PHONOTACTIC AND PROSODIC FEATURES IN NEURAL NETWORK CLASSIFIERS FOR ASSAMESE DIALECT IDENTIFICATION

Abstract

In this study, we investigate phonotactic and prosodic features for dialect identification derived from the voice signal and its transcription. Feedforward Neural Network (FFNN) classifiers are trained to distinguish between dialects using the phonotactic and prosodic features at the trisyllabic level as a representation of dialect characteristics. We show that these traits do indeed include information specific to dialects. We also demonstrate the sufficiency of phonotactic features in terms of broad phonetic categories to express the phonotactic regularities/constraints dialects. For four Assamese dialecta, the effectiveness of the FFNN classifier based on these features is assessed.

Keywords: FFNN; dialect; phonotactic features; prosodic features; Assamese

Author

Hem Chandra Das
Department of Computer Science & Technology
Bodoland University
Kokrajhar, Assam, India
hemchandradas78@gmail.com

I. INTRODUCTION

The task of a machine determining the dialect of speech is known as automatic dialect identification (DID). A multidialectal voice recognizer or dialect translation system can be connected to a DID system as a front end device, allowing the loading of speech recognizers created for that dialect [1, 2]. The construction of a DID system requires an understanding of spoken dialect features. Each dialect uses a finite number of syllables. There is a lot of overlap in the syllable sets since the vocal apparatus used to produce dialects is universal, and there are only a certain amount of syllables in total. However, pronunciations of the same syllable can vary depending on the dialects. Different dialects have very different syllable frequencies. Additionally, there are several phonotactic laws that dictate how various phonemes are put together to make syllables. Different dilects have their own rules regarding the order of syllables. In some dialects, certain phoneme/syllable clusters that are prevalent in some dialects may be uncommon in others. Dialect to dialect variations in word roots and lexicons are common. Every language has a unique vocabulary and way of generating words. The group of words that could come before or after a word in two dialects, even when the word is shared, might be different. Humans have been found to frequently be able to recognize the dialect of a speech even without having a solid grasp of that language's grammar. A listener is likely to rely on lower level restrictions like acoustic-phonetics, phonotactics, and prosody when they lack higher level language understanding. Automatic DID can use any of the following types of information: acoustic-phonetics, prosody, and statistics of sub word units (phonemes or syllables), vocabulary, grammatical and lexical structure, or a combination of these.

The majority of the people of Assam, a state in North East India, and some sections of adjacent states like Meghalaya, Nagaland, and Arunachal Pradesh, speak the Assamese language, which is derived from the Indo-Aryan family of languages. Assamese is the official language of the state of Assam, and it is designated as a Major Indian Language in Schedule-viii of the Indian Constitution. The Assamese language emerged from Sanskrit, but its vocabulary, phonology, and grammar were heavily influenced by the ancient inhabitants of Assam, such as the Bodos and Kacharis [3]. According to recent study, the Assamese language is divided into four dialect groups: The Eastern group, the Central group, the Kamrupi dialect and Goalparia dialect. The Eastern group is spoken in and around the present-day Sibsagar district and its environs, whereas the Central group is spoken in and around the Nagaon district and its environs. Unincorporated Kamrup, Nalbari, Barpeta, Darrang, and a section of Bongaigaon speak the Kamrupi dialect. The Goalparia dialect can be found in Goalpara, Dhubri, and parts of Kokrajhar and Bongaigaon districts.. However, the Central Assamese dialect is now largely regarded as the dominant or standard dialect [4]. There is no major and systematic study of Assamese dialects in the subject of dialect translation. In linguistics, a dialect is a socially different language used by a small number of native speakers who share a common pattern of pronunciation, syntax, and lexicon [5].Human intelligence includes the ability to distinguish between spoken languages [6] [7]. Dialect identification is the first step in developing a dialect-independent voice recognition system for any language. Dialect identification can also help to improve the quality of remote access services like e-health, e-marketing, and e-learning, among others. All dialects share information since they are descended from a single language. Dialect identification is more challenging than language identification since dialects share a lot of reciprocal information.

