Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10203386

DIALECTAL VARIABILITY IN SPOKEN LANGUAGE: A COMPREHENSIVE SURVEY OF MODERN TECHNIQUES FOR LANGUAGE DIALECT IDENTIFICATION IN SPEECH SIGNALS

Er. POONAM KUKANA
Department of Computer Science and Engineering, University School of Engineering & Technology, Rayat Bahra University, Mohali, Punjab, India. Email: poonamkukana@gmail.com

Er. SUKHJINDER KAUR
Department of Computer Science and Engineering, University School of Engineering & Technology, Rayat Bahra University, Mohali, Punjab, India. Email: skaur29100@gmail.com

Dr. PUNEET SAPRA
Department of Computer Science and Engineering, University School of Engineering & Technology, Rayat Bahra University, Mohali, Punjab, India. Email: puneetsapra91@gmail.com

Er. CHIMAN SAINI
Department of Computer Science and Engineering, University School of Engineering & Technology, Rayat Bahra University, Mohali, Punjab, India. Email: chimansaini1994@gmail.com

Abstract

Main fundamental challenge for recent research work on speech based on science and technology is to understand and model the user variants in Spoken Languages. Users have their style of speaking, reliant on various factors, adding the dialect and accent of the speaker as well as the social and economic background of the speaker and contextual attributes like degree of knowledge between the listener, speaker and the position or rank of the speaking condition, from very normal to formal. In the past few decades, an extensive progress has been seen in automatically verifying the language of a speaker offered a sample speech. The main purpose of dialect verification is the recognition of a speaker's region dialect, within a pre-determined language, offered the acoustic signal alone. DR (Dialect Recognition) is a main issue in particular, since even within the similar dialect and accent or register user change may occur. For illustration, In Spontaneous speech, few speakers tend to exhibit more optimizing and alteration of function words than others. The main issue of dialect recognition system has been viewed as challenging than that of language classification or recognition due to the maximum similarity among dialects of the similar language. While, dialects may differ in any dimensions of the linguistic spectrum such as syntactic, lexical, morphological, phonological differences, these changes are likely to be more indirect across dialects than those across languages such as Hindi, Punjabi and English etc.

Keywords: DR (Dialect Recognition), DI (Dialect Identification) ASR (Automatic Speech Recognition), MFCC (Mel Frequency Cepstral coefficient), Linear Discriminant analysis (LDA), (LPC) Linear Prediction Coefficient, Probable Linear Discriminate Analysis (PLDA), Relative spectra (RASTA) filtering, FFMP (feed forward multilayer perceptron), CNN (Convolution Neural Network), RNN (Recurrent Neural Network).

I. INTRODUCTION

Speech is the method of expressing the ideas, expressions or thoughts. Speech is used in the daily conversation. It is the kind of the conversation [1]. Speech plays an essential role in number of actions include socialization. Speech or Language is related to the method of communication used by social beings [2]. It is the method of expressing the ideas and thoughts through vocal sounds. A dialect is mainly a specific kind of the language that specified to area or the human set [3]. It mainly has variations in pronunciation, grammar, syntax, vocabulary. It is not easy to understand because dialect is vocal by living people in specific location [4]. A speech is dialect with the defense forces. The variation among the language and dialect may arrange the variation among speech and dialect. The selection is known as speech that is connected for the identification of edges of the countries, and person determines the speech and following consideration [5].

Speech is one of the essential equipment for the communication among the social being and surroundings. Consequently, an automatic recognition system is done for need of the humans. Speech recognition system is used as distinct and regular models that are reliant or non-reliant on humans [6]. Different methods manage a distinct acoustic method for every word, association of the words that is related to different word speech recognition. A speaker-based network needs that the user traces an instance of the word, sentence or phrase previous to detection by the network [7].

