosodic characteristics are believed to have an essential role in tonal languages. In this direction, the current work focuses to investigate the prosodic characteristics to build a discriminative system in identifying the three Ao dialects. The statistical and Low-Level Descriptors (LLD) of prosodic features are used in this work. The prosodic features such as F_0 , loudness, shimmer, jitter, voiced and unvoiced segment length, etc., are utilized in this study. The experiments are conducted using SVM and attention-based Bi-GRU classifiers in trisyllabic words and passage-level datasets, respectively. The combination of prosodic features outperforms the MFCC (baseline) feature. The Voice Quality and Temporal (VQT) feature set is the best performing prosodic feature. The statistical analysis also shows that the VQT features are statistically significant. The performances of SVM and attention-based Bi-GRU classifiers indicate the significance of prosodic information in classifying the three Ao dialects.

Index Terms: Ao, tonal language, low-resource, prosody, dialect identification, SVM, Bi-GRU, attention

1. Introduction

Ao is an under-resourced language spoken in Nagaland, a North-Eastern state of India [1]. It belongs to the Tibeto-Burman language family consisting of three distinct dialects, namely, Chungli, Mongsen and Changki [2, 3]. The Chungli dialect is known to be the standard dialect of the language [2, 3]. As per the Census of India 2011, the resident population of the language in Nagaland is 227,000 [4]. Ao is a tonal language and has three lexical level tones viz., High (H), Mid (M) and Low (L) tones [5, 3]. However, the tone distribution are different across the three dialects of the language even for words with the same meaning. For instance, Table 1 shows some examples for words differing in tones in the three dialects of Ao. As the language is spoken by one of the most populated tribes in Nagaland, there are various works that investigate different aspects of the language. Several researchers have studied in the linguistic direction of Ao language [5, 2, 3, 6, 7]. However, speech analysis and modeling have not been explored much in Ao. Therefore, this work is attempted to bridge the gap with other widely explored languages of the world, such as Arabic, Chinese and Spanish in Dialect Identification (DID) tasks.

DID is one of the vital topics in the speech research area. The objective of DID task is to distinguish one dialect from the

Table 1: Tone assignment in Ao dialects [3].

Chungli	Mongsen	Changki	Gloss
azək - HL	alík - HL	alík - LH	'necklace'
akuɲ - HH	akuɲ - MM	akuɲ - HL	'shrimp'

other within the same language family [8]. A dialect of a language varies from the standard variety through the use of specific vocabularies and pronunciation [9]. Among the dialects of a language, it is reported to exhibit grammatical, phonological and prosodic differences [9]. The methodology for Language Identification (LID) and DID follows the same principle. However, it is reported that the boundaries of two languages are easily recognizable in comparison to two dialects [10]. It is also described that the overlaps in phonetic and vocabulary systems are more across two dialects than across two languages [10]. In this regard, the DID task is considered more challenging than the LID task [11]. As a specific use case, voice-controlled electronic devices have provided more comprehensive range of applications by using the outcomes of DID systems.

In the literature, several works have attempted DID tasks in major languages of the world, such as Chinese, Arabic and Spanish. There are a number of works in DID task using spectral features such as Mel Frequency Cepstral Coefficient (MFCC) and Shifted Delta Cepstral (SDC) features in languages such as Arabic, Spanish, Hindi and Kannada [10, 12, 13, 14, 15, 16]. Several works have reported improved results in classifying the dialects when prosodic features such as F_0 , energy and duration are combined with spectral features [13, 14, 15]. A number of studies are explored in tonal languages such as Chinese and Vietnamese. These works utilize the spectral features viz., MFCC and SDC with prosodic features such as F_0 and energy [17, 18, 19, 20]. Various works in DID tasks used the filter-bank features outperforming the traditional baseline approaches [11, 21, 22]. Also, a study on German DID system incorporates character, word n-grams and word k-skip bigrams, attaining an accuracy of 62.03% [23]. Some works in low-resource languages like North Sámi and Meeteilon are attempted using prosodic features such as F_0 , energy, intensity and duration, yielding 60% and 61.57% accuracy rate, respectively [24, 25].