II. LITERATURE REVIEW

George Wenker conducted a series of studies to define dialect regions in 1877, which started the research of dialect identification [8]. Baily [9] was a pioneer in identifying the Midland dialect as an unique dialect and establishing it as such. The study's findings led to the conclusion that dialects should not be defined only on the basis of vocabulary, because vocabulary can differ greatly within groups or classes within a geographic area. [9]. Davis and Houck [10] also tried to figure out if the Midland dialect region may be called its own dialect region. The researchers were able to successfully extract phonological and lexical characteristics from 11 cities located along a north-south axis [11].

A lot of work has been done in the field of Arabic Dialect Recognition in recent years [12, 13, 14]. Diab et al. [15] and Watson [16] investigated the Arabic dialect, enumerated its characteristics, established a link between the Standard language and regional variations, and categorised the major regional dialects. GMM is used by Ibrahim et al. [17] to detect Arabic dialects. Malaysian Quranic speakers are taken into account by the authors. For this, spectral and prosodic properties were used. When spectral and prosodic traits were combined, they showed a 5.5 to 7% gain in accuracy. For MFCC and prosodic features like pitch, duration, and so on, the accuracy varied from 81.7 to 89.6%.

In numerous Indian languages, dialect recognition research has been done. GMM and HMM were utilised by Shivaprasad and Sadanandam [18][19] to identify regional Telugu dialects. For this goal, the authors created a Telugu dialect dataset. MFCC and its derivatives, such as ΔMFCC and ΔΔMFCC features, were used for recognition. The work uses GMM and HMM models to analyse 39 feature vectors extracted from each spoken utterance. The GMM model performs better than the HMM model. Certain words with the same auditory features, on the other hand, are not differentiated.

Using only prosodic traits and a few lines from each dialect, Chittaragi and Koolagudi [19] employ the closest neighbour approach to identify Telugu dialects. By only examining prosodic features, the authors were able to achieve a 75% accuracy rate.

III. PROPOSED SYSTEM

1. **Speech database:** The review of the current literature reveals that there is no standard database for the Assamese language and its dialects. A new database has been created with speech samples from all the dialect groups. The same number of speakers has been recorded for each dialect region. The speech data consist of speech samples from 10 speakers (5 male and 5 female) representing each dialect regions. In each of these dialects, approximately 3 minutes of audio recordings were acquired from 10 distinct speakers. As a result, each of the four dialects has roughly 30 minutes of voice signal. A phonetically rich script was prepared to record the speech samples. The same script was used to record all the dialects. The recording has been done at 16 KHz sampling frequency, 16-bit mono resolution. Subjective listening test of the recordings has been done using listeners from the respective dialect groups who were not involved in the recording process. Statistical representation of the speech database is given in Table 1.

Table 1: Statistical Representation of the Speech Database

Number of Speakers	**(Five male and Five female) for each dialect group**
Number of sessions	02
Intersession interval	At least one week
Data Types	Speech
Types of Speech	Read speech
Sampling Rate	16 KHz
Sampling format	Mono-channel,16bit resolution
Speech Duration	Each speaker is recording is for minimum 30 minutes in each session.
Microphone	Zoom H4N Portable Voice Recorder microphone
Acoustic Environment	Laboratory
Total duration of speech data	Minimum 1/2 hours for each dialect

2. Formulation of the Problem of Language Identification using Probabilistics

Let $L = \{l_1 + l_2 + \ldots + l_n\}$ denote a string of phonemes or syllables corresponding to any of the dialects in the collection $D = \{D_1 + D_2 \ldots . + D_M\}$. The goal is to determine the input speech's most likely dialect, D, given its n phonemes and syllables. The issue may be stated as follows:

$$D^* = argmax_i P(L/D_i) \quad (1)$$

Assume for the moment that the input vector L is a member of one of M classes D_i, $1 \leq i \leq M$. The primary goal of pattern classification is to determine which class the given vector L belongs in. The issue can be reduced to maximising the joint density $P(L, D_i) = P(L/D_i)P(D_i)$ using the Baye's Rule. Numerous techniques to calculate likelihoods are described in the literature$P(L/D_i)$. The goal is to select the class D* for which the posterior probability $P(D_i/L)$ is largest for a given L, as per the rule stated in (1). This can also be carried out by utilizing

