INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Excitation Source Feature Based Dialect Identification in Ao - a Low Resource Language

Moakala Tzudir[1], *Shikha Baghel*[1], *Priyankoo Sarmah*[1], *S. R. Mahadeva Prasanna*[2]

[1]Indian Institute of Technology Guwahati, Guwahati - 781039, India
[2]Indian Institute of Technology Dharwad, Dharwad - 580011, India
{moakala, shikha.baghel, priyankoo}@iitg.ac.in, prasanna@iitdh.ac.in

Abstract

Ao is an under-resourced Tibeto-Burman tonal language spoken in Nagaland, India. There are three distinct dialects of the language, namely, Chungli, Mongsen and Changki. The objective of dialect identification is to identify one dialect from the other within the same language family. The goal of this study is to ascertain the potential of excitation source features for automatic dialect identification in Ao. In this direction, Integrated Linear Prediction Residual (ILPR), an approximate representation of source signal, is explored. The log Mel spectrogram of ILPR (S_{Ext}) signal is used to exploit the time-frequency characteristics of the excitation source. This work proposes attention based CNN-BiGRU architecture for automatic dialect identification tasks. Additionally, log Mel spectrogram (S_{VT}), extracted from the pre-emphasized speech signal, is used as a baseline method. The S_{VT} contains the vocal-tract characteristics of the speech signal. A significant performance improvement of (nearly) 6% accuracy is observed when the excitation source feature (S_{Ext}) is combined with the vocal tract representation (S_{VT}). To analyse the effect of segment duration, dialect identification performance is reported for three different durations, viz., 1 sec, 3 sec and 6 sec. The effect of gender in dialect identification task for Ao is also studied in this work.

Index Terms: Ao, tonal language, ILPR, ILPR Log Mel spectrogram, Log Mel spectrogram, dialect identification, CNN, GRU, attention

1. Introduction

Ao is an under-resourced tonal Tibeto-Burman language spoken in Nagaland, a North-Eastern state of India [1]. The Ao language has three lexical tones, viz. High (H), Mid (M) and Low (L) [2, 3]. There are three distinct dialects in Ao; namely, Chungli, Mongsen and Changki [4, 3]. The Chungli dialect is considered the standard dialect of the language. According to the Census of India 2011, the resident population of Ao in Nagaland is 227,000 [5]. Ao is the highest populated community among the Nagas (Nagaland residents), but only a few works are available related to Ao language. These works are mainly focused on the linguistic aspects of the language [2, 4, 3, 6, 7, 8]. However, the analysis and study of signal processing aspects of the Ao language have not been explored much in the literature. Hence, the present work is an attempt in the direction of filling the gap of signal processing analysis of the language. In this direction, tonal characteristics of Ao have been studied in our previous works [9, 10, 11, 12].

Dialect identification (DID, henceforth) is an emerging topic in the speech research community because of its implications for automatic speech recognition (ASR) [13]. The objective of DID is to identify a speaker's regional speech variety, within a predetermined language [13]. DID task is said to be a special case of language identification (LID, henceforth) [14]. However, DID is more challenging in comparison to LID as it identifies the different dialects within the same language family [15]. Moreover, the overlaps in vocabulary and phonetic features are more across two distinct dialects of a particular language than across two distinct languages [16]. The automatic DID is expected to enhance human-computer interaction applications [17]. It is also beneficial in providing new services for e-health and telemedicine, especially for older people [17]. Therefore, this work is motivated to attempt identification of the three dialects of Ao language.

1.1. Related Works

In the literature, a body of work is dedicated to build dialect identification systems in major languages of the world such as Arabic, Chinese and Spanish. Several works explored the phonotactic information based phone-recognition language modeling (PRLM) approach for DID task [13, 18]. The most widely used spectral features, namely, Mel Frequency Cepstral Coefficient (MFCC, henceforth) and Shifted-Delta Cepstral (SDC) features, are studied for DID systems in languages such as Arabic, Chinese, Kannada and Spanish [19, 20, 21, 22]. There are a number of research works in different tonal languages such as Chinese, Ao and Vietnamese for DID task, where spectral features like MFCC and SDC are used along with prosodic features such as pitch [15, 23, 24, 25, 10, 12, 26]. Some existing works explored the bottleneck features (BNF, henceforth) for DID and showed improved accuracy over the baseline MFCC features in Arabic and Chinese languages [27, 28]. Recently, the potential of the filter bank features has been studied for the DID task [14, 29, 30]. The studies conducted for DID are mostly done on high-resourced languages. For example, in case of Arabic dialects, each variety is the standard dialect for different countries, which are used extensively in news broadcasts with available written scripts. However, in case of low resource language like Ao, there is only one standard dialect of the language, i.e., Chungli, with no available written script for the other two dialects.