An automatic DID system in Ao has been attempted in our previous works using spectral features like MFCC and SDC along with tonal information [26, 27, 28]. Excitation source feature is also explored to classify the three Ao dialects in our previous study [29]. As per the literature, majority of the works have explored spectral features such as MFCC and SDC. A number of works have studied the prosodic feature viz., F_0 , energy, intensity and duration. However, prosodic features have not been studied extensively in the literature. As Ao is a tonal language, it is believed that prosodic characteristics will play

a vital role in discriminating the dialects. The variations in the prosodic information may capture the potential differences across the three dialects of Ao. Hence, to see the effectiveness of prosodic information in Ao, the openSMILE toolkit is used to extract multiple prosodic features [30]. In terms of DID tasks, this toolkit has been explored only in [31] to the best of our knowledge. Also, most of the works studied previously are based on high-resource languages. In case of the Arabic language, the Arabic dialects such as Modern Standard Arabic (MSA), Iraqi Arabic and Levantine Arabic are standard dialects of different countries. These dialects are used in broadcast news with available written scripts. However, Ao is a low-resource language where the Chungli dialect is known to be the standard dialect of the language. Hence, the Chungli dialect is used in all formal occasions and gatherings. In particular, all written text is available only in the standard dialect. As such, speech analysis and modeling becomes more challenging in Ao. Accordingly, an automatic Ao DID system is attempted using prosodic information in this work. The contributions of this work are listed below.

- DID in three Ao dialects is attempted, which is a low-resource language spoken by one of the major tribes in Nagaland.
- The present work proposes the use of prosodic information for DID in Ao. First, the statistical prosodic feature is utilized with Support Vector Machine (SVM) based classifier. The classification results establish the importance of prosodic characteristics for the current task.
- Statistical analysis is also conducted to further analyze the efficacy of statistical prosodic features. The result indicates that the temporal context based prosodic features are more significant in the current DID task. This motivates to explore temporal information of different prosodic features for the task.
- The temporal information of prosodic features is learned using an attention-based Bi-directional Gated Recurrent Unit (Bi-GRU) model. The Low-Level Descriptors (LLD) of prosodic information is used with an attention-based Bi-GRU classifier to classify the three Ao dialects. The attention mechanism is performed along the feature dimension. The attention provides higher weights to the features that carry prominent discriminative information across the dialects.

The remaining paper is organized as follows. Section 2 gives a brief description of the speech corpus recorded for this work. The proposed approach is described in section 3. Experiments and results are discussed in 4. Finally, the work is concluded in section 5 with future directions.

2. Speech corpus

The speech data for the three Ao dialects were recorded as spoken in three villages viz., Mopunchuket Village (MV), Khensa Village (KV) and Changki Village (CV). The Chungli dialect was collected from speakers of MV, the Mongsen dialect from speakers of KV and the Changki dialect from speakers of CV. The speech data is recorded at two levels: trisyllabic words and passage-level.

Trisyllabic Words: There were 12 native speakers for the trisyllabic words consisting of 6 females and 6 males across each dialect. The speakers read 40 trisyllabic words common across the three dialects with lexical differences for some

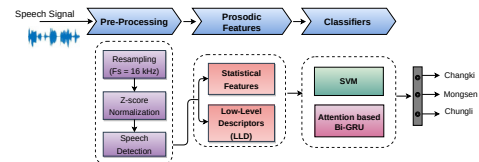


Figure 1: Overall framework of Ao DID system

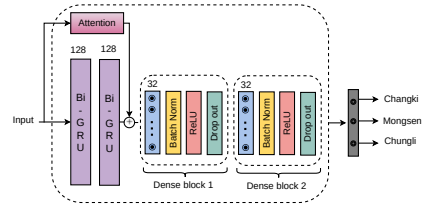


Figure 2: Architecture of attention-based Bi-GRU model

Chungli words. These 40 words differ from dialect to dialect in their tonal specification. The speakers were asked to read the target words in three criteria: meaningful sentences, isolation and phrase. This resulted in a total of 4320 utterances across the three Ao dialects. In addition, the recording was repeated for the same speakers, speaking the same utterances after 2 months to add session variability. Hence, the speech data resulted in 8640 utterances across the 2 sessions.

