RECENT DIALECT IDENTIFICATION TECHNIQUES FOR AUTOMATED SYSTEMS: REVIEW

Poonam Kukana[1], Dr. Neeru Bhardwaj[2]
Department of Computer Science and Engineering
Rayat and Bahra University, Sahauran, Kharar, Punjab, India.

Abstract- The diversity and growth of a language are evident from its several dialects. If several dialects are not identified in technical improvements such as automatic speech, speech recognition, and speech production, then these languages may disappear. The speech field is playing the main role in conserving different dialects of a language from going non-existent. In a manner to construct a full-fledged ASRS (automatic speech recognition system) that identifies different dialects, an ADI (Automatic dialect identification) system acting as the front-end is necessary. This is the same as how linguistic verification systems act as a frontend to ASRSs that manage various languages. In this article, discussed various deep learning methods, dialect recognition system method using MFCC, GMM, CNN, RNN with LSTM, and Bidirectional GRU, etc that create use of various fields of data in the audio signal to construct a system that classifies and recognizes the local dialect and pronunciation of the speaker. In particular, analyse the mel frequency method used to extract the features, frame-based audio, and high-level feature extraction and compare the theoretical and modeling methods. It studied the efficient methods of linguistic verification that have been effectively active by that public, applying them here to dialect recognition in speech. In this review analysis, we discussed the various types of dialect data sets such as Karnataka, IViE, and Hindi, etc.

Keywords- Automatic Speech Recognition System (ASRS), ADI (Automatic Dialect Identification), MFCC (Mel Frequency Cepstral Coefficient), and GMM (Gaussian Mixture Model).

I. INTRODUCTION

Speech or ASR (automatic speech recognition system) is the main procedure of translating a speech signal to a text format or word sequence signal to a text format or word sequence using a method developed as a computer program [1]. Speech communication and processing is the main research area and was inspired by a person's aspiration to construct mechanical structures to match human-verbal communication abilities. It is the simplest form of people communication and SP (Speech processing) has been one of the inspiring fields of signal processing. Voice recognition expertise has created the possibility for machines to survey social speech guidelines and appreciate human dialect or languages. However, the major goal of the SR field is to implement models for speech input samples to computers. Speech recognition is the main resource of communiqué among the people.

It created main improvements in SM (Statistical Modeling) of voice, ASRSs today search worldwide application in tasks that need HMI (human-machine interface) like in telephonic systems, and query-based data networks updated travel data, stock market prices (up and downs), data-entry, voice dictation, weather reports, information access such as banking sectors, automobile portals, voice transcription, blind persons in super-market, online reservations (Airplane and Railways), etc. Several uses of the SR domain have been described in the following table 1.

TABLE 1: SPEECH RECOGNITION DOMAIN APPLICATIONS [2]

Issues	Applications	Input	Design classes
Voice recognition/telephonic sector communication	Telephonic directory inquiry without OA (Operator Assistance)	Wave format	Spoken words
Education Field	Tech in foreign dialects to pronounce, vocab. Correction	Wave format	Spoken words
Outside Edu. Field	Computer Video games and gambling	Wave format	Spoken words
Domestic Field	Dishwasher, washing machine, oven, microwaves	Wave format	Spoken words
Army filed	Fighter aircrafts Helicopters Battle management Training in air traffic control Telephonic communication	Wave format	Spoken words
AI (Artificial Intelligence) field	Robots	Wave format	Spoken words

Dialect is the main source of communiqué among people. A similar dialect is used for different ASRSs for various motives. The dialect of a linguistic is measured as the main issue when using an SRS. SRS covers of speech processing signal consuming its structures and identification depend on the further classification of SRS. Pashto is one of the languages that is spoken by about 80 million people in regions of Pakistan, Iran, and Afghanistan [3]. There are more than 20 languages where selected are Yousafzai, Afridi, and Banuchi, etc. Though the languages are related to each other, still there is less understanding of the unique dialect. And, in maximum cases, an utterer of the real

linguistic may not be capable of appreciating the sound of the used language.