1.2. Motivation and Contribution

Most of the existing works have explored the Vocal Tract (VT, henceforth) information to discriminate different dialects of a language. The MFCC, a VT representation, is widely used in the existing works related to DID. However, the present work believes that the excitation source also encapsulates essential characteristics of a tonal language. In a tonal language, fundamental frequency F_0 plays an essential role. As such, Ao is a tonal language with three lexical tones, namely, High (H), Mid

http://dx.doi.org/10.21437/Interspeech.2021-1672

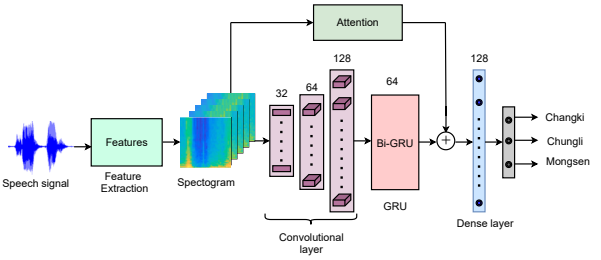


Figure 1: *Proposed DID system: Attention based CNN-BiGRU architecture.*

(M) and Low (L), where the tone assignment in lexical words may differ among the three dialects of Ao [3]. From Table 1, it can be observed that the tonal specifications of the same word are different based on the dialects [11]. These tonal differences across the three dialects motivated us to investigate the characteristics of the excitation source for DID in Ao.

Excitation source features have been explored for LID tasks under clean and degraded environment [31, 32]. D. Nandi et al. [32] found that the characteristics of excitation and articulatory source are distinct for each sound unit. Thus, the characteristics of the same sound unit of a tonal language may differ due to the co-articulation effects and the different assignment of lexical tones across the dialects. Considering this, the excitation source may carry dialect-specific information which can be utilized to distinguish the three dialects in Ao. Hence, excitation source characteristics are explored for an automatic DID system for Ao. In the existing works related to DID in Ao, traditional classifiers such as GMM is used. However, the present work proposes a novel deep learning based architecture for the current task. The novelty of the present work are listed below.

- The ILPR, an approximate representation of source signal, is studied in this work. The time-frequency representation is used for DID in Ao.
- An attention based Convolutional Neural Network-Bidirectional Gated Recurrent Unit (CNN-BiGRU, henceforth) model is proposed. The proposed architecture utilizes the spatial (learned using CNN)-temporal context (learned using Bi-GRU) patterns of the spectrogram for the classification. Since Ao is a tonal language, the attention mechanism is used along the frequency direction. The attention mechanism provides higher weights to the frequency bins, which carry the discriminative patterns across dialects.
- The F_0 is an essential characteristic of a tonal language. However, F_0 varies with gender. Hence, this work explores the effect of gender on the DID task in Ao language.
- The segment duration also plays a critical role in DID. Therefore, this work reports classification results for three different segment durations, namely 1 sec, 3 sec and 6 sec.

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

Table 1: *Examples of trisyllables with different tone assignment across the three dialects.*

Words	Changki	Mongsen	Chungli	Gloss
Temesen	HMM	HHL	MHL	'liver'
Wamaba	LHM	LHL	HHL	'slice into pieces'