$$D^* = \arg max_i P(L/D_i)\, P(L_i) \quad (2)$$

Where $P(L/D_i)$ is the likelihood probability of S related to dialect D_i and P(D_i) stands for the presumed uniform a priori linguistic probability for all dialects. As a result, the issue is simplified to

$$D^* = \arg max_i P(L/D_i) \quad (3)$$

As a result, the LID problem is now the estimation of the posterior probability as per (1) or the likelihood probability as per (3).

3. Classifiers in neural networks using phonetic features: Using neural network classifiers, the syllable-based characteristics of the dialects can be utilized for dialect identification. When test characteristics are applied as input, the FFN based classifier determines the posterior probability of various dialects.

The sequential constraints at the subword level (how phonemes are joined to produce syllables and syllables are combined to generate words) serve as the property for categorising dialects in the FFN model depicted in Figure 1. For the purposes of syllable coding, it is assumed that each syllable has four components, and each component is assigned a specific code. As a result, four codes will be created, one for each syllable. The lack of a syllable serves as a representation of each word's border. Each syllable's code, along with the one before and one after it, can be normalised between 0 and +1. It can be utilized as input data for neural network-based classifier training, with linguistic identification as an output, as shown in Figure 1. The output for the training dialect is set to + I, while all other outputs are set to 0. Back propagation technique can be used to train the network. The classifier during testing can be fed the trisyllabic code created from the test data and the appropriate output score can be added up for a continuous stream of n syllables. Using the highest overall score at the output, the language's identity will be determined.

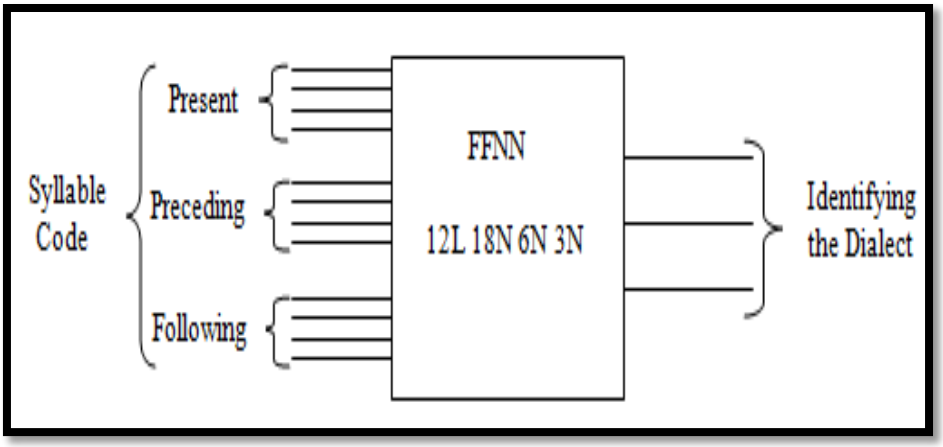

Figure 1: Using Phonotactic Features, Train a Neural Network Classifier to Identify Dialect

Take into account a test sequence $L = \{r_1, r_2, \dots, r_n\}$ that has n trisyllabic units, where $r_{j,}1 \leq j \leq n$ stands for each trisyllabic unit. Then

$$P(D_i/L) = \prod_{j=1}^{n} P\left(D_i/r_j\right) \quad (4)$$

This is the same as adding up the log-likelihood probability

$$\log\ P(D_i/L) = \sum_{j=1}^{n} \log\left(P\left(D_i/r_j\right)\right) \quad (5)$$

Based on the output's highest cumulative log likelihood, the dialect can be identified. The counterpart of this is to identify the dialect that maximizes the log likelihood.

$$D^* = \arg\max \log P(D_i/L) \quad (6)$$

4. **Dialect identification prosodic features:** The aspects of speech known as suprasegmental prosodic features deal with auditory characteristics of sound and go beyond phonemes and syllables.