In addition, the speech handling area is enormous, due to non-linearity and difficulty with space and time [4]. Hence, current proceedings of the investigators have moved near the precise linguistic dialect detection. This is controlled by numerous dialects of Karnataka, Sanskrit, IViE, and Bangladeshi using properties of MF (Mel frequency), and classifying it using HMM, GMM, and SVM algorithms. Moreover, operated on the specific dialect of the Hindi language by mean of prosodic and spectral features, it is placed in neural networks. The precision was attained about seventy-nine percent when collective feature sets were utilized for the recognition scheme.[5] They used the mel frequency cepstral coefficient method to extract the properties from the phonetic data set. The different classes were identified using the discriminant analysis method that presented a precision of eighty-three percent train and test data from the data set.

This paper is ordered as: II section describes the existing related work and study. III section elaborates on the modeling and theoretical in dialect recognition methods. IV section explains the various dialect recognition dataset such as Kannada, Hindi, IViE, etc. Section V clarifies the deep learning methods, structures while the survey paper is concluded in VI sections.

II. RELATED WORK

M. Nanmalar et al., 2019 [6] implemented a novel approach to verify the classified Tamil dialects like colloquial and literary Tamil. It extracted the acoustical features rather than phonotactics and phonetics were used. One of the main benefits of this proposed approach was that it didn't need an annotated quantity; therefore, it could simply adapt to other languages. GMM (Gaussian Mixture Models) using MFCC algorithm (Mel Frequency Cepstral Coefficients) characteristics were used to evaluate the classification. The simulation result calculated an exception error rate of 12 percent. **Saud Khan et al., 2017 [7]** developed SVM (Support Vector Machine) classification model that was developed for Content based Dialect Classification (CBDC) and recovery. This proposed method was a concept of an on-going determination to identify the requirements of the novel under resourced dialects. The speech dialect classification model would research proposal for the importance and prosperity of the Pashto communication persons and would help in protecting the languages active by this procedure. Speech samples were composed of the motive of creating a data set that contains languages from various age and gender categories. The extracted vectors from the data set contain cepstral coefficients (CCs) and numerical metrics (SPs) which were well-defined by optimal group edges using SVMs. Simulation analysis defines that SVM model-based speech dialect recognition system provides valuable results in precisely unique between various dialects. **Chen-Yu Chiang et al., 2018[8]** defined an effective cross-dialect edition model for building prosodic structures for china language text to speech models. The dialect prosodic structures were modified from a prior mandarin speaking rate (MSR) based on the hierarchical prosodic model (HPM). The basic system depends on the cross-language connections among Chinese and Mandarin languages in the form of prosodic and syntactic models. In this survey, two major issues are identified: (i) It pertained to the use of cross-language connections in the project and edition of the Dialect speech value reliant on HPM. Some other issues were the data scarcity produced by the inefficiency of an edition quantity cover vital language prosodic and contexts activities and also worldwide SRR (speaking rate range). This issue was resolved by retaining the physical extreme a posteriori approach that relates to managing the DSR dependent HPM metrics into DTs (Decision trees) to ease metric estimations. This proposed method was calculated by simulations on two different dialects such as HAKKA and MIN. Moreover, the subjective calculations established that the prosodic characteristics created by the DSR reliant HPMs were normal in several speaking rates (SRs) extending from 3.31 to 6.72 per sec. **Rita Rahmawati et al., 2017 [9]** discussed the main motive of the spoken dialect recognition system in different languages. The basic study of this topic was to compare the valuable modeling characteristics and methods for the classification of Sunda and Java dialects in INDONESIAN speech using the Mel frequency cepstral coefficient. The Mel frequency method and pitch feature values were compared with the GMM model and vector modeling methods. This analysis was used in the machine learning (ML) method for the training procedure and tests the model. **Prashant Upadhyaya et al., 2017 [10]** defined the continuous Hindi speech recognition model (HSRM) using Kaldi Tool-Kit (KKT). For the classification, feature extraction using Mel frequency and perceptual linear prediction characteristics were removed from 1k phonetically managed Sanskrit or Hindi sentence formation from AMUAV quantity. Audio Model was evaluated using GMM, HMM, and decryption was evaluated on known as HCLG which was built from weighing fixed stage transducers (WFSTs). Evaluation of together tri-phone and mono-phone methods using the Ngram language model was described which was evaluated in the form of WER (Word Error Rate). The main reduction in word error value was supposed through the triphone method. In the future, it was searched that the Mel frequency cepstral coefficient (MFCC) feature extraction model gives a higher accuracy rate than other features (PLP). The main goal was to define the presentation of Hindi or Sanskrit language using the current model.