The speech analysis is done through the speech input. Then the feature vector values are determined. After that, the data processing is done from the selected feature vector. The resulted speech recognition is achieved from the reference model. The process of the automatic recognition system is described as;

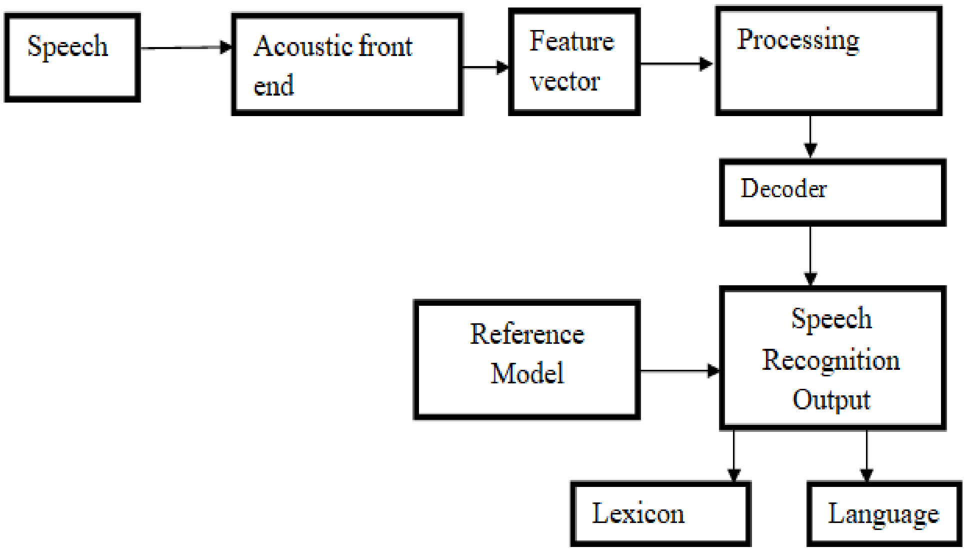

Figure 1: Speech Recognition System [7]

It is done in training and recognition stage. During the training stage, the pre-processing and recording of the speech is done where the feature extraction take place. In other phase, acoustic study of the indefinite signal is done in recognition stage.

1. Feature Extraction: In this stage, noise robust evaluation is probable on the high correlation coefficients. However, the minimum value of the noise speech signal autocorrelation is related to the elimination of the major noise elements [8].
2. Acoustic Model: It is the main element of the automatic recognition system that is responsible for the high load and performance of the network. It is established for identifying the vocal fact. The establishment includes the usage of the audio database of the speech and the text of the scripts. After that, compile the data into statistic demonstration of speech that construct the words [9].
3. Lexical Method: The accent of every word in required language is provided where lexicon is established. The different linking of the phones is determined by lexicon method for providing the suitable words.
4. Language Approach: It is the main element operates on large number if the words, comprising of large metrics and through accent vocabulary, created words series in sentence.

II. DIALECT IDENTIFICATION IN THE LITERATURE

A. From a linguistic point of view

The most difficult part of speech recognition is dialect, which refers to the language features of a given geographical population. The lack of databases and the time-consuming analytical method continue to hamper research in dialect linguistics. As technology has improved, a robust speech recognizer capable of dealing with unstable settings such as noise and accent fluctuatsion has become an imperative requirement.

Dialects are not fixed; they fluctuate with location and time. In other words, dialects may differ owing to language changes throughout time, just as they do with physical borders. Today, we find that new generations employ terms and language that were not previously available or used. As a result, languages evolve with time, as do their dialects. Furthermore, each language has multiple dialects. The United States, for example, has one dialect border that runs between the North and South along what is known as the Mason-Dixon Line. American English, on the other hand, has a variety of dialect barrier. There are several dialects in each of the fifty states, and each state may have a multiple dialect.

1) DI Early Studies

Dialect research began in 1877, when George Wenker conducted a series of surveys to define dialect areas Then Bailey made one of the initial attempts to describe the Midland dialect and determine whether or not it exists. The study concluded that distinguishing dialects should not be based on vocabulary because it varies depending on group or class within the same geographical region. Davis and Houck were also attempting to determine if the Midland area may be considered as a separate dialect region or not. Their research was effective in extracting phonological and lexical data from 11 cities along a north-south axis. The result was that the Midlands area cannot be called an intermediate region and that there's a linear link between the dialects of the far South and the dialects of the

Southern region. In contrast to the latter idea, Johnson demonstrated that integrating phonological and lexical variables is inaccurate since it negatively impacts data patterns and hence generates incorrect findings. Some terms were used to demonstrate that there is a distinct difference in the dialects of the North and the Midlands, as well as the dialects of the South and the Midlands, but not among the dialects of the North and the South. The result was that the Midland area is distinct from both the North and South regions. These were the initial steps in recognizing and categorizing dialects. The second stage was to establish how much dialects correspond to each other after defining them. What are the similarities and differences across dialects of an identical language, and via what method can they be evaluated?