2. Speech Dataset

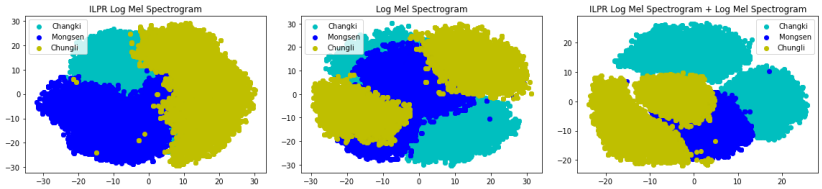
The speech data was collected from speakers of the three dialects, namely, Chungli, Mongsen and Changki. The Chungli dialect (spoken in Mopungchuket village), Mongsen dialect (spoken in Khensa village) and Changki dialect (spoken in Changki village) is used for this work. The three villages fall under the Mokokchung district in Nagaland, India. Among the three dialects of Ao, the Chungli dialect is considered the standard variety of the language. Hence, the Chungli dialect is used when it comes to formal occasions and writings, as the literature is available only on the standard dialect. The speech database was recorded from 8 native speakers (4 males and 4 females) for each dialect. All the speakers were asked to read a passage from the bible, "The parable of the prodigal son". As the text in the bible is written in the standard dialect, the passage was translated for the speakers of Mongsen and Changki in their respective dialects. The passage was recorded in four sessions for all the speakers across the three dialects. Hence, the database consisted of 96 passages (approximately 6 hours) in total by speakers of all three dialects. The speech data were recorded using a TASCAM DR-100 MKII, 2-channel portable digital recorder. It was connected to a Shure SM10A head-mounted microphone for high-quality recordings. The recordings took place in a real-world environment scenario. All the speakers spoke English and Nagamese, a creolized variety of the Assamese language apart from their native Ao dialects.

3. Proposed DID System for Ao language

This section briefly describes the feature extraction (sub-section 3.1) and the proposed DID system (sub-section 3.2).

3.1. Integrated Linear Prediction Residual (ILPR)

The analysis and processing of speech signals are composed of separate representations for vocal tract systems and excitation source. In Linear Prediction (LP, henceforth) analysis, the LP coefficients (LPC, henceforth) constitute the time-varying vocal tract information and the residual signal represents the excitation source characteristics [33]. In this work, ILPR signal is explored as an approximate representation of the voice source signal. The extraction of ILPR signal is briefly explained next. First, the speech signal $x(n)$ is pre-emphasized in order to enhance the high-frequency components. This pre-emphasized speech $x_e(n)$ is then used to predict the LPC values using LP analysis. In order to obtain the ILPR signal, the original speech signal $x(n)$ is passed through the inverse filter obtained from LPCs [34]. The LP order p is set to $p = f_s + 4$, where f_s is the sampling frequency in kHz [34]. In this study, the time-domain ILPR signal of each frame is transformed to frequency-domain by computing the Mel spectrogram. The Mel filter bank contains 40 overlapping triangular filters. The logarithm is applied on the Mel spectrogram and the resultant time-frequency representation is considered as ILPR log Mel spectrogram (ILPR-LMS, henceforth). Hence, the ILPR-LMS representation exploits the time-frequency characteristics of the excitation source signal for DID task. In this work, the speech signal is processed



(a) (b) (c)

Figure 2: *t-SNE plot for (a) S_{Ext}, (b) S_{VT}, (c) Combination of S_{Ext} and S_{VT}.*

using a frame size of 20 ms and a shift of 10 ms with a 16 kHz sampling frequency.

3.2. Attention Based CNN-BiGRU Classifier

The architecture of the attention based CNN-BiGRU Model is illustrated in Figure 1. The proposed architecture learns spatial information of the spectrogram using CNN architecture. The temporal context of different dialects is captured by the Bi-GRU layer. Since Ao is a tonal language, the proposed model is motivated to use an attention mechanism [35] along frequency direction. This provides higher weights to those frequency bins, which have more discriminative information for the classification task. Hence, the proposed architecture utilizes spatial, temporal and frequency based attention for DID in Ao language. The model consists of three convolutional layers with 32, 64 and 128 number of kernels, respectively. The size of kernel is $(3, 3)$ with a stride of $(2, 1)$. The output of each convolutional layer is batch normalized and passed through a Maxpooling layer with a pool size of $(2, 1)$. The output of the last convolutional layer is passed through a bidirectional GRU layer with 64 units in both directions. Simultaneously, the input spectrogram is passed through an attention mechanism where it emphasizes the essential information of the spectrogram along frequency direction. The output of the attention module and Bi-GRU is concatenated. The concatenated output is flattened and fed to a dense layer with 128 neurons. A dropout of 0.4 is used after the dense layer. Finally, the output layer contains three nodes with softmax activation. The model is trained for 50 epochs with a batch size of 32. An early stopping criteria is used to avoid overfitting of the model. An initial learning rate of 0.0001 and categorical cross-entropy loss are used for training the model. All the layers (except the output layer) use ReLU activation.