Passage-Level: There were 8 native speakers for the passage level consisting of 4 females and 4 males from each dialect. The speakers read a short bible passage, “The parable of the prodigal son”. As Chungli is the standard dialect, the text of the bible is available only in the standard dialect. In regard to this, the passage was translated for the speakers of Changki and Mongsen dialects. The passage was recorded in 4 sessions for all the speakers. This resulted in approximately 6 hours of recordings across the three dialects. For all the recordings, a TASCAM DR-100 MKII with a 2-channel portable digital recorder was used. For high-quality recordings, a shure SM10A head-mounted microphone was connected to the recorder. All the recordings took place in a real-world environment scenario. Apart from that, all the speakers could speak English and Nagamese (*lingua franca*) besides their native Ao dialects.

3. Proposed work

The proposed framework of the Ao DID system is illustrated in Figure 1. The input speech signal is initially pre-processed by resampling to 16 kHz, applying z-score normalization and detecting the speech region by removing the silence regions. Next, the pre-processed speech is passed through the prosodic feature extraction block. The extracted features are then fed to the classifiers for the three-class classification task. The description of prosodic features and classifiers are presented next.

3.1. Prosodic Features

For the DID in Ao, the variation in suprasegmental characteristics of the speech signal are believed to carry potential differences across the three dialects. As Ao is a tonal language, the suprasegmental features capturing variation in different aspects of tone are expected to be helpful for the task. Therefore, different statistical features of F_0 semitone along with jitter and shimmer are explored. Similarly, the present work also focuses on exploring other prosodic features, such as loudness, voice quality and temporal features to capture the suprasegmental variation across the dialects. In this work, the prosodic features are extracted using the openSMILE toolkit [30, 32, 33]. The

Table 2: Classification performance of Ao dialects using statistical prosodic features in trisyllabic words reported in ($\mu \pm \sigma$).

Statistical Features	Accuracy	F1-score			
		Changki	Mongsen	Chungli	Average
MFCC (Baseline)	43.21±5.07	30.67±9.05	49.36±2.88	43.24±6.32	41.09±6.07
F_0 ST	36.91±4.83	37.66±7.34	32.93±2.07	39.07±5.08	36.86±5.64
Loudness	38.84±3.64	39.87±5.90	30.61±6.44	44.98±1.61	36.96±4.16
VQT	46.08±1.65	41.30±1.93	43.88±1.86	52.41±1.84	44.16±1.35
F_0 ST+Loudness	45.80±5.02	42.91±7.60	42.64±2.82	51.29±5.28	45.61±4.97
VQT+ F_0 ST	44.80±4.88	42.00±6.49	38.22±6.30	53.48±2.41	44.57±5.01
VQT+Loudness	43.69±4.74	42.06±6.97	41.15±1.98	47.44±5.93	43.55±4.63
VQT+ F_0 ST+Loudness	46.30±2.03	42.02±2.11	43.68±2.02	52.66±2.40	46.12±2.07
VQT+Loudness+ F_0 ST+MFCC	49.06±5.28	42.83±7.31	47.12±4.16	56.53±4.86	48.82±5.40

prosodic features are extracted on two levels: Functionals and LLD. In the case of functionals, 30 statistical prosodic features in terms of mean (μ) and its standard deviation (σ) are extracted and are categorized into 3 groups:

1. F_0 SemiTone (F_0 ST) features : 10 statistical features, namely, percentile 20, 50, 80 of F_0 ST, μ and σ of F_0 ST, F_0 ST rising slope and F_0 ST falling slope.
2. Loudness features: 10 statistical features viz., percentile 20, 50, 80 of loudness, μ and σ of loudness, rising slope of loudness and falling slope of loudness.
3. Voice Quality and Temporal (VQT) features: Voice quality comprises of 4 statistical features viz., μ and σ of jitter, shimmer. While, the temporal features include 6 statistical features, namely, loudness peaks per sec, voiced segments per sec, μ and σ of voiced segment length and unvoiced segment.

