Malaya Journal of Matematik, Vol. 8, No. 4, 1743-1750, 2020

https://doi.org/10.26637/MJM0804/0070

Accent based speech recognition: A critical overview

Rizwana Kallooravi Thandil[1]* and K. P. Mohamed Basheer[2]

Abstract

An incredible amount of research has been conducted in speech recognition and accent-based speech recognition during recent decades. Automatic Speech Recognition in various dialects in any natural language is examined as one among the most complicated domains in Automatic Speech Recognition (ASR). The increasing significance of speech recognition in any dialect is attributable to the ever-developing interest for applications that handle human-machine interaction through geographically influenced natural languages. The objective of this paper is to provide an overview of recent developments in dialect or accent-based speech processing. This paper concentrates on the study of accent-based speech recognition techniques in various languages and the technologies used for the same.

Keywords

Spoken language identification, Dialect identification, Accent recognition, Speech recognition, acoustic modeling, HMM, DNN, Acoustic features.

AMS Subject Classification

68T10, 62H30.

[1,2] *Department of Computer Science, Sullamussalam Science College, Areekode-673639, Kerala, India.*

***Corresponding author**: [1] ktrizwana@gmail.com; [2]mbasheerkp@gmail.com

Article History: Received **11** July **2020**; Accepted **18** September **2020**

Contents

1. Introduction

Speech has been always the predominant medium to intercommunicate and share information since time began. Around thousands of languages are spoken by people around the globe. Every language is rich in various forms of dialects. It is quite a difficult task to develop dialect-specific acoustic models since the dialect-specific data is inadequate and there are numerous dialects for each language.

As per Chambers JK et. al. [1], dialectology is the study of dialects and their geographic and social distribution. Dialect identification (DID) requires the recognition and patterning of prominent speech features of linguistic and paralinguistic information. Dialectology concentrates on sounds, pitch, stress, words, grammar, and morphological variation within a language. The Automatic Dialect Identification (ADI) recognizes dialects of a language from different speakers which is input as raw audio data. Dealing with dialects of any language in ASR is the most important and most complicated challenge to be addressed. The recent ten years have perceived significant advancements in automatically identifying the dialects/accent of a speaker, the emotions of the speaker, and the speaker itself from his/her speech. Another major hurdle in dialect identification is that it must deal with speakers who pronounce the same words with different accents. Dialect specific information in speech is contained in all the speech segments: at the entire segmental, subsegmental, or suprasegmental level. Different accents in any language or differences in the utter-

ance are due to the varying size and shape of the vocal tract of the speakers and the influence of a particular geographical area. Practically speaking, it is extremely complicated to cover all the dialects of any language. Speech signals modeled using Gaussian Mixture Models (GMMs) based on the Hidden Markov Models (HMMs) are used to implement Traditional ASR systems. Conventional ASR systems were modeled using statistical and probabilistic tools.
The speech signal is a piece by piece successive short-term stationary signal which can be considered as a Markov model for many stochastic processes. Every HMM utilizes a mixture of Gaussian to model a spectral representation of the acoustic signals. These methods are considered statistically inefficient for modeling non-linear or near non-linear functions [2], [3]. On the other hand, ANNs perform conditional modeling more effectively and efficiently. Many studies prove that better results are obtained when Deep Neural Networks are used other than conventional models like HMM.
A Dialect Identification (DID) system usually classifies the incoming speech utterances from a known language into one of the dialects, or more generally, classes of interest within that language. The terminologies "dialect", "accent" and "variety of pronunciation in a language" were used indistinguishably in the literature Liu et al. [4]. Most of the studies are found to use dialect and accent interchangeably. We have used the terminologies "dialect" and "accent" synonymously all over this paper to present our findings.

2. Approaches In Accent Identification And Speech Recognition

Researchers have explored several techniques for developing speech-based systems from the beginning. There is no such rule that one can only follow a certain methodology for conducting the experiments in ASR. Every approach has its own merits and demerits. The basic ASR procedure falls broadly into the acoustic-phonetic method, pattern recognition method, machine learning, and artificial intelligence method. Phonemes from the acoustic signal are selected and labeled in the acoustic-phonetic method. The acoustic signals are segmented and labeled in this approach which is further used for determining the word accurately [5]. Pattern training and pattern matching are the major steps involved in the pattern recognition approach. Here the ASR system will be trained with labeled data and some test data will be used to match against the trained data to predict the correct word [5][6]. The coordinated effort of the acoustic-phonetic approach and pattern recognition approach jointly makes the artificial intelligence and the machine learning approach. This modeling is performed either by the probabilistic approaches or by machine learning approaches. The utterances input to the system are compared against the model and is defined by analyzing the class to which the utterances fall into. The classification rules can be designed based on the linguistic, spectrogram, or phonetic knowledge of speech signal or we can have an unsupervised or supervised learning approach [6][19].
The general approach for accent-based speech recognition can be described as:

1. The acoustic signals comprising of different dialects will be given as input to the system.

2. The system then removes noise and records the pattern of the signal to train the system by labeling each pattern.

3. Later any signal that is input to the system is tested for a match in the system with already recorded patterns and are grouped under the labels to which they closely match which we refer to as classification.

4. Finally, the acoustic signals are recognized based on the classification to which they belong

Several methods are adopted for accent-based speech recognition in ASR. Researchers used conventional statistical methods like HMM and GMM [12] [13] [14] and few researchers recently rely on artificial neural networks and deep neural networks for DID to get better results [15][16]. It is also noticed that very few researchers used unsupervised learning techniques by using autoencoders for dialect-based speech recognition [19].
Several approaches can be adopted for extracting the features of the acoustic signals such as probabilistic methods, spectral methods, model-based methods, transform-based methods, and pattern recognition methods [7]. Among the above feature, extraction approaches pattern recognition methods are adopted widely. Principal Component Analysis, Independent Component Analysis, Zero Crossing detection, Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), Relative Spectral Filtering (RASTA), Mel Frequency Cepstral Coefficient (MFCC), Wavelet Transform(WT), GMM, HMM, Dynamic Time Warping(DTW), Vector Quantization (VQ), Neural Network(NN) and Convolutional Neural Network(CNN) are the prominent feature extraction techniques used by researchers in ASR[45]. In ASR different classification algorithms are used to identify the utterances accurately by comparing with the pre-trained model. Some of the classification algorithms used in ASR are K-Nearest Neighbor Classifier (KNN), Support Vector Machine (SVM) and Artificial neural network (ANN) classifiers.
For speech recognition systems with a small dataset, the statistical feature extraction methods like PCA, ICA, and ZCR are found to give better results though prone to noise. For medium, to large-sized datasets HMM, MFCC, PLP, LPC are found to give better results if the data is less noisy. Relative Spectral Filtering (RASTA) or RASTA-PLP gives better results when working with noisy data. The neural networks require huge data to give better results. It takes too much time for training the system to develop a model and it requires additional hardware like GPUs for excessive processing of data.

3. Recent Advancements In Dialect Identification

So many researches are being carried out in ADI both by academia and industry for its significance. Several works are going on in speech recognition in regional languages. A few languages have succeeded in developing ADI systems despite all the obstacles.

Dialect study started back in 1877 when George Wenker had conducted a series of surveys and studies to identify regional dialects [1, 37]. One of the earliest attempts were carried out by Bailey [38] to define the dialect and concluded that identifying dialects should be independent of the vocabulary, since the dialects may differ according to community or class within the same geographic locale.

ASR can be broadly classified into Speaker Independent (SI) speech recognition and speaker-dependent (SD) systems. The SI systems are more complex to model than SD systems since it requires to model speaker variations in addition to the acoustic information within a single model [2][41].

3.1 Statistical or Probabilistic Approach for Dialect Identification Systems.

Irakli Kardaval et al. [12] proposed a method to create an ASR for Georgian and related languages. The paper carries out an observation of isolated words in three accents. The authors built Monophone acoustic models using HMMs. They used MFCC (Mel Frequency Cepstral Coefficients) to extract acoustic features from the audio data. The features were normalized using cepstral mean normalization. The acoustic models were context, speaker, and gender independent for the phone recognizer. The paper also discusses an improved model for dialect recognition for speech recognition systems which throws light on the fact that its inversion can be used for text to speech synthesis in any dialect of a predefined language. F. Biadsy[13] in his Ph.D. thesis discusses that accent recognition is the process of identifying a speaker's regional accent of a known language identified by the acoustic signal alone. He discusses that accent recognition is more challenging and complicated than speech recognition due to greater similarity in accents of words in a language. The author presented different approaches and modeling techniques for identifying regional dialects among Arabic speakers. The work discusses on extracting frame-based acoustic, phonetic, phonotactic, and high-level prosodic features for identifying four broad Arabic dialects (Levantine, Gulf, Iraqi, and Egyptian). The paper also demonstrates how the Arabic dialect recognition system improves the Arabic speech recognition system. Pedro A. Torres-Carrasquillo et al. [14] presented a solution for dialect/accent identification for three sets of dialects: Chinese, English (American and Indian accented), and Arabic (discrimination between three dialects). The paper is based on GMM which was a dominant technology for dialect recognition at that time. The paper discusses discriminative training using maximum mutual information (MMI) and feature compensation using Eigen-channel compensation via factor analysis. Experimental results are shown for 30-second test utterances. The paper discusses three experiments for all tasks. 1) a GMM-UBM baseline system, 2) a GMM-UBM with VTLN and fLFA, and 3) a GMM-UBM with VTLN and fLFA plus 10 MMI iterations. The experiment showed the best results when worked with the fLFA-MMI system in general. The paper also shows that improvement in the performance of each class depends on various factors depending on each class. They also introduced an attempt for open set dialect scoring. The experiment faced issues like false alarm rates in the experiments and regarding the size of the training set selected.