In table 2 discussed the basic methods were used in the existing speech dialect recognition system, performance metrics, tools, and issues, etc.

TABLE 2. COMPARATIVE ANALYSIS

Author Name	Title Name	Journal Name	Methods/ Tools	Language	Tools /Parameters	Issues
Nanmalar et al., [6] 2019	Literary and Colloquial Dialect Identification for Tamil using Acoustic Features	*TENCON 2019*	GMM MFCC Linguistic Tool	Tamil (Colloquial and Literary)	Accuracy (%ge) Error rate (%)	Complexity increases Classification Error (When Samples are similar).
Saud Khan et al., [7] 2017	Pashto Language Dialect Recognition using Mel Frequency Cepstral Coefficient and Support Vector Machines	*IEEE*	MFCC SVM GMM	Khattak Banuchi Afridi and Yousafzri	RMSE (%) Accuracy Rate WER (Word Error Rate)	-
Chen-Yu Chiang [8] 2018	Cross-Dialect Adaptation Framework for Constructing Prosodic Models for Chinese Dialect Text-to-Speech Systems	*IEEE Transactions*	HMM PLM SR-HPMs	Hakka Min Chinese	Pitch Value Utterance Count Duration standard deviation Mean SR denormalization HTS-2.2 toolkit	Cross dialect similarities Data Sparseness is caused by insufficient.
Rita et al., [9] 2017	Java and Sunda Dialect Recognition from Indonesian Speech using GMM and I-Vector	*IEEE*	MFCC Hybridization(MFCC+Pitch) feature extraction with GMM and i-vector	Java Sunda	Classification Error rate	-
Upadhyaya et al., [10] 2017	Continuous Hindi Speech Recognition Model Based on Kaldi ASR Toolkit	*WISPNET 2017*	GMM HMM PLP	N-gram Language Hindi	WER (Word Error Rate) SRILM Toolkit Kaldi Toolkit	Higher error rate and Complexity

III. MODELING AND THEORETICAL METHODS USED IN SPEECH DIALECT RECOGNITION SYSTEM (SDRs)

Speech Dialect is an alteration of linguistics that marks how a pronoun is spoken by a human being. SRS converts speech into texture format, and SD (Speech dialect) may mark the consequences of detection and recognition [11]. DR (Dialect Recognition) has been done with various methods such as MFCC, GMM, HMM, ANN, and SVM classifiers. Several procedures are explained in pointwise:

A. DRS (Dialect Recognition System)

DRS normally starts with a front-end procedure which is a voice considered procedure by phone. Before, the signal from the voice will be evaluated using feature extraction methods such as MFCC, and GFCC, etc. Dialect modeling (DM) is completed by utilizing valuable modeling and theoretical methods for demonstrating languages.

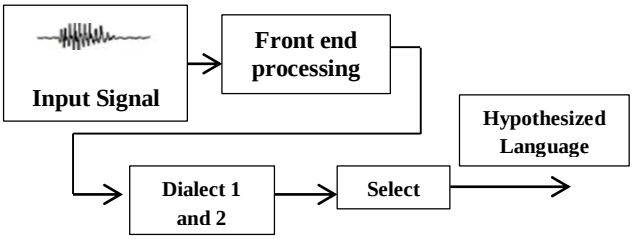

Figure 1. The basic diagram of DRS (Dialect Recognition System)[11]

After that, [11] Machine learning classification of the model using SVM, ANN, and other algorithms by choosing the max_value that depends on a certain score. Then, the last and final procedure of the DRS is the pre-defined dialect category defined in fig 1.

B. MFCC used in Feature Extraction Process

This is used for feature extraction. It means to calculate the spectral feature sets. FE initials from speech signals segmentation into small frame division normally ranges from 20 to 25 milliseconds. Before, introducing the MFCC window procedure from voice division and changing the voice signal into frequency layout by using a fast Fourier transformation (FFT) approach. Moreover, the sign of the

voice will be modified to the mel scale and generate a logarithmic mel. The last phase modified the logarithmic mel which is also known as a cepstrum as discussed in fig 2 [12].