In the case of LLD, 10 prosodic features are extracted, namely, F_0 final, sum of auditory spectrum (loudness), sum of RASTA-filtered auditory spectrum, Root Mean Square (RMS) energy, Zero-Crossing Rate (ZCR), probability of voicing, log Harmonic-to-Noise Ratio (HNR), jitter (local & Δ) and shimmer (local) [33].

3.2. Classifiers

The present work performs the classification task using two different classifiers, namely, SVM and attention-based Bi-GRU.

Support Vector Machine (SVM): SVM classifier (with RBF kernel) is trained using statistical prosodic features extracted from trisyllabic words. The optimum values of the kernel parameters, c and γ are obtained using the grid-search mechanism. The parameters c and γ are considered in the range of $c = [10^{-1}, 10^0, \dots 10^{+2}]$ and $\gamma = [10^{-3}, 10^{-2}, \dots 10^0]$ for the grid search.

Attention-based BiGRU: The architecture of the attention-based BiGRU model is illustrated in Figure 2. This architecture is motivated from our previous work [29]. Temporal information plays a vital role in capturing the prosodic details of a speech signal. Therefore, the proposed work is motivated to use Bi-GRU based architecture to learn the temporal context of different dialects. The proposed model also uses an attention mechanism [34] along the feature direction. The attention mechanism gives higher weights to those features that provide essential information for the classification task. The model consists of two Bi-GRU layers with 128 units each. The output of the attention module and Bi-GRU is concatenated. The concatenated output is fed to two dense layers with 32 neurons each. All the dense layers have *ReLU* activation and a dropout rate of 0.4. Finally, the output layer (size = 3) is activated with *Softmax* function. The model is trained for 50 epochs with a mini-

batch size of 33. An early stopping criteria is used to avoid overfitting of the model. The model is trained with categorical cross-entropy loss and an initial learning rate of 0.0001.

4. Experiments and Results

This section describes the experimental setups and results obtained for the present work. The LLD features are calculated with a window size of 20 ms and a hop size of 10 ms.

4.1. Baseline Method

MFCC feature is broadly used in DID task to capture the vocal tract information [10, 12, 16, 35]. Therefore, MFCC is considered as the baseline method. The 13-dimensional MFCC features with their Δ and $\Delta\Delta$ are extracted using openSMILE toolkit [30]. The statistical features (μ and σ) of 39-dimensional MFCC are also extracted from the trisyllabic words. While, 39-dimensional MFCC features extracted from the passage-level data are considered as MFCC-LLD.

4.2. SVM Based Classification

Initially, DID task in Ao is conducted with statistical prosodic features extracted from the trisyllabic words. The trisyllabic data is divided into three non-overlapping folds consisting of 2 females and 2 males in each fold. Each fold consists of different sets of speakers resulting in a speaker-independent framework. To demonstrate the usefulness of prosodic information in Ao, statistical features of F_0 ST, loudness and VQT are used. The optimum values of c and γ parameters obtained after grid-search are 10 and 0.1, respectively. The results are presented in terms of μ and σ calculated from three-fold cross-validation performances. The classification performance is reported in Table 2.

Comparable average F1-scores are obtained for F_0 ST and loudness features. The MFCC feature outperforms the F_0 ST and loudness features. The best performance is obtained for the VQT feature compared to other individual features. The decent performance of statistical prosodic features encourages to further explore them in combination. The features are fused

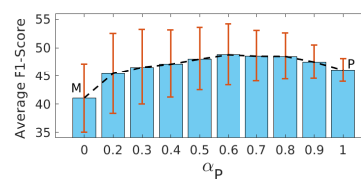


Figure 3: α variation for the 4 features combination reported in Table 2. M = MFCC, P = Prosody.

Table 3: Statistical significance analysis using ANOVA for the best five VQT features.