4. Experiments and Results

This section describes the experimental settings used to evaluate the present work.

4.1. Baseline Method

MFCC features are one of the most widely used vocal tract representations of speech in DID task [19, 20, 21, 22]. Therefore, the Mel-spectrogram used to extract MFCC features is considered as the baseline method in the present work. The Mel-spectrograms are extracted using a Mel filter bank containing 40 overlapping triangular filters. The pre-emphasized speech signal is used to obtain the Mel-spectrograms. The logarithmic operation is performed on Mel-spectrograms and the resultant time-frequency representation is considered as the log Mel spectrogram (LMS, henceforth). The LMS is used to provide complementary information with ILPR-LMS.

4.2. Classification Using Attention Based CNN-BiGRU

To demonstrate the effectiveness of the proposed approach for DID in Ao, the following experimental setup is used. The speech data is divided into four non-overlapping folds by ensuring equal number of male and female speakers' data in each fold. For each iteration, any three folds' data is considered as the train set and the remaining fold is the test set. The train set is further split into 70 : 30 ratio to get the train and validation set, respectively. Therefore, four different sets of train, validation and test data are obtained from four-folds. Hence, the proposed approach is evaluated four times. The results are reported in terms of mean (μ) and standard deviation (σ) of the performances obtained from four iterations. All the folds contain different sets of speakers. Thus, the speakers of the test set are different from the train set speakers. Hence, the proposed approach is evaluated in a speaker-independent framework. To utilize the temporal context of the speech, the model is trained for a segment duration of 1 sec. The performance is reported in terms of overall accuracy and F1-score measures.

The classification performance of 4-fold cross-validation is reported in Table 2 for a segment duration of 1 sec. The highest F1-score is obtained for Mongsen dialect than Changki and Chungli in case of ILPR-LMS (S_{Ext}, henceforth). The overall performance of S_{Ext} is lower than the LMS (S_{VT}, henceforth). However, the decent performance of S_{Ext} motivated this study to further explore the complementary information of S_{Ext} in combination with S_{VT}. The S_{Ext} carry excitation source characteristics, while S_{VT} contains vocal-tract information. The combination of S_{Ext} and S_{VT} representations is expected to perform better than individual features. Figure 2 shows the t-SNE plots for S_{Ext} (Figure 2(a)), S_{VT} (Figure 2(b)) and their combination (Figure 2(c)). The data of one fold is used for t-SNE plots. In Figure 2, the embeddings (obtained from learned classifier) of individual features are used to show the separability across the dialects. It is observed that there are overlaps in the three dialects for individual features. However, when the embedding of S_{Ext} is concatenated with the embedding of S_{VT}, the separability across the three dialects increases.

For the feature combination, the S_{Ext} and S_{VT} spectrograms are converted into embeddings by using their respective trained model. The concatenated output of the Bi-GRU and attention model is considered as the embedding. The embeddings of S_{Ext} and S_{VT} are further concatenated and fed to a DNN

Table 2: *Classification performance of Ao dialects using S_{Ext} and S_{VT} features for 1 sec segment duration.*

Measures		S_{Ext} ($\mu \pm \sigma$)	S_{VT} ($\mu \pm \sigma$)	S_{Ext}+S_{VT} ($\mu \pm \sigma$)
Overall accuracy		75.32±9.21	85.92±9.31	91.89±4.10
F1-score	Changki	71.93±15.73	84.41±11.47	92.12±7.05
	Mongsen	79.58±2.31★	85.81±8.71	89.12±3.34
	Chungli	72.40±13.37	87.66±8.39	**93.77±5.19**
	Average	74.63±9.07	85.96±9.34	91.67±4.07

Table 3: *Classification performance of Ao dialects using S_{Ext} and S_{VT} features for 3 sec and 6 sec segment durations.*

