The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Dialect Identification Using Tonal and Spectral Features in Two Dialects of Ao

Moakala Tzudir[1], *Priyankoo Sarmah*[2,3], *S. R. Mahadeva Prasanna*[1,2,3]

[1] Department of Electronics and Electrical Engineering,
[2]Department of Humanities and Social Sciences
[3]Center for Linguistic Science and Technology
Indian Institute of Technology Guwahati, Guwahati-781039, India
[4]Department of Electrical Engineering,
Indian Institute of Technology Dharwad, Dharwad-580011, India

(moakala, priyankoo, prasanna)@iitg.ac.in

Abstract

Ao is an under-resourced Tibeto-Burman tone language spoken in Nagaland, India, with three lexical tones, namely, high, mid and low. There are three dialects of the language namely, Chungli, Mongsen and Changki, differing in tone assignment in lexical words. This work investigates if the idiosyncratic tone assignment in the Ao dialects can be utilized for dialect identification of two Ao dialects, namely, Changki and Mongsen. A perception test confirmed that Ao speakers identified the two dialects based on their dialect-specific tone assignment. To confirm that tone is the primary cue in dialect identification, F_0 was neutralized in the speech data before subjecting them to a Gaussian Mixture Model (GMM) based dialect identification system. The low dialect recognition accuracy confirmed the significance of tones in Ao dialect identification. Finally, a GMM-based dialect identification system was built with tonal and spectral features, resulting in better dialect recognition accuracy.

Index Terms: Ao, Changki, Mongsen, tone language, dialect identification.

1. Introduction

Dialect identification is one of the important research topics in the speech research community because of its implications to automatic speech recognition (ASR). The task of dialect identification is to recognize a speaker's regional speech variety, within a predetermined language [1]. The problem of dialect identification has been viewed as more challenging, than that of language identification, due to the phonetic similarities between the dialects of a particular language [2]. Overlaps in vocabulary and phonetic features are more across two distinct dialects of a particular language than across two distinct languages [3]. In the current work, dialect identification is explored in two dialects of Ao, a Tibeto-Burman language, spoken in the Northern part of Nagaland, India. Ao has three major dialects namely, Chungli, Mongsen and Changki [4]. It is a tone language and is reported to have three lexical tones, namely, high (H), mid (M) and low (L) [5]. While the number and types of tones do not differ across the Ao dialects, tone assignment is distinct, even for the same words among the dialects [6]. For instance, the high tone of Changki dialect corresponds to a low tone of Mongsen and Chungli dialects and vice-versa [6]. However, there are not many acoustic analysis on the varieties of Ao that describe the assignment of tones. Nevertheless, the canonical F_0 patterns for Changki and Mongsen tones have been described earlier, where the three tones can be categorized, based on the height of the tones [7].

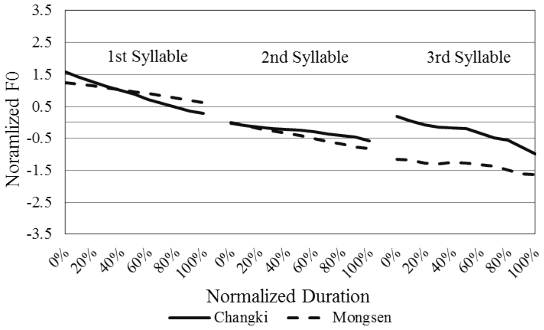

Figure 1: *Dialect-wise tone variation in the trisyllabic word 'jemrepba'- trampled, with normalized duration and pitch*

While Ao is an under-described language, there are a few works on Ao language, including grammatical descriptions, [8, 9, 10], an early dictionary [11], later updated in 2013 and a description of tones in the Chungli dialect [12]. Similarly, there are a few works available for the Mongsen dialect [5, 4, 13, 14, 15], however, Changki dialect is the least documented one with only one work available on automatic discrimination of Ao dialects [7]. All these works are descriptive in nature and except for one [7], no previous attempts have been made in Ao automatic dialect identification.