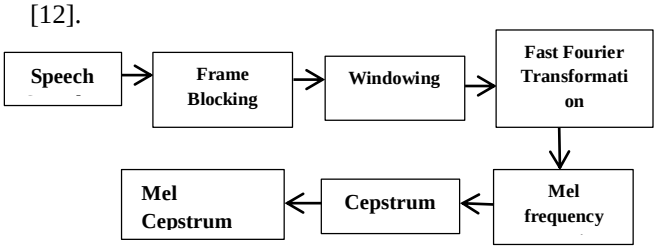

Figure 2. MFCC algorithm Steps

C. Pitch Feature Extraction Process [13]

This is a prosodic feature (PF) that is given as central frequency (F0), pitch-feature set completed by using the robust method for pitch tracking model. PF comprises three major coefficients of delta-pitch, voicing, and normalized pitch features extracted from speech signals [14].

D. GMM in Modeling Method (Gaussian Mixture Model)[15]

This modeling method is the amount of the Gaussian module that has been often wt. the Gaussian model mathematical formulation can be defined as below:

$$Q(y|\lambda) = \sum_{j=1}^{N} wt_j \; g(y|\mu_j, \sum j) \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(i)}$$

Here eq (i) is distinct as y is the dimension of the vector statistics, wt_j, i= 1,2,……..N, is the mixture wt., and $g(y|\mu_j, \sum j)$, j = 1, 2, ……N is the GD (Gaussian Division). In speech dialect recognition, λ is a metric covering $\{wt_j, \mu_j, \sum j\}$. where wt_j is the CGM (Co-efficient Gaussian Mixture), μ_j defines the spectral format of the ith audio category in the ith Gaussian and $\sum j$ defines the alteration of the spectral-form sum calculation. The high quality of wt_j value is completed by EM (Empirical Method), the highest the wt_j value is the maximum Gaussian wt.

E. I-Vector Modeling Method [16]

This concept is defined in the context of signal identification. This method is given as a complete inconsistency space inspired by the achievement of join element analysis where the utterer and sub-intersection space is generated distinctly. This method has all vital inconsistencies in a similar minimum dimensionality. The EVA (Eigen Voice Adaptation) method is ended with the supposition that all the inconsistency characteristics are quartered by t matrix which is a minimum dimensionality matrix. Gaussian Mixture Model super vector for speech dialects could be modeled by eq (ii):

$$N = n + T_{wt.} + \epsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(ii)}$$

Here eq (ii) shows the n is the super vector universal background model, Ivector $wt.$ an RNDV (Random Normalized Distributed Vector) of N(0,1) and $\varepsilon \sim N(0, \sum)$ modeling the interference inconsistency that excepted t matrix[17]. This has the feature of perception voice. So, it will be dependable for data assessment that must interference.

IV. VARIOUS LANGUAGES DATASETS USING IN DIALECT RECOGNITION SYSTEM

The fourth phase defines the detailed description of various dialect data sets such as Tamil, Assamese, and Kannada, etc. It comprises the details of the prior data set used for investigation. A language contextual of Karnataka linguistic and the process modified through the group of Karnataka data set are obviously defined in subsections and the same details of IViE data set are defined:

A. Kannada Language Data set

It is highly concatenated and rich in linguistics with the effect of Hindi in it. Same as some other dialects, the concatenate feature comprises the generation of novel-words along with suffixes to the root-word. Later, the difficult words are designed by additional expressive words together deprived of modifying them in predicting. Morpho-Syntax (MS) is regulated by the command where suffixes are involved in the original word. It has forty-nine microphones; of which fourteen are long-short like vowels and thirty-five are non-vowels. Table 3 shows the Kannada dialect describing all information such as dialect name, male, female, duration time and speaker, etc.