2) DI using Vowels

Peterson and Barney made one of the earliest efforts, focusing on the vowel spacing feature. They discovered that perceptually distinct vowels occupy various areas in formant space; also, the identical vowel produced by different persons appears in varied spots in formant space. Finally, it was determined that participants' capacity to create and interpret a particular vowel is influenced by their history. This work was noteworthy in the field of recognition since it was the very first to emphasize the significance of dialects. The one disadvantage of this investigation was the variability of individuals, which made proper characterization of vowel spacing alterations difficult. As a result, Hillenbrand, Clark, and Wheeler performed the same tests to achieve more precise findings for vowel gap changes using a single homogenous class. More vowels and diphthongs have been introduced to those covered in Peterson and Barney's study.

In addition, for each vowel, the spectral shifts, its duration, which stem and stable state F1, F2 were measured. All of these efforts were made in order to deliver more accurate and up-to-date data. Michigan was represented by the contestants. According to the data, the individuals had identical vowel locations but with less vowel spacing in the vowel space. The study concluded that lowering the distance between vowels had no effect on their perception. Furthermore, our findings corroborated prior findings that F1 and F2 measures are insufficient to characterize vowel spacing.

Hajiwara duplicated the findings of Peterson, Barney, Clark, and Wheeler. The contestants this time were from Southern California. Hajiwara was on the lookout for formant modifications in this new dialect. Hajiwara's research discovered that a common dialect like Southern California seldom generates a fully rounded vowel, which was the major cause for having greater F2 for some vowels. This study discovered a new dialect (South Californian) that may be compared to existing dialects. Furthermore, Hajiwara thought that further study for different dialects within a single group should be done in the future by expanding the number of of both genders' participants.

3) DI using Consonants

Consonants, on the other hand, have been identified as dialect informational identifiers due to their ability to disclose accents from other countries and social status. This prompted William Labov to conduct sociolinguistic research on accent variations. As a new measuring metric that indicates personal origin, he employed rhoticity: the sound of 'r' when it occurs after a vowel (as in bar, sort, churn) and is known as post- vocalic 'r'.

Rhoticity was seen as a low-status trait in the United Kingdom, but it was regarded as an important characteristic of pronouncing in the United States. This made distinguishing between American and British English simple. As a result, Labov conducted the first sociolinguistic study, demonstrating that New York accent differences center around postvocalic "r" use. He began with a brief poll to assess the dependability of the fresh testing procedures. Labov did not interview speakers as previous studies had done; instead, he wandered through three Manhattan department shops pretending to be a consumer. He began by checking what things were on the 4th floor of each store, and then he asking the sales associates where he could find these items. In each store, he repeated the experiment. He chose the fourth level because it had two tokens of postvocalic "r," and by appearing not to hear, he got every participant to say the two syllables a second time, once impulsively and once attentively, resulting in four tokens from each informant. The collected data revealed that Labov's theory was correct, and that the postvocalic "r" fluctuates depending on social class, speaking style, and linguistic context connected with each store's customers. Higher and lower-ranking employees were readily identified in this study by evaluating the post-vocalic "r" pronunciation. Labov demonstrated that the rhotic usage of the letter "r" signified social status and aspiration from the as well as that it was more prevalent in younger speakers.

According to Labov's research, the New York accent changed towards the end of WWII. Furthermore, varieties of American English lack the sound of the post-vocalic "r". This approach was a simple and effective way to detect American accents. Through sociolinguistic investigations, Labov was the first to develop new measurement methods for identifying dialects.