Measures		3 sec segment			6 sec segment		
		S_{Ext} $(\mu \pm \sigma)$	S_{VT} $(\mu \pm \sigma)$	S_{Ext}+S_{VT} $(\mu \pm \sigma)$	S_{Ext} $(\mu \pm \sigma)$	S_{VT} $(\mu \pm \sigma)$	S_{Ext}+S_{VT} $(\mu \pm \sigma)$
Overall accuracy		82.08±11.66	89.88±13.83	97.29±1.61	82.15±11.72	90.02±13.92	97.33±1.60
F1-score	Changki	79.08±13.96	87.89±15.86	**97.54±2.53**	79.40±13.81	88.04 ±16.02	**97.53±2.55**
	Mongsen	92.74±6.64★	89.60±13.72	96.99±3.03	92.80±7.08★	89.60±13.73	97.06±2.96
	Chungli	73.17±18.61	91.98±12.69	97.30±2.72	73.08±18.66	92.25±12.77	97.35±2.69
	Average	81.66±12.05	89.83 ±13.99	97.27±1.58	81.76±12.09	89.96±14.09	97.31±1.57

architecture (exactly same as used in the proposed model). The DNN architecture is trained separately over the concatenated embeddings. The results for the combination are reported in Table 2. A significant improvement of (nearly) 6% accuracy is obtained for the combination of S_{Ext} and S_{VT}. The highest F1-score of 93.77 ± 5.19 is obtained for the Chungli dialect. One possible reason could be that the speakers of the Mongsen and Changki dialect may have the influence of the standard dialect. Hence, it may be prone to capture the segmental information of the standard dialect for short segment duration like 1 sec. The μ values of all the measures are improved and σ decreases for the combination. This signifies that S_{Ext} carry essential complementary information about the different dialects which is not being captured by S_{VT} representations. Thus, S_{Ext} is worth exploring for DID in Ao language.

4.3. Effects of Gender

Ao is a tonal language and the tone information of speech signal changes according to different gender. Thus, this work is motivated to study the effect of gender on dialect identification tasks. The classification is performed with four-fold cross-validation as described in sub-section 4.2. Each fold contains one female speaker and one male speaker data. The models trained over 1 sec duration (used in Table 2) are used for analyzing the effect of gender. During training, data from both the gender are used. However, gender-specific data is tested separately using the trained models. The male and female performances in terms of F1-score are shown in Figure 3. It can be observed that the performance of male speakers is better than female speakers in all the dialects. This trend is consistent across individual feature and their combination (except S_{VT} in mongsen dialect). In general, the fundamental frequency of female speakers is higher than the male speakers. The female speakers are prone to have higher variations among the different tones in a conversation in comparison with male speakers. As such, male speakers are more likely to incline towards a stationary signal while computing in a frame. This could be one of the reasons for better dialectal information in male speakers. The performance difference between both the gender decreases for the feature combinations ($S_{Ext} + S_{VT}$) in comparison to individual features. This again verifies the potential of excitation source features for the current task.

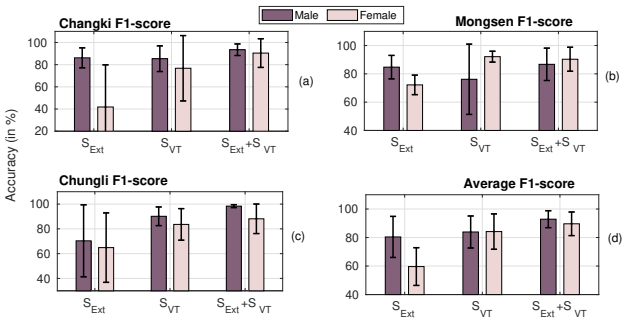


Figure 3: *Classification performance of F1-score across the three dialects for the effects of gender. Note that the ranges of vertical axes are different in (a), (b), (c) and (d).*

4.4. Effects of Segment Duration

The perceptual difference across dialects can be easily identified by humans if a speech segment of sufficient duration is available. However, it would become difficult for humans also, if a short segment of speech signal is given. Hence, the segment duration is expected to be an important factor in DID. Therefore, this work is motivated to study the effect of segment duration on the proposed approach. The trained models (over a segment duration of 1 sec) are used to evaluate the performance over 3 sec and 6 sec. The scores for each 1 sec context window are predicted using the trained models. A shift of one frame is used for moving 1 sec (99 frames) context window over the whole speech signal. Further, the mean scores of all the frames of a segment duration are calculated and the final class label is assigned based on the highest score. The performance for 3 sec and 6 sec durations are reported in Table 3. A prominent performance improvement of 7% (approximate) is observed for 3 sec duration in comparison to 1 sec duration in case of S_{Ext}. An increment of 4% accuracy is observed for S_{VT}. However, a performance improvement of 6% accuracy is obtained for the combination of S_{Ext} and S_{VT}. The important point to notice is that the σ is significantly reduced for the combination. This again justifies the importance of excitation source characteristics in DID. The performance obtained for 6 sec segment duration is almost comparable with that of 3 sec segment duration. Hence, it can be said that 3 sec segment duration can be a good segment choice for detecting three dialects in Ao language.