These are elements that we can utilise in spoken communication without giving them any thought. It is demonstrated that prosodic features outperform cepstral features in their resistance to auditory and environmental mismatches *[20]*.

The prosodic aspect of rhythm is known to provide a lot of information about the identity of a language. The regularity of a particular dialect's grammatical unit's patterns creates rhythm. When producing speech, the articulation is directly tied to the rhythm and CV patterns [21]. In rhythm modeling, the syllable structure is crucial. The syllabic structure can be represented by the number of components in each syllable, $N_t$and the number of components before and after vowel,N_{c1} andN_{c2}, respectively. The length of the syllable is utilized instead of computing the durations of the syllable's consonants and vowels because doing so is difficult and may help to better convey the syllable's rhythmic qualities. A syllable cannot be connected to a typical rhythm on its own. A series of syllables that are strongly tied to specific linguistic characteristics make up the rhythm. Therefore, by integrating the characteristics of the three neighbouring syllables, the rhythm can be roughly approximated at the trisyllabic level.

Normally, intonation is defined as the variation in pitch or fundamental frequency (F_0) over a broader span (bigger than a syllable or as vast as a phrase or sentence). Pitch analysis creates pitch curves that have finer variances. However, intonation is frequently viewed as the contour's inclinations and inflections rather than its subtle fluctuations. There are numerous quantitative measurements that can describe the dialect-specific pitch characteristics, including, the F_0 syllable's range of pitches, the placement of the maximal F_0 in relation to the beginning of the syllable, which is crucial for identifying the accent, the syllable's representative pitch is the average pitch F_0 and the syllable's specific placement inside the word and phrase. These measurements of the previous and next syllables can be used to create a trisyllable level representation, which takes into account regional variations in pitch.

5. **Phonetic and prosodic features used in a neural network classifier:** Over time, humans develop their knowledge of prosody. The mechanism by which this occurs cannot be described or expressed in terms of algorithms. Therefore, it is challenging to express them in a way that a machine can learn. In light of this, rule-based methods to prosody are impracticable. The implicit prosodic qualities of a dialect can be captured using neural networks to help identify that dialect. Figure 2 illustrates how a segmentation and transcription database employing an FFNN may be used to study the function of prosody and phonotactic features in syllable sequence for language identification.

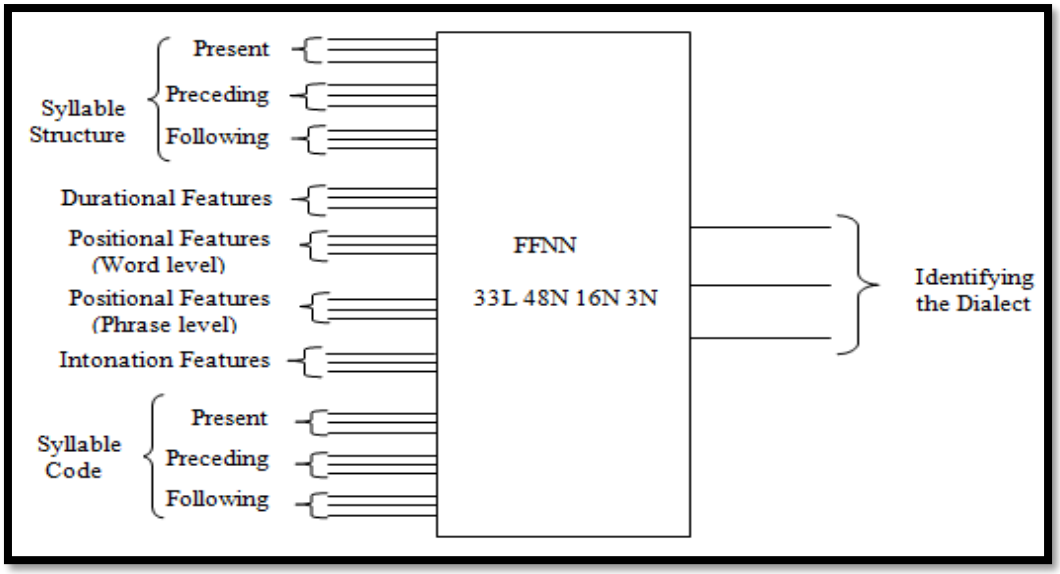

Figure 2: P**honotactic and Prosodic Features are used in a Neural Network Classifier for Dialect Recognition.**