To summaries, the past studies were experiments to assign the basic building pieces for Dialect Identification (DI) research. The research was largely effective in demonstrating that vowel gap and consonants are both crucial factors in categorizing dialects. Following the discovery that acoustic properties varied across dialects, researchers conducted investigations to uncover how these differences differ between regional dialects as well as between various genders and ages. This prompted Byrd to investigate variances in the TIMIT corpora while categories utterances based on area and gender. TIMIT contains 630 female and male speakers reflecting all eight American dialects. The study sought to identify gender variations in glottalization, number of the flaps, palatalization, central vowels, and speaking pace. The results revealed that the release stop, glottal stop, and vowel contraction rates varied amongst dialects, but the number of flaps did not. Furthermore, men tend to talk quicker than women, with fewer stressed vowels. However, there was no substantial difference in dialects, and it was suggested that this might be due to a lack of data for both genders. This demonstrates that there are several parameters/features that might influence dialect variations. It is not a simple process to extract these parameters. Furthermore, because this is a database-dependent problem, it is hard to pinpoint the specific characteristic that is that causes a particular dialect shift. A useful database must include individuals of all ages, both genders, and various affiliations. This useful database, however, is still missing other hidden characteristics that change with dialect, like moods and physical state.

4) DI using Acoustic and Phonetic Features

Clopper acknowledged and recognized six American dialects in some successful attempts. The acoustic aspects of a dialect were researched initially, followed by the perceptual features of listeners. In this paper, several words that differ between dialects were chosen and their acoustic-phonetic features were investigated. Before proceeding with the overview of both articles, it is useful establishing certain key parameters. Fricatives are consonants generated by pushing air through a small channel formed by the close proximity of two articulators. In the instance of [f], the articulators are the lower lip and the upper teeth. Frication refers to turbulent wind. Another crucial criterion is vowel back-ness, which refers to the location of the tongue during vowel articulation in relation to the back of the mouth. However, vowels are classified as rear or front based on the frequency of the second formant (F2) rather than the actual articulation. The higher the F2 value, the closer the vowel is to the front; the lower the F2 value, the further back the vowel is. Di- pthongization, also known as vowel breaking, is the transformation of a mono-phthong into a di-phthong or tri-phthong. Di-phthongization refers to the transformation into a di-phthong. Nagy and Zhang were interested in novel metrics such as r- fullness to distinguish between rhotic and non-rhotic dialects, vowel brightness to check for "r" insertion in a word, and fricative voicing and length to determine if individuals pronounce words with fricatives. Experiments were conducted after evaluating and determining the above listed factors to determine which qualities listeners utilize for recognizing a dialect. According to the findings of these investigations, listeners usually employ four characteristics in perceptions and dialect identification. R-fullness, blackness, vowel brightness, and di-pthongs were the four criteria. Listeners were also able to distinguish the dialects of the South and New England from others. The capacity of an individual listening to recognize a dialect is determined by where he or she has lived throughout his or her life.

Later, in another study, the goal was to provide a thorough characterization of major American English dialects utilizing the previously extracted acoustic-phonetic elements. The unexpected finding here was the variation of traits across the same dialect. It is hard to generalize characteristic variation on a specific dialect group. Unexpected characteristics among Southern males included fronting and rising of the "u," as well as merged vowels on occasion. Finally, it was concluded that there exist phonological distinctions across dialects and that we can determine a dialect based on well-chosen groupings of strongly correlated lexical sets. Human perceptions of dialects use the same processes, but occasionally fails owing to dialect variance within a single group.

5) DI using Words and Lexical Sets

Wells chose several terms and defined them as sets with an unchanging pronunciation pattern, so supporting the same premise that certain words are ideal dialect identifiers. A set of keywords was created, each with its own lexical set. A lexical set is a group of words that all have the same vowel pronunciation. During studies, it was discovered that two lexical sets blended inside an accent. As a result, the finding was that a dialect may be distinguished using certain mergers and lexical set pronunciation.

Zhang and Nagy investigated the reciprocal information between distinct lexical sets in order to classify dialects based on phonological characteristics. This approach was evaluated on 168 binary parameters defining the pronunciation of English vowels and consonants for speakers from 35 different nations. This approach yielded clusters that were comparable to those created by typical clustering methods. A comparison of clustering approaches was also performed, and the variations in clustering effects were investigated. This study was crucial in giving a succinct description for several English dialects, as it identified all lexical set variances.