Features	df	F-value	p-value
jitter (μ)	2	27.1	<0.001
shimmer (σ)	2	37.6	<0.001
loudness peaks per sec	2	31.1	<0.001
voiced segments per sec	2	121.7	<0.001
unvoiced segments (μ)	2	140.1	<0.001

at score level according to the following equation,

$$S_{comb1} = \alpha_P S_{f1} + (1 - \alpha_P) S_{f2} \quad (1)$$

where, S_{f1} and S_{f2} are the prediction scores obtained for feature $f1$ and $f2$. The value of α_P varies from 0-1. The average F1-score for the combination of F_0 ST and loudness features is higher than the individual MFCC feature. Comparable performances are obtained for VQT+ F_0 ST and VQT+loudness. However, a lower σ is obtained for the VQT feature compared to VQT+ F_0 ST and VQT+loudness. The best performance is obtained for the combination of VQT+ F_0 ST+loudness. An improvement of $\approx 2\%$ in average F1-score is observed for the VQT+ F_0 ST+loudness combination than individual VQT. This improvement is obtained for $\alpha_P = 0.98$ value for the VQT feature. This implies that a higher weight is assigned to the VQT feature compared to other features. This signifies that VQT features are the most important features for the task. Moreover, higher performance for most statistical prosodic features than MFCC justifies the significance of the prosodic information for the current task.

The prosodic features (VQT+ F_0 ST+loudness) are finally fused with the MFCC feature to utilize the complementary information for the task. The performance for this combination at different α_P is illustrated in Figure 3. The performance is better for the α_P range of 0.6 – 1, i.e., the higher weight for the prosodic features in comparison to the MFCC features. The best performance is obtained for $\alpha_P = 0.6$ as shown in Figure 3. Here, a weight of 0.6 is assigned to the VQT+ F_0 ST+loudness combination and 0.4 weight is assigned to MFCC. The performance obtained for $\alpha_P = 0.6$ is reported in Table 2. The combination of prosodic and MFCC features resulted in a further improvement of $\approx 2\%$ in the average F1-score. However, the σ value increases significantly compared to the VQT+ F_0 ST+loudness combination. Hence, the performance at $\alpha_P = 0.9$ can be considered as the best result considering both the μ and σ values. These observations justify the effectiveness of prosodic features, specifically VQT, in identifying the three Ao dialects.

Statistical Analysis: ANalysis Of VAriance (ANOVA) is conducted to observe the statistical significance of the best performing VQT features. Table 3 shows the top 5 statistical significant VQT features with its p-value and F-value. The p-values indicate that the features are statistically different. Also, the F-value is related to the p-value inversely. Higher F-value indicates a significant p-value. It is noticed from the table that the temporal features have the highest F-value. Hence, this motivated us to use Bi-GRU to learn the temporal context of the LLD prosodic features.

4.3. Data Augmentation

To increase the dataset and avoid overfitting for the classification process, passage-level data were augmented to telephonic and reverberated speech. G.191 software was implemented to convert the original speech data into telephonic speech [36]. A pipeline process reported in [37] is used for simulation. On

Table 4: Classification performance of Ao dialects using LLD features in 3 sec segment duration reported in ($\mu \pm \sigma$).

Measures		Prosody (P)	MFCC (M)	P+M
Overall accuracy		58.53 \pm 5.81	45.78 \pm 4.17	59.09\pm5.56
F1-score	Changki	52.82 \pm 9.64	27.26 \pm 16.85	53.27\pm7.99
	Mongsen	57.79 \pm 23.35	43.33 \pm 10.28	58.05\pm22.88
	Chungli	58.57 \pm 13.71	61.06 \pm 8.52	60.06\pm13.10
	Average	56.39 \pm 7.18	43.88 \pm 5.05	57.13\pm6.96

the other hand, the dataset was also augmented to two types of reverberated speech using Roomsim toolbox [38]. The two categories of reverberated speech vary in terms of source and room sensor configurations. After data augmentation, the speech data resulted in ≈ 24 hours consisting of 384 passages across the three Ao dialects.