Let's say $A = \{a_1, a_2, \ldots, a_n\}$ a series of feature vectors is used to represent the test speech, where the jth trisyllabic unit's prosodic and phonotactic characteristics are represented by $a_{j,} 1 \leq j \leq n$ .Then

$$\log P\,(D_i/A) = \sum_{j=1}^{n} logP\big(D_i/a_j\big) \qquad (7)$$

At the FFFN classifier's output terminal, the highest cumulative log-likelihood probability can be used to determine the dialect's identification

IV. RESULTS AND DISCUSSION

The database includes the Eastern group, Central group, Kamrupia dialect, and Goalporia dialects. There are various small voice snippets that have been divided up and transcribed in the database; they last about 3 seconds apiece. By human experts, this is further broken down into syllables. The major purpose of selecting this database was to explore the phonotactics of these dialectss without any transcription or segmentation problems. Additionally, since the word boundaries are known, it is possible to model frequently occurring words with less than four syllables using the trisyllabic structure. Approximately 40,000 syllables from each language were used to train the classifier. An average of 600 test instances, each 20 syllables long, and 250 test cases, each 50 syllables long, were used during testing.

1. **Making use of explicit syllable codes:** By explicitly coding the syllables and making the assumption that each syllable has four parts, the input features were taken from the database of transcriptions. A specific code stood in for the absence of any ingredient. The phonotactic regularities were modeled using a trisyllabic framework. Table 2 provides the results of the neural network-based classifier's use of phonotactic features. As part of the rank-based procedure, dialects were ordered according to the classifer output value for

each syllable in the test speech. The duration of the test syllable sequence is measured by the number of first ranks attained.

The accumulation of evidence is done as pr 5 in the second technique. When evidence is accumulated, the DID system is seen to perform better.

Table 2: The Output of the Neural Network-Based Classifier Using Phonotactic Characteristics

Methods	**20 syllables**		**50 syllables**	
Dialects	Rank-based	Accumulation	Rank-based	Accumulation
Eastern Dialect	99.4%	98.5%	100%	100%
Central Dialect	96.2%	97.5%	100%	100%
Kamrupia	72.4%	85.8%	92.5%	98.6%
Goalporia	72.4%	85.8%	83.8%	95.6%

2. **Making use of syllable codes for broad category:** By substantially segmenting the syllables in the transcription database, the input features were collected, and they were subsequently coded. Vowels, nasals, semivowels, fricatives, unvoiced unaspirated stops, unvoiced aspirated stops, and voiced aspirated stops were the basic phonetic categories used to categorise the components of syllables. Syllables were coded according to this wide classification to acquire the input features. On the basis of how they were articulated, the stop consonants were categorized. The phonotactic regularities were modeled using a trisyllabic structure. Table 3 displays the results of the NN-based classifier for DID utilizing the broad phonotactic characteristics. According to data derived from broad phonotactic features, categorizing syllables in respect of broad categories is sufficient to describe the phonotacic restrictions of dialects. As a result, neither the neural network's testing nor its training will require accurately transcribed speech. The broad phoneme categories used to categorize the components of the syllables should be optimized for each dialect in the identification task.