TABLE 3. SPEECH DIALECT IN KANNADA LANGUAGE[18]

Dialect Name	Age (yrs)	Participants		Participants		No. of Speakers
		Males	Duration (in min)	Female	Duration (in min)	
CENK Central Kannada	20 to 85	18	65	12	47	30
STHK South Kannada	21 to 76	16	78	13	50	29
MUBK Mumbai Kannada	25 to 80	12	85	14	45	26

B. IViE Speech data set

IViE data set full form is intonational variation in Eng. The nine dialect alternatives of British eng. Spoken in 9 various areas in the British isles are considered in the database. The amount has been restored in a manner to study the cross-variant, style changes, and inflection designs crossways 9 languages of eng. 9 British areas comprised are shown in table 4:

- ID1 : Belfast
- ID2 : Bradford
- ID3: Cardiff
- ID4 : Cambridge
- ID5: Dublin
- ID6: Leeds
- ID7: Liver pool
- ID8: London and
- ID9 : New castle

TABLE 4: IViE DATA SET [18]

Name of Dialect	No. of participants (M+F)	Time (in min)
ID1	12	32
ID2	12	31
ID3	12	35
ID4	12	37
ID5	12	33
ID6	12	31
ID7	12	26
ID8	12	38
ID9	12	31

C. Assamese data set

This data set comprises a telephonic voice of eight speakers with three mood changes like loud, normal, and angry, etc. Individual speakers are requested to utter 5 different sentences at least 3 times which comprise 2, 4, 5, and 6 phoneme words resp. Thus, 3 different data sets of total samples 360 in the database each for the train, learning, and validation process. In the testing process, there is a set of eighteen hundred samples. This is attained by including GNs (Gaussian Noise) having various SNRs (Signal to Noise Ratios).

V. DEEP LEARNING TECHNIQUES STRUCTURE

This process discussed the network model regarded as the deep learning model that is used for CDI (Chinese Dialects Identification). Various deep learning models are discussed as follows:

A. Convolutional Network Model in DL (Deep learning)

The end to end technology is used for application in deep neural networks (DNNs), it could evaluate the end-to-end modeling straightforward from the filtered-bank characteristics to the voice language identification that depends on the convolutional neural network (CNN) model. The HL (Hidden Layer) of convolutional networks normally comprises two phases: (i) the convolution layer (CL) and the pooling layer (PL) [19]. (ii) The CL consists of various CUs (Convolutional units), and the metrics of individual CU are improved by the BP (Back Propagation) approach to extract various characteristics of the effort. Providing the filter size, the individual layer is combined to the native cell sub-set of HL lower. Before, relating ReLU to roll the effort and filter, w is a wt. Enhances the biased form (b) which could be shown as:

$$HL = \max(0, wt. * y + a) \dots\dots\dots\dots\dots\dots\dots\dots\dots (i)$$

eq (i) HL is the non-linear production of the input-vector (y) after the lined change. The pooling layer is a feature-set (FS) with high-dimension attained after the convolution layer, by captivating the max-rate and avg-value to attain a novel FS with a minimum measurement.

B. BiDirectional GRU (Gated Recurrent Unit)

For input series y = (y1,y2,….. yt), the standard recurrent NN, evaluates the state vector series of the HL (Hidden Layer) hl = (hl1,hl2,……hltt) and the output vector z = (z1,z3,……ztt) by iterating from 1 to tt:

$$hltt = \sigma_{hl}\,(W_{yhl}\ y_{tt} + W_{hlhl}\ h_{tt-1} + b_{hl} \dots\dots\dots (ii)$$

$$Z_{tt} = \sigma_z\,(W_{yhl}\ hl_{tt} + b_z) \dots\dots\dots\dots\dots\dots\dots (iii)$$

Here, W defines the wt. matrix among the different layers, b_{hl} and b_z are off-set vectors of the HL and the OL resp; σ_{hl} and σ_z are used for AF (Activation Function)[20].

Gated Recurrent Unit is an irregular long short term memory (LSTM) that has a normal model and better convergence. Gated recurrent unit comprises an update and resets of the gate. The model of the Gated recurrent unit is defined in fig 3.

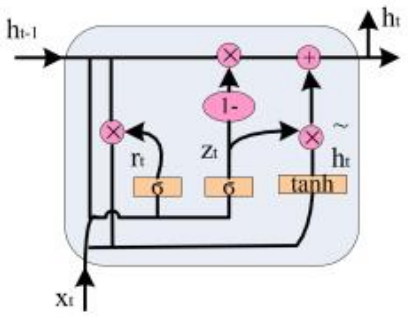

Figure 3. The basic model of GRU cell

The updation procedure of Gated recurrent unit (GRU) is defined as follow:

$$R_{tt} = \sigma\,(W_r\ y_{tt} + U_r\ hl_{tt-1} + b_r) \dots\dots\dots\dots\dots\dots\dots\dots (iv)$$