5. Conclusion and Future Work

A novel deep learning based DID system is proposed in this work to automatically identify the three dialects in Ao using excitation source features. An attention based CNN-BiGRU classifier is proposed. The results presented in this work confirm that the excitation source information, i.e., S_{Ext}, can be used to build a more discriminative system in combination with vocal tract information. Ao being a tonal language, the effects of gender is studied in this work where the performance of male speakers is better than the female speakers. This gender-specific trend is consistent across individual and combined features. Additionally, the effects of segment duration are also studied over 3 sec and 6 sec. A notable performance improvement is obtained for 3 sec segment duration than 1 sec duration.

Prosodic characteristics play an important role in tonal languages. Thus, it is expected that the prosodic information of the three dialects carry dialect-specific information. In the future, this work can be extended by exploiting the prosodic features with additional spectral and excitation source features. Since the spoken and textual resources are limited in this work, another possible direction would be the extension of the database.

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, "The Acoustic and Perceptual Features of Tone in the Tibeto-Burman Language Ao Naga." in *ICSLP*, 1998.

[3] T. Temsunungsang, "Tonal correspondences in Ao languages of Nagaland," in 22^{nd} *Himalayan Languages Symposium.*, 2016.

[4] 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.

[5] Directorate of census operation Nagaland, *District Census Handbook Mokokchung*, Nagaland, 2011. [Online]. Available: https://www.censusindia.gov.in

[6] M. M. Clark, *The Ao Naga Grammar*. Assam Secretariat Printing Department, 1893.

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

[8] T. Temsunungsang, *The structure of Mongsen: Phonology and Morphology*. Hyderabad: Hyderabad Central University MPhil thesis, 2003.

[9] 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.

[10] 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.

[11] P. Gogoi, M. Tzudir, P. Sarmah, and S. Prasanna, "Automatic tone recognition of Ao language," in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 1005–1008.

[12] 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.

[13] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proceedings of the EACL 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

[14] 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.

[15] 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.

[16] W.-H. Tsai and W.-W. Chang, "Chinese dialect identification using an acoustic-phonotactic model," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[17] A. Etman and A. L. Beex, "Language and dialect identification: A survey," in *SAI Intelligent Systems Conference (IntelliSys), 2015*. IEEE, 2015, pp. 220–231.

[18] M. A. Zissman, T. P. Gleason, D. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech," in *ICASSP*, 1996.

[19] 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.

[20] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using gaussian mixture models," in *Odyssey*, 2004.

[21] 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.

[22] 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.

[23] 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.

[24] W.-H. Tsai and W.-W. Chang, "Discriminative training of gaussian mixture bigram models with application to Chinese dialect identification," *Speech Communication*, vol. 36, no. 3, pp. 317–326, 2002.

[25] G. Mingliang and X. Yuguo, "Chinese dialect identification using clustered support vector machine," in *2008 International Conference on Neural Networks and Signal Processing*. IEEE, 2008, pp. 396–399.

[26] 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.

[27] Q. Zhang and J. H. Hansen, "Dialect recognition based on unsupervised bottleneck features." in *INTERSPEECH*, 2017, pp. 2576–2580.

[28] ——, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 873–882, 2018.

[29] 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.

[30] 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.

[31] D. Nandi, D. Pati, and K. S. Rao, "Implicit excitation source features for robust language identification," *International Journal of Speech Technology*, vol. 18, no. 3, pp. 459–477, 2015.

[32] ——, "Parametric representation of excitation source information for language identification," *Computer Speech & Language*, vol. 41, pp. 88–115, 2017.

[33] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[34] S. Baghel, S. M. Prasanna, and P. Guha, "Excitation source feature for discriminating shouted and normal speech," in *2018 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2018, pp. 167–171.

[35] S. Goel, S. K. Pandey, and H. S. Shekhawat, "Analysis of emotional content in indian political speeches," *arXiv preprint arXiv:2007.13325*, 2020.