Cross-linguistically, there are numerous works in dialect identification, both for tone and non-tone languages. Attempts at dialect identification are classified into two modelling schemes namely, phonotactic modeling such as phone recognizer language modeling (PRLM), where phone recognizers are used to tokenize speech data from target languages or dialects to be classified. The target language or dialects are assigned likelihood scores based on the language model of an independent language. On the other hand, in acoustic modeling, spectral features are modelled directly for languages or dialects [16].

Automatic dialect identification efforts in languages, such as, Arabic are conducted using PRLM in Arabic dialects [1]. In case of Spanish, automatic dialect identification in two dialects of Spanish used the parallel PRLM approach [17]. Use of GMM with shifted-delta cepstral features in the identification of two Spanish dialects is reported in a previous study [18]. For Arabic and Spanish dialects, KLD-GMM and FSD-GMM with MFCC features are reported previously [19]. It is also reported that a hybrid of SVM-GMM was used in three Spanish dialects for features such as formants, LSP (Line Spectral Pairs) and MEPZ

10.21437/SLTU.2018-29

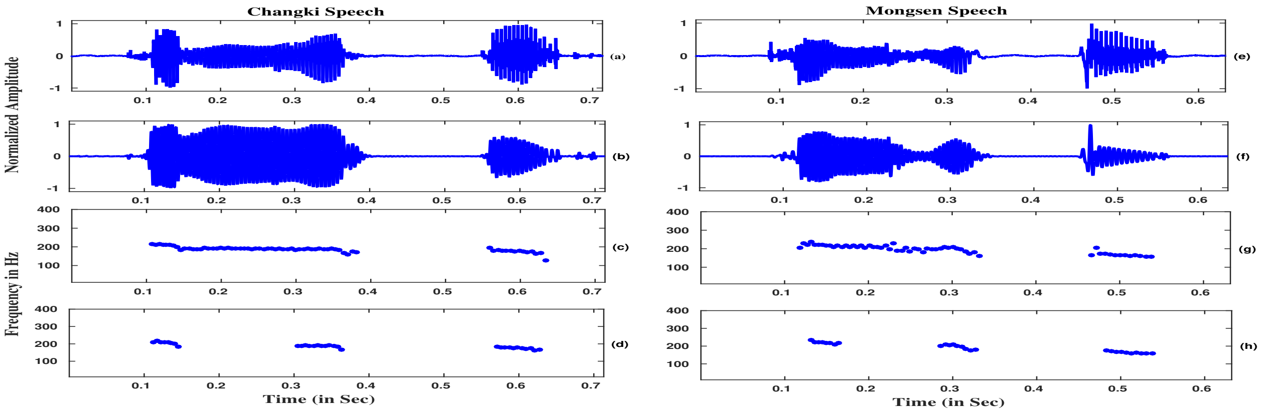

Figure 2: *ZFF derived pitch tracks for Ao tones. (a) Changki speech signal from a female speaker for the target word 'jemrepba'. (b) ZFF output of (a). (c) Pitch contour of (a). (d) Pitch contour of (a) for the vowel regions. (e) Mongsen speech signal from a female speaker for the target word 'jemrepba'. (f) ZFF output of (e). (h) Pitch contour of (e). (g) Pitch contour of (e) for the vowel regions.*

(MFCCs + Energy + pitch) [20].

In case of tone languages, attempts are made at dialect identification primarily in the Chinese languages. In Chinese dialect identification, previous studies have used both pitch and Mel Frequency Cepstral Coefficient (MFCC) feature in Gaussian Mixture Model (GMM) based classification [2, 21]. Similarly, Hidden Markov Model is used in automatic dialect identification in previous works on three Chinese dialects [22].