Sreeraj V.V et. al. [18] put forward a method by using the feature-level fusion of MFCC and Teager Energy Operator (TEO) based features. The paper discusses the systematic evaluation of four dialects in the Malayalam language created in a studio environment with 300 speech samples each. The paper reveals that SVM was used for the classification in the experiment. The authors focused on combining the features from MFCC and TEO in the feature level and gained an accuracy of 78% which is higher than the results they obtained when using MFCC and TEO alone.

Lokesh et. al. [23] proposed an ASR using enhanced MFCC with windowing and framing method. The paper discusses using Laplace smoothing technique for language modeling to recognize the audio signals. The authors claim that the WER of the ASR is low when compared with wavelet-based feature extraction and artificial neural network-based feature extraction methods for speech recognition in the experiment carried out by them.

Kumpf and King proposed a method to use linear discriminant analysis (LDA) for the identification of accents in Australian English [25]. Chen et al. [26] had experimented on the effect of the number of components in GMMs on classification. Deshpande et al. [29] used GMMs for frequency feature extraction to discriminate between American and Indian accents in the English language. Tang and Ghorbani et. al. [28] in their paper discuss the performance of HMMs with Support Vector Machine (SVM) for accent classification.

3.2 Deep Learning Approach for Dialect Identification Systems.

N. D. Londhe et al. [15] focused on machine learning solutions for speech recognition in the Chhattisgarhi dialect. The paper presented a comparative study among the various classification algorithms like HMM, SVM, and ANN for speech recognition. The paper focused on feature extraction using the MFCC algorithm. The experiment discussed in the paper showed that the best results were obtained when SVM was used and ANN outperformed HMM in speech recognition.

Ahmed Ali1 et. al. [16] discusses the experimenting details in dialects in Arabic speech. The dataset utilized was audio data from 19 different programs from Aljazeera Arabic TV channel in the period (2005-2015). They analyzed the speech transcription and word alignment of five Arabic dialects. It

can be considered as a good reference paper for dialect identification tasks using various methods. The experiment included data pre-processing, data selection, acoustic modeling (AM), and language modeling (LM), as well as decoding for the challenge, via a GitHub repository. The paper reveals that the best results were obtained from the collaboration of numerous deep and sequential neural networks for acoustic modeling. The authors worked with Recurrent Neural Networks (RNN) with *n*-gram modeling for language modeling. The authors successfully developed a baseline system with good precision and recall value.

Nassif, A. B et. al. [20] conducted a systematic review in speech recognition using neural networks right from 2006 till 2018. They conducted an overall survey in all aspects of speech recognition. From the survey, the authors found out that using neural networks in speech recognition very well outperforms the conventional statistical HMM/GMM based approaches.

Hinton et al. [21], focuses on the use of deep neural networks that contain many numbers of hidden layers. The authors summarize the advantage of a feed-forward neural network that has quite a few frames of coefficients as input and produces subsequent probabilities over HMM states as output. The experiment has shown that deep neural networks with many hidden layers that are trained by new techniques outperform GMMs - HMMs on a variety of speech recognition benchmarks.

Yishan Jiao1 et. al. [24] proposed a technique for the classification of eleven accents speech acoustics. The authors proposed a system that collaborates Long Short Term Memory(LSTM) features of the speech signal using DNNs and RNN. The authors focused on identifying local language given accents. The paper reveals that the proposed method with DNN and RNN for ADI surpasses the performance SVM-based baseline system.