Kessler's approach of grouping by analyzing phonetic transcription was another. The approach, like the isoglosse method, was based on determining linguistic distance between two places. Isoglosses were the usual way for determining dialect borders. This was accomplished by drawing border lines between locations with a group dialect; individuals who pronounce the same manner. The disadvantage of isoglosses is the inconsistency of border lines caused by lexical feature change within the same area. The baseline metric and the Levenshtien metric were the new distance metrics used to aggregate phonetic measures. In the baseline measure, two locations on the map had "0" distance for phone strings that were similarly comparable, and "1" distance otherwise. The Levenshtien range between phonetic string was measured in the Levenshtien metric, where the Levenshtien range is defined as the most inexpensive sequence of deletions, insertions, and substitutions required to transform one phone string into another. The phonetic and features string comparisons were utilised; in the former, each operation turns one phone string into another phone string, when in the latter, each operation changes the characteristic feature. The distance was calculated by averaging the distances between several landmarks. The ultimate finding was that the phone string is superior for automated dialect identification grouping. Furthermore, the term was defined as an ADI unit that may be separated into string to get exact measurements.

The last DI approach is that supplied by Huckvale, who proposed ACCDIST, a new measure for the quantitative evaluation of a speaker's accent similarities. The idea behind ACCDIST is that the author was able to group together accents of the speakers by storing the inter-segment distance measurements for each speaker in a table and then detecting autocorrelation measurements between the data collected and the ones under test. The ACCDIST speaker classification approach captures a speaker's pronunciation qualities rather than voice parameters.

III. SPEECH DIALECT SYSTEM

In speech dialect recognition scheme, the recognition of the dialect or accent is done through the training speech utterance of special dialect utilizing the single phone detector. Speech recognition is the capability of the program to recognize the words or phrase in spoken language and transform them machine readable format. In the given figure, the speech input is analyzed for the selection of the template. After that the features are extracted where the decoded procedure is done through the different models. The models are acoustic, linguistic or pronunciation. The final output value is achieved or pronunciation. The final output value is achieved.

The main types of the speech dialect system are described as;

1. Acoustic Phonetic Approach: In this method, the different phonetic unit in vocal speech and the speech is measured through the group of the acoustic features. The features of the phonetic elements ate mainly changeable with speaker and closest sound. It is unspoken in this method that regulates the leading the changeability are simple and educated through equipment [10]. The initial stage of this method is language spectral study. In the other stage, the conversion of the spectral dimensions in to group of the feature is done during feature extraction that explained the acoustic features of various phonetic units. The other stage is the division and marking phase where the language signal is divided in to remote areas follow by linking single or more markers to every divided area [11]. The final phase searches the suitable word from phonetic mark series created by segmentation to marking. The input the speech signal. After that, speech signal pre-processing is experienced. And then the recognition of the phonetic unit is done by the acoustic modeling. The decoded data is practiced.

2. Pattern Recognition Approach: This method is used for the extraction of the pattern dependent on desired situations and to divide single class from other. It is based on four phases as feature dimension, training and classification the pattern. The test pattern is defined, a series of the dimensions is completed on inner signal [12]. The related pattern is developed by acquiring more than pattern related language sounds. It can be given in the speech pattern. It is analyzed on phrase, speech or word. During the pattern classification phase of this method, a straight comparison is done among the undesired test pattern. Input the speech signal and features are analyzed for the classification of the pattern along with language, word lexicon and acoustic model. The utterance is verified to receive the final value.

3. Template Method: In this technique, the undesired language is comparable opposite to group of templates for searching desired match [13]. The group of the prototype language pattern are recorded as location pattern demonstrating the vocabulary of words of the applicant. The recognition are done by matching of undesired language word along with every of the related template and choosing the class of suitable matching pattern. This technique presents the high recognition rate. Input the speech signal and the pre-processing and feature extraction is done. Moreover, the template and pattern classification are done by testing. In addition, acoustic and language are determined by to get the speech transcript through data classification.

4. Artificial Intelligence Method: It is mixture of the acoustic phonetic method and pattern detection system. A number of the investigators established detection method through acoustic phonetic data to establish the classification regulation for language sounds. Despite the fact that the template dependent technique presents the small relation about the social speech processing, but these methods are more efficient in design of the language recognition methods. In contrast, the language and phonetic

survey of social language processing is done. On other hand, this method has one sided achievement because of the complexity to quantify the experience data [14].