The goal of this study is to determine if the two dialects of Ao, namely, Changki and Mongsen, can be automatically identified depending on the differences in tone assignment in words that are common across the two dialects. In order to demonstrate that tonal information is used to identify the two Ao dialects, z-score normalized average F_0 plots of the tones in the trisyllabic word 'jemrepba', for both dialects are plotted in Figure 1. The plot contains three parts, where the first part of $0\% - 100\%$ represents the F_0 plot of the first syllable of the word, the second and the third parts represent the F_0 values of the second and the third syllables, respectively. As seen in the plot, the tone assignment is different for the same word in the two dialects. For the same word, the tones assigned in the three syllables of Changki are HMM, whereas, the same is HML in the Mongsen dialect. Based on this observation, in this study, F_0 is considered to be the primary phonetic feature for dialect identification. In order to confirm the role of F_0 in dialect identification and to demonstrate the effectiveness of including F_0 information in dialect identification, three experiments were conducted in this work:

- A perception test on Ao native speakers to confirm phonetic salience in two dialects of Ao
- Automatic dialect identification with F_0 information neutralized
- Automatic dialect identification with both tonal and spectral features

The rest of the paper is arranged as follows: Section 2 describes the methodology and details the experiments conducted. Section 3 provides information regarding the database and describes the results of the experiments conducted. Finally, Section 4 summarizes and concludes the work.

2. Methodology

2.1. Method of the perception test

For the perception test, 5 Changki subjects and 5 Mongsen subjects, 3 females and 2 males for each dialect, were considered. The stimuli contains 40 underived trisyllabic words, common between the two dialects [23]. The words were spoken by 12 speakers from each of the two dialects, in citation form, sentence frames and in an example sentence, resulting in a total of 1440 tokens for each dialect. From each dialect, out of 1440 tokens, 50 were selected as stimuli for the perception test. Hence, a total of 100 tokens for the two dialects, repeated 3 times, resulting in a total of 300 stimuli were presented to 10 speakers of the two Ao varieties. The experiment was a Multiple Forced Choice experiment, implemented using Praat [24] on a laptop computer where the speakers had three choices to choose from, namely, 'Changki', 'Mongsen' and 'Can't Decide'. The stimuli were presented randomly to the subjects and the subjects were asked to listen to and click one of the three options that appeared on the screen. Each speaker took about 30 minutes to complete the task. After the experiment was completed, the results were extracted to an excel sheet for further analysis.

2.2. Feature extraction from the production data

F_0 is the primary phonetic feature associated with tones. Hence, to estimate F_0, zero-frequency filtering (ZFF) approach was used as this approach is considered to be more effective compared to the conventional pitch estimation approaches. In ZFF, the instant of significant excitation signifying the location of epochs, provides accurate estimation of the instantaneous fundamental frequency (F_0) [25]. From the estimated F_0 using ZFF, a set of features are used to capture average F_0, slope of F_0 and rate of change of F_0 in the tones of the two Ao dialects in this study. Considering the variation of tones in the Ao dialects, these features are expected to facilitate dialect identification in Ao. Apart from average F_0, ΔF_0 and $\Delta\Delta F_0$ features are used as they help to capture the rate of change of the F_0 contour. ΔF_0 is the difference between successive F_0 values and $\Delta\Delta F_0$ is the difference between successive ΔF_0. Finally, as shown below, a set of tonal feature vectors are arranged to determine their effectiveness in dialect identification in Ao.

- F_0 : $3D$ feature vector
- ΔF_0 : $3D$ feature vector
- $\Delta\Delta F_0$: $3D$ feature vector
- F_0 + ΔF_0 : $6D$ feature vector
- F_0 + $\Delta\Delta F_0$: $6D$ feature vector
- ΔF_0 + $\Delta\Delta F_0$: $6D$ feature vector
- F_0 + ΔF_0 + $\Delta\Delta F_0$: $9D$ feature vector

In order to capture the spectral changes between the two dialects, MFCC features are extracted. MFCCs have proven to be one of the most successful features that capture the vocal tract information of speech [26]. The perception of frequency components by the human auditory system is on a logarithmic scale. Therefore, nonlinear mel filter has been designed to emphasize the lower frequency components over the higher ones. MFCCs are extracted from a speech signal using normal block processing approach. A frame size of 20 ms with a shift of 10 ms was used. Hamming window was used during the framing of the speech signal. Cepstrum gives the static vocal tract shape (VTS) information but Δcepstrum gives the change in VTS information and ΔΔcepstrum gives the rate of change in VTS information. Hence, $39D$ MFCC feature set is considered for this study.