4.4. Attention-based Bi-GRU Classification Results

The attention-based Bi-GRU classifier is trained separately for 10 LLD prosodic features and 39-dimensional MFCC-LLD features. The original and augmented passage-level datasets are considered for training the classifier. While the testing is done using only the original speech. A four-fold speaker-independent cross-validation training approach is conducted to obtain the classification results. Each fold consists of 1 female and 1 male. Further, the training set is split into a 70 : 30 ratio to get the training and validation set, respectively. The model is trained for a segment duration of 3 sec to capture the temporal context of the speech signal. From Table 4, it is observed that the prosodic features give an improved performance in comparison to the results reported in Table 2. It is also noticed that the prosodic features outperform the baseline MFCC-LLD features. Additionally, for two features combination, the α_P value is varied using equation 1. The best performance is achieved for $\alpha_P = 0.7$, assigning higher weight to prosodic features. Also, there is an increase of about 1% in the average F1-score. Hence, these results substantiate the importance of prosodic information by capturing dialect-specific characteristics to classify the three dialects of Ao.

5. Discussion and Conclusions

This work presents a DID system to automatically identify the three Ao dialects using prosodic information. First, the statistical prosodic features extracted from trisyllabic words are used with SVM classifiers. Secondly, the attention-based Bi-GRU model learns the temporal context of LLD prosodic features. In addition, statistical analysis is conducted for the best performing feature, i.e., VQT, confirming the importance of temporal features. The overall classification performance with SVM and attention-based Bi-GRU justify the efficacy of prosodic information for the current task. The important point to notice is that the Chungli dialect gives the highest performance compared to Changki and Mongsen dialects (Table 2 and 4). The reason can be attributed to the following point. Chungli is known as the standard dialect of the language, as described in section 1. As a result of that, the Changki and Mongsen speakers tend to switch to the standard dialect in formal gatherings. Also, the written form is available only in the Chungli dialect and the speech data collected was in read speech. This might have biased the pronunciation of some speakers from Changki and Mongsen dialects towards the standard dialect. Hence, DID tasks in Ao can be extended in spontaneous speech data in the future. Also, as Ao is a tonal language with tone bearing unit in the vowel regions, formant analysis can be explored.