IV. INDIAN LANGUAGES

About languages and number's (Indian constitution recognizes):

Language is normal method and it may not be essential to get direct boundaries that were simple for the definition and identification. Language is related to communal, educational factors instead of intrinsic real and realistic computation of language characteristics of speech [15].

The Indian constitution determines the eighteen representative Indian speech or language. These languages include various dialect or the changes of that speech. In addition, the regional languages are considered by central government and worked as official language. The state limits of the state generated was dependent on restrictions of the major Indian language that is considered by Indian Constitution.

Dialect is the kind of the speech that is vocal in one region which varies from another kind of the similar speech (words). It demonstrated the characteristics of the syntax and terminology and aspect of accent. It is the kind of the language that is vocal in single region which may change from other kind of similar speech. There is no universal criterion for differentiating and distinction is subject of point. It is not easy to present direct demonstration for speech and dialect [16].

The language when utilized by the people from various areas may be analyzed to check the selection of the words [17]. If some standard forms of the words are spoken then the variation in the spectral features of sound created. Every person creates a method of speaking at the initial age. The method is based on the native language. The standard form of the language is spoken through native language, and then the speaker acquires the traits of the method. The mode may be affected by the human language or region surrounding of the speaker. The speaking variations worked as accent. Various accents provide the variation in the understanding of the utterance. During the acoustic training, the effectiveness of the automated system recognition has been improved by dialect.

V. FEATURE EXTRACTION TECHNIQUES

Generally, the speech feature extraction is utilized for decreasing the size of the input vector, whereas the management of the sensitive energy of the signal. Feature extraction is the specific type of the database and extraction of the specified features are acquired. The features received structure of the desired data about the speech [18]. The major kinds of the feature extraction group are MFCC (Mel Frequency Cepstral coefficient), Linear Discriminant analysis (LDA), (LPC) Linear Prediction Coefficient. However, MFCC is the major method for feature extraction. Feature extraction is the major kind of the speech recognition scheme. It is measured as the major part of the network.

The main speech signal is forwarded to the per emphasis stage. In this stage, the decreasing of the maximum frequency region of the acoustic was muted in the social audio development method [19]. The speech after the pre-emphasis sounds deeper with less volume rate. The speech signal is forwarded to maximum arrangement filter. In the frame blocking, an audio digital signal is segmented in to the amount of the frames and few portions of the frames overlapped to each other. Every frame may be estimated without the loss of the data. In the given table 1, types of feature extraction techniques, characteristics, advantages and disadvantages are given.

Table 1: Advantages and Disadvantages of Feature Extraction Techniques.

Feature Extraction Method	Characteristics	Advantages	Disadvantages
Linear Predictive Coding (LPC)	Regression based speech Feature	Accurate and Robust Method. Encode speech at less bit rate	Does not differentiate words with the same vowels.
MEL Frequency Cestrum (MFCC)	Used in speech Processing	Provide high accuracy. MFCC presents the major features of the phone of speech.	The performance is influenced by the amount of the filters.
Probable Linear Discriminate Analysis (PLDA)	Used the variables based on HMM.	It is flexible acoustic method that used variable amount of the related frames.	Class discrimination may be inaccurate.
Relative spectra (RASTA) filtering	It is used in the noisy speech signals.	It acquires the frequency with less modulation related to speech.	An error occurs during the filtering.

Feature selection is the stage when an automatic or the manual selection of the features that took place in prediction of the variable or the desired outcome. The unrelated features reduce the accuracy models. Feature selection is the necessary stage in the establishment of the network for the speech recognition.

Presently, the interference among the features created from single audio resource is rarely measured, that creates the redundant features and improves the calculation price [20]. The selection of the features used for the automated speech recognition influenced the performance rate. The features are extracted through the input data. After that, the feature set re arranged with the optimum feature subset. In addition, the speech features are classified as word, phrase, language or phenomena.

Advantages:

1. Less redundant information with minimum opportunity for the selection based on the noise and it decreases over fitting.
2. Improved accuracy with minimum misleading information.
3. Reduced training time where the minimum data trained at faster rate.
4. Removing the unrelated and redundant features, the performance of the prediction model can be improved.
5. The improvement through the elimination of the effect of dimensionality.
6. Improved the general performance rate.