Table 3: Results of the NN-Based Classifier for DID Utilizing the Broad Phonotactic Characteristics

Methods	**20 syllables**		**50 syllables**	
Dialects	Rank-based	Accumulation	Rank-based	Accumulation
Eastern Dialect	99.8%	98.55%	100%	100%
Central Dialect	96.2%	98.5%	88.71%	97.52%
Kamrupia	77.4%	85.8%	83.8%	95.6%
Goalporia	35.4%	85.8%	55.8%	95.6%

3. **Making use of phonetic and prosodic features:** The structure of a syllable and its duration serve as a representation of dialect's rhythm. The average syllabic pitch, pitch range, and maximum pitch location are all indicators of intonation. A trisyllabic structure is used as the fundamental unit because the data relating to a single syllable alone are insufficient to capture the prosodic pattern. By combining the characteristics of the present syllable with those of the previous and next syllables, the trisyllabic structure is created. The classifier was trained using characteristics derived from roughly 25,000

syllables for each language. The FFFN-based classifier is trained using prosodic data as well as phonotactic features in terms of explicit syllable codes, and it outputs dialect identification. Table 3's findings demonstrate that even while training with fewer samples, the classifier performs better when prosodic features are added in addition to phonotactic variables.

Table 4: The Results Of the DID System NN-Classifier Based Phonotactic Characteristics and Prosody Feature

Methods	**20 syllables**		**50 syllables**	
Dialects	Rank-based	Accumulation	Rank-based	Accumulation
Eastern Dialect	96.8%	97.45%	99.66%	100%
Central Dialect	98.50%	100%	100%	100%
Kamrupia	77.4%	85.8%	97.8%	100%
Goalporia	67.3%	83.8%	95.8%	100%

V. CONCLUSIONS

Through the use of phonotactic and prosodic features, we have demonstrated in this research that neural network-based classifiers are capable of conducting language identification. The phonotactic regularities and limitations of dialects, as well as dialect-specific prosody given in terms of numerical measures, allow the FFNN classifiers to identify between dialects. It was also attempted to represent the phonotactic regularities of dialects by using phonotactic elements in terms of broad phonetic categories. More dialects of Assamese language could be included in this study. The next step is to create a system that can recognize dialects by directly deriving features from speech signals without the aid of syllable transcription.

REFERENCE

[1] M. A. ZISSMAN AND K. M. BERKLING, "AUTOMATIC LANGUAGE IDENTIFICATION," SPEECH COMMUNICATION VOL. 35, NO.
[2] 1-2, PP. 115--124, 2001. WAIBEL, P. GEUTNER, L. M. TOMOKIYO, T. SCHULTZ AND M. WOSZCZYNA, "MULTILINGUALITY IN SPEECH AND SPOKEN
[3] LANGUAGE SYSTEMS," PROCEEDINGS OF THE IEEE, VOL. 88, NO. 8, PP. 1297--1313, 2000.
[4] S. JOTHILAKSHMI, V. RAMALINGAM AND S. PALANIVEL, "A HIERARCHICAL LANGUAGE IDENTIFICATION SYSTEM FOR INDIAN
[5] LANGUAGES," DIGITAL SIGNAL PROCESSING, VOL. 22, NO. 3, PP. 544--553, 2012.
[6] M. SHARMA AND K. K. SARMA, "LEARNING AIDED MOOD AND DIALECT RECOGNITION USING TELEPHONIC SPEECH," IN 2016
[7] INTERNATIONAL CONFERENCE ON ACCESSIBILITY TO DIGITAL WORLD (ICADW), 2016.
[8] WIKIPEDIA, "ASSAMESE LANGUAGE," 10 OCT 2019. [ONLINE]. AVAILABLE: HTTPS://EN.WIKIPEDIA.ORG/WIKI/ASSAMESE_LANGUAGE.
[9] G. A. LIU AND J. H. L. HANSEN, "A SYSTEMATIC STRATEGY FOR ROBUST AUTOMATIC DIALECT IDENTIFICATION," IN 2011 19TH
[10] EUROPEAN SIGNAL PROCESSING CONFERENCE, BARCELONA, SPAIN, PP. 2138-2141, IEEE, 2011.
[11] H. LI, B. MA AND K. A. LEE, "SPOKEN LANGUAGE RECOGNITION:," IN PROCEEDINGS OF THE IEEE, VOL. 101, PP. 1136-1159, 2013.