$$Z_{tt} = \sigma\,(W_r\ y_{tt} + U_z\ hl_{tt-1} + b_z) \dots\dots\dots\dots\dots\dots\dots\dots (v)$$

$$hl'_{tt} = tanh\left(W_r\ y_{tt} + U_{hl}\left(r_{tt} \odot hl_{tt-1}\right) + b_{hl}\right) \dots\dots\dots\dots (vi)$$

$$hl_{tt} = (1 - z_{tt}) \odot hl'_{tt} + z_{tt} \odot hl_{tt-1} \dots\dots\dots\dots\dots\dots\dots\dots (vii)$$

Where σ is the sigmoid and tan is the hyperbolic tangent method. Weight Matrix is U for the prior hidden state vector hl_{tt-1}, hl'_{tt} is the applicant stimulation method and $\odot$ is the element-wise multiply. The vector r_{tt}, z_{tt} denotes the re-set and re-new gate vector.

Moreover, the voice itself has a particular framework co-relation. The linguistic structure in the old SRS has in-sufficient memory space ability for ancient data, and can't fully study the significance of the speech series. BiGRU networks are explained above and the model is shown in fig 4.

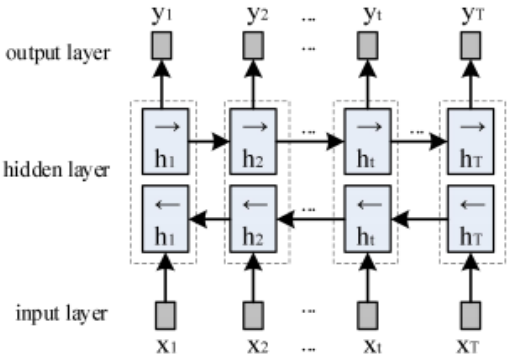

Figure 4. The BiGRU Model Structure

C. RNN using DL Classification Method

Recurrent Neural Network is a type of ANN (Artificial Neural Network), where links between nodes form a directed graph along with a series. It is normally a series of NN sections that are connected like a sequence.

Suppose an initial input data Y = (Y1,…….Yt), is a SRNN (Standard Recurrent Neural Network) calculates the HVS (Hidden Vector Sequence) V = (V1,……...Vt) and OVS (Output Vector Sequence) Z = (Z1,……..Zt) by repeating the subsequent eq (i) and (ii) from t = 1 to T*:

$$V_t = \mathrm{H}\,(W_{yV}\,V_t + W_{VV}V_{t-1} + b_V \ldots\ldots\ldots\ldots (i)$$

$$z_t = W_{VZ}\,V_t + B_V \ldots\ldots\ldots\ldots\ldots\ldots (ii)$$

Here, the W has belonged wt.matrix (an example W_{yV} is the input_hidden wt. matrix) and the B defines BV (Bias Vector) or Hidden_Bias_Vector and h is defined as HL (Hidden Layer) method.

H is normally an element-wise use of an SM (Sigmoid Method). But, it has been searched that the LSTM model [21], which uses motive construct MCs (Memory Cells) to saved data, is better at searching and using maximum range context. The novel version of the LSTM algorithm is used in this survey paper H [22] is developed by the subsequent CF (Composite Function):

$$j_l = \sigma\,(W_{Vl}V_t + W_{hl}h_{t-1} + W_{dl}D_{t-1} + B_{l\ldots\ldots\ldots(iii)}$$

$$F_l = \sigma\,(W_{VF}V_t + W_{hF}h_{t-1} + W_{dl}D_{t-1} + B_{F\ldots\ldots\ldots(iv)}$$

$$D_l = F_t\,D_{t-1} + l_t \tanh(\,W_{VD}V_t + W_{hD}h_{t-1} + B_{D\ldots\ldots\ldots(v)}$$

$$O_t = \sigma\,(W_{VO}V_t + W_{hO}h_{t-1} + W_{dO}D_{t-1} + W_{dO}D_t + B_{O\ldots\ldots\ldots(vi)}$$

$$h_t = O_t \tanh(D_t) \quad \ldots\ldots\ldots(vii)$$

Here, l, F, O and D are defined as the Input_gate[22], forget_gate, output_gate, and cell_activation_vectors and all of which are the similar size as the HV(Hidden vector) h. The wt. matrix from the cell of GVs (Gate Vectors) is traverse, so element m in individual gate vector-only gets input from the element m of the CV (Cell vector). The Logistic SM is σ.