7. Improved the learning process.

VI. DEEP LEARNING MODELS

Deep Learning is the method as the subgroup of the machine learning that that is the internal subgroup of the artificial intelligence. Over the past few decades, deep learning is the technique of the learning information demonstrations. It may be supervised, unsupervised or semi supervised. It presents the group of the algorithms and methods that learns the features and jobs directly from the information [21]. The information may be structured unstructured that includes the pictures, textual data or the sound. DL is related as the point-to-point learning for the direct demonstration of the information. In addition, the deep learning methods worked in absence of the human intervention and it supports an accurate outcome. DL is mostly utilized in regions such as computerized vision, natural language processing, pattern and object detection. Generally, the deep learning methods processed the data in same way to the human brain that may acquire to multiple job by the people. It is mainly used in image pre-processing, normal language processing and speech recognition.

The deep learning architecture followed the approach as,

- Build network containing the input and hidden layer along with the required hops.
- Trained the system.
- Addition of the other hidden layer at the higher level of the learned system for creating a new system.
- Retrain the system.
- Addition of maximum layers and retrains the system.

Advantages of DL:

1. Higher level features are learned.
2. Identify direct relations.
3. High measurement information.
4. Unlabelled information

Some of the methods of the deep learning models are described as [22];

1. **Auto-Encoder:** It is an artificial neural network that is helpful in different coding patterns. The normal type of the auto encoder is same as the multiple perceptron comprising an input layer or hidden layer. The major variation among the multiple layer perceptron, FFNN and auto-encoder is the amount of the hops at the external layers. The external layer comprises the similar kinds of the hops as internal layer.
2. **Deep Belief Network:** It is the output to the issue of management of the non-convex selective values and local minima through specific multiplayer perceptron. It is the kind of the deep learning with multiple layers of the latest variables with the interconnection among the layers. This network may be displayed as the Boltzmann

equipment when every hidden layer of the sub system works as observable internal layer for the adjoining layer of the system.

3. **Convolution Neural Network:** It is the other alternative of the FFMP (feed forward multilayer perceptron). It is kind of the FFNN in which single neuron are arranged for complete overlapping areas. CNN works as successive model of data and linking in the system. The boundaries of the initial layer are identified and create the templates for the edge recognition. After that, succeeding layer tried to link in easy design and templates of the diverse locations. The last layer matched as input picture with complete templates and last possibility worked as addition of the weighted. Hence, the deep CNN method provides high exact predictions.

4. **Recurrent Neural Network:** It worked as the permanent amount of the input value, creates as static dimension vectors as output value along already defined stages. This network permits for operating the series of the vectors in input and output. The interconnection among the units creates as directed cycle in RNN. In this network, the input and output values are not automated but interrelated. In addition, RNN connects the normal metrics at each layer. RNN trained in traditional neural system through back propagation technique.

5. **Re-enforcement Learning to NN:** This network is the type of the hybridization of the dynamic programming and supervised learning. The distinctive element of the model is background, agent, measures, guidelines and cost values. The agents work as the regulator of the network; guidelines recognize the operations, and output value determines the selective of the re-enforcement learning issue. This learning provides the performance with maximum output for the desired stage.

VII. CONCLUSION AND FUTURE SCOPE

The main issue of dialect recognition system has been viewed as challenging than that of language classification or recognition due to the maximum similarity among dialects of the similar language. While, dialects may differ in any dimensions of the linguistic spectrum such as syntactic, lexical, morphological, phonological differences, these changes are likely to be more indirect across dialects than those across languages such as Hindi, Punjabi and English etc.

DI mainly helps in ASR system since speakers with various dialects pronounce few words like differently, consistently changing certain phones, and even morphemes. This is proof, like example with the Hindi Language, which has various variants the formal written SL (Standard Language) of the education, media and culture and information spoken dialects that are the preferred technique of communication in a daily life. Although, there are commercially available ASR (Automatic Speech Recognition) systems for recognizing few standard dialects with minimum error rates, these recognizers failure, when a native Hindi speaker speaks in her/his regional dialect.

The deep learning is very efficient technique used for this purpose as in this technique the higher-level features are learned, it can identify direct relations, high measurement information and can deal with unlabelled information.