[12] A. NTI, "STUDYING DIALECTS TO UNDERSTAND HUMAN LANGUAGE," M.E. DISSERTATION, DEPT. OF ELECT. ENG. AND
[13] COMPUTER SCIENCE, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, MASSACHUSETTS, ME, 2009.
[14] BAILEY AND C. J. N, "IS THERE A "MIDLAND" DIALECT OF AMERICAN ENGLISH?," ERIC CLEARINGHOUSE, 1968.
[15] L. M. DAVIS AND C. L. HOUCK, "IS THERE A MIDLAND DIALECT AREA?—AGAIN," AMERICAN SPEECH, VOL. 67, NO. 1, PP. 61-70, 1992.
[16] ETMAN AND A. L. BEEX, "LANGUAGE AND DIALECT IDENTIFICATION: A SURVEY," IN 2015 IEEE SAI INTELLIGENT SYSTEMS CONFERENCE (INTELLISYS), LONDON, UK, 2015.
[17] SHOUFAN AND S. AL-AMERI, "NATURAL LANGUAGE PROCESSING FOR DIALECTICAL ARABIC: A SURVEY," IN PROCEEDINGS
[18] OF THE SECOND WORKSHOP ON ARABIC NATURAL LANGUAGE PROCESSING, BEIJING, CHINA, PP. 36-48, 2015.
[19] I. GUELLIL, H. SAÂDANE, F. AZOUAOU, B. GUENI AND D. NOUVEL, "ARABIC NATURAL LANGUAGE PROCESSING: AN OVERVIEW," JOURNAL OF KING SAUD UNIVERSITY - COMPUTER AND INFORMATION SCIENCES, VOL. 33, NO. 5, PP. 497-507, 2021. [14] A. ELINAGAR, S. M. YAGI, A. B. NASSIF, I. SHAHIN AND S. A. SALLOUM, "SYSTEMATIC LITERATURE REVIEW OF DIALECTAL ARABIC:," IEEE ACCESS, VOL. 9, PP. 31010-31042, 2021.
[20] M. DIAB AND N. HABASH, "ARABIC DIALECT PROCESSING TUTORIAL," IN IN PROCEEDINGS OF THE HUMAN LANGUAGE TECHNOLOGY CONFERENCE OF THE NAACL, COMPANION VOLUME: TUTORIAL ABSTRACTS, NEW YORK, 2007.
[21] J. C. WATSON, "50. ARABIC DIALECTS (GENERAL ARTICLE)," IN THE SEMITIC LANGUAGES: AN INTERNATIONAL HANDBOOK, BERLIN, WALTER DE GRUYTER, 2011, PP. 841-896.
[22] N. J. IBRAHIM, M. Y. I. IDRIS, M. YAKUB, N. N. A. RAHMAN AND M. I. DIEN, "ROBUST FEATURE EXTRACTION BASED ON SPECTRAL AND PROSODIC FEATURES FOR CLASSICAL ARABIC ACCENTS RECOGNITION," MALAYSIAN JOURNAL OF COMPUTER SCIENCE, NO. 3, PP. 46-72, 2019.
[23] S. SHIVAPRASAD AND M. SADANANDAM, "IDENTIFICATION OF REGIONAL DIALECTS OF TELUGU LANGUAGE USING TEXT INDEPENDENT SPEECH PROCESSING MODELS," INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY, PP. 251-258, 2020.
[24] N. B. CHITTARAGI AND S. G. KOOLAGUDI, "ACOUSTIC FEATURES BASED WORD LEVEL DIALECT CLASSIFICATION USING SVM AND ENSEMBLE METHODS," IN 2017 TENTH INTERNATIONAL CONFERENCE ON CONTEMPORARY COMPUTING (IC3), NOIDA, INDIA, 2017.
[25] E. THYME-GOBBEL AND S. E. HUTCHINS, "ON USING PROSODIC CUES IN AUTOMATIC LANGUAGE IDENTIFICATION," IN PROC. ICSLP, 1996.
[26] J. L. ROUAS, J. FARINAS, F. PELLEGRINO AND R. ANDRE-OBRECHT, "MODELING PROSODY FOR LANGUAGE IDENTIFICATION ON READ AND SPONTANEOUS SPEECH," IN 2003 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2003