6. References

- [1] G. A. Grierson, *Linguistic survey of India*. Office of the superintendent of government printing, India, 1906, vol. 4.
- [2] A. R. Coupe *et al.*, *A phonetic and phonological description of Ao: A Tibeto-Burman language of Nagaland, North-East India*. Pacific Linguistics, Research School of Pacific and Asian Studies, 2003.
- [3] T. Temsunungsang, “Tonal correspondences in Ao languages of Nagaland,” in *22nd Himalayan Languages Symposium.*, 2016.
- [4] Directorate of census operation Nagaland, *District Census Handbook Mokochung*, Nagaland, 2011. [Online]. Available: <https://www.censusindia.gov.in>
- [5] A. R. Coupe, “The Acoustic and Perceptual Features of Tone in the Tibeto-Burman Language Ao Naga.” in *ICSLP*, 1998.
- [6] M. M. Clark, *The Ao Naga Grammar*. Assam Secretariat Printing Department, 1893.
- [7] D. Bruhn, “The tonal classification of Chungli Ao verbs,” *UC Berkeley PhonLab Annual Report*, vol. 5, no. 5, 2009.
- [8] F. Biadsy, J. Hirschberg, and N. Habash, “Spoken Arabic dialect identification using phonotactic modeling,” in *Proceedings of the EAACL 2009 Workshop on Computational Approaches to Semitic Languages*, ser. Semitic '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 53–61. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1621774.1621784>
- [9] J. K. Chambers and P. Trudgill, *Dialectology*. Cambridge University press, 1998, vol. 2nd edition.
- [10] Y. Lei and J. H. L. Hansen, “Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 85–96, 2011.
- [11] W. Lin, M. Madhavi, R. K. Das, and H. Li, “Transformer-based Arabic dialect identification,” in *2020 International Conference on Asian Language Processing (IALP)*. IEEE, 2020, pp. 192–196.
- [12] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, “Dialect identification using gaussian mixture models,” in *Odyssey*, 2004.
- [13] K. S. Rao and S. G. Koolagudi, “Identification of hindi dialects and emotions using spectral and prosodic features of speech,” *IJSCI: International Journal of Systemics, Cybernetics and Informatics*, vol. 9, no. 4, pp. 24–33, 2011.
- [14] S. S. Agrawal, A. Jain, and S. Sinha, “Analysis and modeling of acoustic information for automatic dialect classification,” *International Journal of Speech Technology*, vol. 19, no. 3, pp. 593–609, 2016.
- [15] N. B. Chittaragi and S. G. Koolagudi, “Automatic dialect identification system for kannada language using single and ensemble svm algorithms,” *Language Resources and Evaluation*, vol. 54, no. 2, pp. 553–585, 2020.
- [16] S. Shon, A. Ali, Y. Samih, H. Mubarak, and J. Glass, “ADI17: A fine-grained Arabic dialect identification dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8244–8248.
- [17] B. Ma, D. Zhu, and R. Tong, “Chinese dialect identification using tone features based on pitch flux,” *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006.
- [18] W.-W. Chang and W.-H. Tsai, “Chinese dialect identification using segmental and prosodic features,” *The Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1906–1913, 2000.
- [19] G. Mingliang, X. Yuguo, and Y. Yiming, “Semi-supervised learning based Chinese dialect identification,” in *2008 9th International Conference on Signal Processing*. IEEE, 2008, pp. 1608–1611.
- [20] P. N. Hung, N. T. Ha, T. Van Loan, V. X. Thang, and N. D. Chien, “Vietnamese dialect identification on embedded system,” *UTEHY Journal of Science and Technology*, vol. 24, pp. 82–87, 2019.
- [21] Z. Qi, Y. Ma, M. Gu, Y. Jin, S. Li, Q. Zhang, and Y. Shen, “End-to-end Chinese dialect identification using deep feature model of recurrent neural network,” in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*. IEEE, 2018, pp. 2148–2152.
- [22] Q. Zhang, Y. Ma, M. Gu, Y. Jin, Z. Qi, X. Ma, and Q. Zhou, “End-to-end Chinese dialects identification in short utterances using cnn-bigru,” in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. IEEE, 2019, pp. 340–344.
- [23] A. M. Ciobanu, S. Malmasi, and L. P. Dinu, “German dialect identification using classifier ensembles,” in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, 2018, pp. 288–294.
- [24] S. Kakouros, K. Hiovain, M. Vainio, and J. Šimko, “Dialect Identification of Spoken North Sámi Language Varieties Using Prosodic Features,” *arXiv preprint arXiv:2003.10183*, 2020.
- [25] T. C. Devi and K. Thaoroijam, “Vowel-based meeteilon dialect identification using a random forest classifier,” *arXiv preprint arXiv:2107.13419*, 2021.
- [26] 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.
- [27] 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.
- [28] M. Tzudir, P. Sarmah, and S. M. Prasanna, “Analysis and modeling of dialect information in Ao, a low resource language,” *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 2976–2987, 2021.
- [29] M. Tzudir, S. Baghel, P. Sarmah, and S. M. Prasanna, “Excitation source feature based dialect identification in ao—a low resource language,” *Proc. Interspeech 2021*, pp. 1524–1528, 2021.
- [30] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
- [31] T. Kislser, R. Winkelmann, and F. Schiel, “Styrian dialect classification: Comparing and fusing classifiers based on a feature selection using a genetic algorithm,” in *INTERSPEECH*, 2019, pp. 2393–2397.
- [32] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
- [33] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, “The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language,” in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016)*, Vols 1-5, 2016, pp. 2001–2005.
- [34] S. Goel, S. K. Pandey, and H. S. Shekhawat, “Analysis of emotional content in indian political speeches,” *arXiv preprint arXiv:2007.13325*, 2020.
- [35] N. B. Chittaragi, A. Limaye, N. Chandana, B. Annappa, and S. G. Koolagudi, “Automatic text-independent Kannada dialect identification system,” in *Information Systems Design and Intelligent Applications*. Springer, 2019, pp. 79–87.
- [36] R. G.191 ITU-T, “Software tools for speech and audio coding standardization, Int. Telecom. Union, Geneva, Switzerland, 2005.” [Online]. Available: <https://www.itu.int/rec/T-REC-G.191/en>
- [37] —, “ITU-T software tool library 2009 users manual,” Int. Telecom. Union, Geneva, Switzerland, 2009.”
- [38] E. Vincent and D. Campbell, *Roomsimove*.