2.3. Automatic dialect identification with neutralized $\mathbf{F_0}$

To confirm the importance of tonal features in dialect identification, an experiment was conducted by attempting dialect identification devoid of F_0 variation. F_0 was flattened to a static F_0 value of 100 Hz for the three syllables in trisyllables. The F_0 values were computed using ZFF approach. GMM is the baseline of acoustic modeling. The basis of using GMM is that the distribution of feature vectors extracted from speech signal can be modeled by a mixture of Gaussian densities [27]. The parameter of GMM was estimated using the iterative expectation-maximization (EM) algorithm. In the dialect identification process, a GMM is created for each dialect by taking the flattened F_0 features with 32 mixtures. During testing, a test dialect is given, which is represented by a sequence of feature vectors, and the log likelihood score of each model is calculated. The process with the highest likelihood score determines the hypothesized dialect for the flattened F_0 feature. If tonal features are crucial for dialect identification, we expect that normalization of F_0 will result in poor dialect identification. On the other hand, if tonal features are redundant, we expect to see better dialect identification, even when tonal differences are eliminated by neutralizing F_0 variations.

2.4. Automatic dialect identification using tonal and spectral features

Assuming that tone assignment differences in the same lexical items signal dialectal information in Ao, we consider F_0 features to be effective in identifying Ao dialects. In order to estimate F_0, the ZFF approach is used. F_0 is computed for the vowel regions as they are the Tone bearing units (TBU) in Ao. Figure 2 (d), (h) shows the pitch contours for both the dialects for the target word 'jemrepba'. In order to normalize F_0 across genders, z-score normalization is used to convert the F_0 values to z-score values. This method is considered to be one of the best F_0 normalization methods [28, 29]. Apart from average F_0, ΔF_0 and $\Delta\Delta F_0$ are computed for each tone and used as tonal features in dialect identification. In the dialect identification process, a GMM is created for each dialect by taking the feature vectors as shown in Figure 3. The number of mixture components is empirically chosen for the dataset. While considering the tonal feature as the feature vector, 32 mixtures were used whereas, 1024 mixtures were used for MFCC as the feature vector. During testing, a test dialect is represented by a sequence of feature vectors, and the log likelihood produced by the model is calculated. The Changki and Mongsen dialects are represented by the models λ_1 and λ_2 respectively. The log likelihood ratio $S(x)$ for a test feature vector x is computed as,

$$S(x) = log\,p(x|\lambda_1) - log\,p(x|\lambda_2) \tag{1}$$

where, $p(x|\lambda_1)$ and $p(x|\lambda_2)$ are the probability density function of the variable x given λ_1 and λ_2. By computing the log likelihood of these processes, the process with the highest likelihood determines the dialect for both tonal and MFCC features. Further, for better dialect modeling the combined scores S_c of the classifier, trained using MFCC and F_0 is obtained by,

$$S_c = \alpha S_m + (1 - \alpha) S_f \tag{2}$$

where, S_m and S_f denotes the scores obtained using MFCC and F_0 features. α represents a scalar value which ranges from 0 to 1 at an interval of 0.05 in Equation 2. Based on the highest likelihood combination scores, the process determines the hypothesized dialect.