References

1) Chittaragi, N. B., Limaye, A., Chandana, N. T., Annappa, B., and Koolagudi, S. G.,"Automatic text-independent kannada dialect identification system", In *Information Systems Design and Intelligent Applications* (pp. 79-87). Springer, Singapore, 2019.

2) Dharmale, G., Thakare, V. M., &Patil, D. D., "Implementation of Efficient Speech Recognition System on Mobile Device for Hindi and English", *International Journal of Advanced Computer Science and Applications (IJACSA),* Vol. 10, No. 2, 2019.

3) Graves, A., Mohamed, A. R. and Hinton, G., "Speech recognition with deep recurrent neural networks," *In 2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). IEEE, (2013).

4) Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G. And Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing,* vol22, no. 10, pp. 1533-1545, 2014.

5) Pulz, G., Blanchard, H. E., Lewis, S. H., Zhang, L, "System and method for recognizing speech with dialect grammars." *U.S. Patent 9,361,880*, issued June 7, 2016.

6) Biadsy, F.,"Automatic dialect and accent recognition and its application to speech recognition", (Doctoral dissertation, Columbia University), 2011.

7) Biadsy, F., Hirschberg, J. And Habash, N. "Spoken Arabic dialect identification using phonotacticmodeling." In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, pp. 53-61. Association for Computational Linguistics, 2009.

8) Hermansky, H., Ellis, D. P and Sharma, S. Tandem connectionist feature extraction for conventional HMM systems." In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3, pp. 1635-1638. IEEE, 2000.

9) Lamel, L., Gauvain, J. L and Adda, G. (2002), "Unsupervised acoustic model training." In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I-877.IEEE, 2002.

10) Singer, E., Torres-Carrasquillo, P. A., Gleason, T. P., Campbell, W. M. and Reynolds, D. A. "Acoustic, phonetic, and discriminative approaches to automatic language identification." In *Eighth European conference on speech communication and technology*. 2003.

11) Truong, K "Automatic pronunciation error detection in Dutch as a second language: an acoustic-phonetic approach." (2004).

12) Desai, N., Dhameliya, K. and Desai, V. "Feature extraction and classification techniques for speech recognition: A review." *International Journal of Emerging Technology and Advanced Engineering* 3, no. 12 (2013): 367-371.

13) Chou, W. "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition." *Proceedings of the IEEE* 88, no. 8 (2000): 1201-1223.

14) Alhawiti, K. M. " *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 9, no. 6 (2015): 1351-1354.

15) Cooper, H., Holt, B. and Bowden, R. "Sign language recognition." In *Visual Analysis of Humans*, pp. 539-562.Springer, London, 2011.

16) Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J and Renals, S. Automatic dialect detection in arabic broadcast speech." *arXiv preprint arXiv:1509.06928* (2015).

17) Cavell, S. *Conditions handsome and unhandsome: the constitution of Emersonian perfectionism: The Carus Lectures, 1988*. University of Chicago Press, 2018.

18) Narang, S and Gupta, M. D. "Speech feature extraction techniques: a review." *International Journal of Computer Science and Mobile Computing* 4, no. 3 (2015): 107-114.

19) Chiţu, A. G., Rothkrantz, L. J., Wiggers, P and Wojdel, J. C. "Comparison between different feature extraction techniques for audio-visual speech recognition." *Journal on Multimodal User Interfaces* 1, no. 1 (2007): 7-20.

20) Liu, Z. T., Wu, M., Cao, W. H., Mao, J. W., Xu, J. P., & Tan, G. Z. "Speech emotion recognition based on feature selection and extreme learning machine decision tree." *Neurocomputing* 273 (2018): 271-280.

21) Yu, D and Deng, L. "Deep learning and its applications to signal and information processing [exploratory dsp]." *IEEE Signal Processing Magazine* 28, no. 1 (2010): 145-154.

22) Ruder, S. "An overview of multi-task learning in deep neural networks." *arXiv preprint arXiv:1706.05098* (2017).

23) Sainath, T. N., Weiss, R. J., Wilson, K. W., Li, B., Narayanan, A., Variani, E., and Misra, A. "Multichannel signal processing with deep neural networks for automatic speech recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, no. 5 (2017): 965-979.