VI. CONCLUSION

Various approaches have been discussed to recognize dialects of various languages but no well-known effort has been accepted for Pashto dialects (Afghanistan, Pakistan, and so on) language. It discussed the feature extraction method using MFCC (Mel Frequency Cepstral Coefficient), the GMM model and evaluated the mathematical metrics and prosodic features of the speech signals which are recognized using CNN, Bidirectional GRU, and LSTM classifiers. Gaussian mixture model-based classification method for verifying colloquial and literary Kannada and IViE is constructed. Gaussian models have been used in the existing research for language and dialect recognition systems. But it has not been used in classification in Karnataka Dialects. various classification methods discussed in this article. Several deep learning models like CNN, RNN with LSTM, Bidirectional GRU, and GMM algorithms are used to classify the dialect data sets. It has improved the training data, and sources an increase in system performance. The solution attained is easy and well-organized.

In further analysis, dialect recognition systems can be constructed with records and the consequences can be compared with the present algorithms. A deep analysis of the main motive of verbal nasalization in dialect recognition can be performed, achieve a high accuracy rate, and reduce the word error rate.

VII. REFERENCES

[1] Furui, S. (2005). 50 years of progress in speech and speaker recognition research. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, *1*(2), 64-74.

[2] Anusuya, M. A., & Katti, S. K. (2010). Speech recognition by machine, a review. *arXiv preprint arXiv:1001.2267*.

[3] Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*(5), 2421-2424.

[4] Das, P. P., Allayear, S. M., Amin, R., & Rahman, Z. (2016, February). Bangladeshi dialect recognition using Mel frequency cepstral coefficient, delta, delta-delta, and Gaussian mixture model. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)* (pp. 359-364). IEEE.

[5] Sinha, S., Jain, A., & Agrawal, S. S. (2014). Speech processing for Hindi dialect recognition. In *Advances in Signal Processing and Intelligent Recognition Systems* (pp. 161-169). Springer, Cham.

[6] Nanmalar, M., Vijayalakshmi, P., & Nagarajan, T. (2019, October). Literary and Colloquial Dialect Identification for Tamil using Acoustic Features. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 1303-1306). IEEE.

[7] Khan, S., Ali, H., & Ullah, K. (2017, April). Pashto language dialect recognition using mel frequency cepstral coefficient and support vector machines. In *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)* (pp. 1-4). IEEE.

[8] Chiang, C. Y. (2017). Cross-dialect adaptation framework for constructing prosodic models for Chinese dialect text-to-speech systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 108-121.

[9] Rahmawati, R., & Lestari, D. P. (2017, October). Java and Sunda dialect recognition from Indonesian speech using GMM and I-Vector. In *2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA)* (pp. 1-5). IEEE.

[10] Upadhyaya, P., Farooq, O., Abidi, M. R., & Varshney, Y. V. (2017, March). Continuous Hindi speech recognition model based on the Kaldi ASR toolkit. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 786-789). IEEE.

[11] Torres-Carrasquillo, P. A., Gleason, T. P., & Reynolds, D. A. (2004). Dialect identification using Gaussian mixture models. In *ODYSSEY04-The Speaker and Language Recognition Workshop*.

[12] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366.

[13] Talkin, D., & Kleijn, W. B. (1995). A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, *495*, 518.

[14] Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., & Khudanpur, S. (2014, May). A pitch extraction algorithm tuned for automatic speech recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2494-2498). IEEE.

[15] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1-3), 19-41.

[16] Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association*.

[17] Mak, M. W., Pang, X., & Chien, J. T. (2015). Mixture of PLDA for noise robust i-vector speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(1), 130-142.
[18] Trong, T. N., Hautamäki, V., & Lee, K. A. (2016). Deep Language: a comprehensive deep learning approach to end-to-end language recognition. In *Odyssey* (pp. 109-116).
[19] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
[20] Chittaragi, N. B., & Koolagudi, S. G. (2019). Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms. *Language Resources and Evaluation*, 1-33.
[21] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.
[22] Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of machine learning research*, *3*(Aug), 115-143.