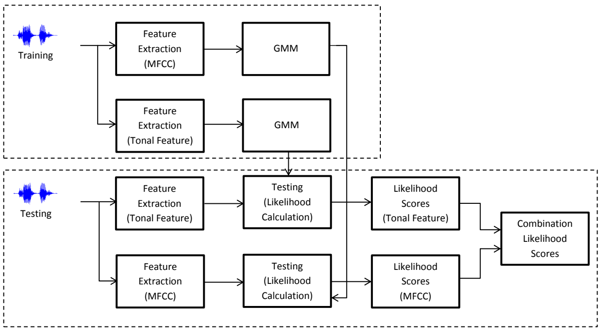
Training
Feature Extraction (MFCC)
GMM
Feature Extraction (Tonal Feature)
GMM
Testing
Feature Extraction (Tonal Feature)
Testing (Likelihood Calculation)
Likelihood Scores (Tonal Feature)
Feature Extraction (MFCC)
Testing (Likelihood Calculation)
Likelihood Scores (MFCC)
Combination Likelihood Scores

Figure 3: *Overall dialect identification block diagram*

3. Database and Results

For this study, Changki dialect spoken in Changki village in the western Changkikong range and Mongsen dialect spoken in Khensa village in the southern Onpangkong range were taken into consideration. For each of the two 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 both the dialects, resulting in a total of 1440 trisyllabic utterances in each dialect. In order to account for session variability, the recording process was repeated with the same set of speakers after 2 months. The old session data was used for training and the new session data was used for testing. 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 [24]. However, for this study, the speech signal is re-sampled to 8KHz for pitch estimation using ZFF.

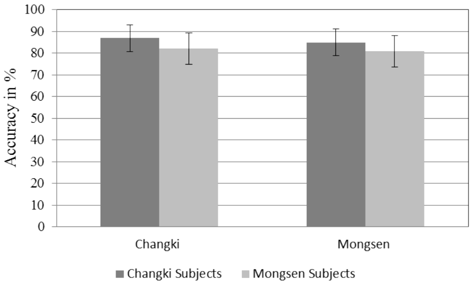

Figure 4: *Averaged accuracy rate for perception test from 5 subjects for each dialect with 1500 tokens*

3.1. Perception Test

The results of the perception test showed that the speakers of each dialect are able to identify Ao dialects based on the tone specifications on the trisyllables. Overall, speakers of both dialects were able to identify the dialects with an average accuracy of 83.8%. As shown in Figure 4, the Changki speakers correctly identified the dialect of the Changki speakers with an accuracy of 87% and that of Mongsen speakers with an accuracy of 85%. On the other hand, the Mongsen speakers correctly identified the Changki dialect with an accuracy of 82% and the Mongsen dialect with an accuracy of 81%. The Changki subjects are slightly better than the Mongsen speakers in dialect identification, however, both dialect speakers have a robust accuracy in identifying the dialects, performing way above the chance level.

3.2. Automatic dialect identification with neutralized F_0

Automatic dialect identification for this study is a binary classification as there are two dialects namely, Changki and Mongsen. Hence, the chance probability for each dialect is 50%. For the experiment conducted by flattening F_0 contours to static F_0 values, it is observed that the accuracy rate for Mongsen dialect is less than the chance probability as shown in Table 1. The low accuracy of identification with pitch information neutralized confirms that absence of tone information did affect the recognition of dialects in Ao.

Table 1: *Accuracy rate for neutralized tone*

Features	Changki Accuracy (%)	Mongsen Accuracy (%)
F_0	57.6	48.5
ΔF_0	53.6	47.0
$\Delta\Delta F_0$	53.7	49.5
F_0 + ΔF_0	55.2	48.6
F_0 + $\Delta\Delta F_0$	59.3	48.2
ΔF_0 + $\Delta\Delta F_0$	52.7	49.0
F_0 + ΔF_0 + $\Delta\Delta F_0$	56.8	48.1

3.3. Automatic dialect identification using tonal and spectral features

Considering the importance of tonal features, as demonstrated in Section 3.2, it was decided to consider tonal features along with traditional MFCC features for dialect identification on Ao dialects. Table 2 shows the results of the dialect identification system, where, tonal and spectral features are introduced separately. When tonal features are used, the combined feature vector of average F_0 and ΔF_0 gives the best accuracy in identifying the Changki dialect. On the other hand, in case of Mongsen dialect, the average F_0 feature gives the best accuracy. Considering the results summarized in Table 2, we combined the tonal features with the MFCC features as shown in Table 3. The α value in Equation 2 were varied from 0 to 1 at an interval of 0.05 and the best alpha accuracy was considered for reporting. As shown in Table 3, it is noticed that by combining all the features, MFCC + F_0 + ΔF_0 + $\Delta\Delta F_0$ we get the best accuracy in Ao dialect identification, with an average accuracy rate of 86.2% across the two dialects. The optimum α value obtained using Equation 2 for Changki dialect was 0.85 and for Mongsen dialect was 0.95.

Table 2: *Dialect identification accuracy rates using tonal features*

Features	Changki Accuracy (%)	Mongsen Accuracy (%)
F_0	64.5	62.3
ΔF_0	51.9	55.4
$\Delta\Delta F_0$	50.8	55.5
F_0 + ΔF_0	66.8	61.5
F_0 + $\Delta\Delta F_0$	63.1	59.5
ΔF_0 + $\Delta\Delta F_0$	53.8	54.7
F_0 + ΔF_0 + $\Delta\Delta F_0$	64.1	60.0

Table 3: *Dialect identification accuracy rates with tonal and spectral features*

Features	Changki Accuracy (%)	Mongsen Accuracy (%)
MFCC	85.5	84.7
MFCC + F_0	86.3	85.3
MFCC + ΔF_0	85.2	84.5
MFCC + $\Delta\Delta F_0$	85.1	84.6
MFCC + F_0 + ΔF_0	86.8	84.6
MFCC + F_0 + $\Delta\Delta F_0$	85.9	85.1
MFCC + ΔF_0 + $\Delta\Delta F_0$	84.9	85.0
MFCC + F_0 + ΔF_0 + $\Delta\Delta F_0$	87.3	85.1

4. Conclusions and Future Work

The results of the perception test reported in this work confirmed that the native speakers of Ao are able to distinguish the two varieties of the language, namely, Changki and Mongsen, based on tonal features. Hence, the dialect identification system proposed in this work was designed to identify two Ao dialects based on spectral and tonal features. While, dialect identification, based only on spectral features yielded satisfactory results (85.1%), the addition of tone information in automatic dialect identification improved the results to 86.2%. In future, we plan to exploit the tonal features further for more robust dialect identification in Ao and other tone languages.

5. Acknowledgements

The authors would like to thank the Ao speakers who voluntarily took part in this study. Also, sincere gratitude to Sarfaraz Jelil of IIT Guwahati for his comments and suggestions.

6. References

[1] 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

[2] 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.

[3] W.-H. Tsai and W.-W. Chang, "Chinese dialect identification using an acoustic-phonotactic model." in *EUROSPEECH*, 1999.

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

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

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

[7] 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.

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

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

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

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

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

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

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

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

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

[17] 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.

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

[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] R. Chitturi and J. H. Hansen, "Multi-stream dialect classification using svm-gmm hybrid classifiers," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on.* IEEE, 2007, pp. 431–436.

[21] 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.

[22] Y. Yan and E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 5. IEEE, 1995, pp. 3511–3514.

[23] L. Sunup Aonok, *A Beginner's Changki-English Dictionary.* Dimapur Mission School, Dimapur, 1994.

[24] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[25] 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, 2009.

[26] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[27] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[28] P. Rose, "Considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech communication*, vol. 6, no. 4, pp. 343–352, 1987.

[29] ——, "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?" *Speech Communication*, vol. 10, no. 3, pp. 229–247, 1991.