3.3 Development of High Performance DID systems

Levent M. et. al. [8], Torres-Carrasquillo PA [9], Hanani A et. al. [10] conducted experiments on various acoustic-phonetic signals for dialect identification. The audio signals are represented by spectral features and paralinguistic information otherwise known as prosody features. The spoken words can be identified by the help of lexical and sub-lexical units of the audio signal. Prosody helps in detecting word and syllable boundaries in continuous speech utterances.

Benzeghiba, M et.al.[11] conducted an elaborated study on vital references to literature based on endogenous variations of the speech signal and their importance in ASR. The authors focused on methods for diagnosing weaknesses in speech recognition approaches. The authors had put forward an overview of general and specific techniques for handling variation sources in ASR in a better manner.

Yoo, S., Song, I. et. al. [17] proposed a novel technique for modeling a multi dialect using a single acoustic model. In the paper, the authors discussed applying Feature-wise Linear Modulation (FiLM) transformations to the AM based on dialect information. The authors used the combination of both external and internal information in feature-wise transformations to make the AM more adaptive to deal with multiple dialects. They also proposed methods to handle unknown dialects during the training period.

Tan Z et. al. [40] worked on Code-Switching (CS) in Mandarin-English language spoken by the Chinese natives. CS is the tendency of the speakers to include multiple languages in a single communication while speaking. The authors proposed methods to enhance the quality of the recognition of code-switched audio data. The paper concentrates on efforts to minimize disparity in accents of a language. Not so many works have been proposed in CS-based ASR since it is extremely complicated. Zhang, Q et. al. [19] focused on experimenting in Language/Dialect Recognition (LID/DID) based on unsupervised deep learning methods. The authors proposed two strategies to enhance the performance: First, an unsupervised bottleneck feature extraction solution was proposed, and secondly, two types of autoencoder modeling were introduced for speech feature extraction. Variational autoencoder and adversarial autoencoder were used in the study. They were the first to make attempts for speech signal processing using autoencoders. The paper focuses on the result that the unsupervised bottleneck solution consistently outperforms MFCCs even under noisy situations and the autoencoder applied at the frame level outperforms that when applied to utterance level. The paper throws light on the fact that adversarial autoencoder is more effective for similar dialects.

Rao, K. et. al. [22] proposed a method that used grapheme based acoustic models for ASR using a hierarchical RNN architecture with connectionist temporal classification (CTC) loss. The paper focuses on training the system with a single multi-dialect model using the dataset with US, British, Indian, and Australian accents for the English language and then familiarize the model using the US accent alone. The authors concluded that when used a single pronunciation dictionary for all the data degraded the model performance. The authors concluded that for single dialect models the grapheme-based models are inferior to the phoneme models for all languages and on the contrary, in multi-dialect models, the grapheme-based models significantly outperform the phoneme-based models. The list of papers considered for the study has been summarized in the Table 1.

4. Dialect Identification In Literature

This section discusses the contribution of various researchers in dialect identification in ASR. Table II: the most significant DID methods found in the literature are summarized [7][35][36][39][40][42][44]. The summary of ASR and DID in literature is given in table 2

Author(s)	Language	Ref.	Feature Extraction	Classifier	Dataset	SD/SI	Results
Irakli Kardava et al.-2016	Georgian	[12]	MFCC	Hidden Markov Models	Isolated words (3 accents)	SI	Presents an improved method to solve the accents problems of speech recognition systems.
F. Biadsy-2011	Arabic dialects, American English vs. Indian English. and three Portuguese dialects	[13]	1. MFCC 2.PLP with cepstral mean and variance normalization (CMVN) CC	HMM, binary Multi-Layer Perceptron, GMM-UBM SVM	1. Gulf Arabic Speech database 2. Iraqi Arabic. 3. Arabic CTS Levantine 4. e CallHome Egyptian 5.TDT4 6. LDC in 2003. 7. GALE data collection	SI	Classification accuracy of 86.3 on 2-minute-long utterances for phonotactic features that used only prosodic modeling. HMM-72.0 Four broad Arabic dialects could be identified.
Pedro A. Torres Carrasquillo et al. 2008	1)Chinese 2)English 3)three Arabic dialects.	[14]	GMM using shifted delta cesptra (SDC)	A set of diagonal covariance Gaussian classifiers	2007 NIST LRE corpus LDC Arabic set		Experimental results are shown for 30-second test utterances. The experiment showed the best results with fLFA-MMI system
N. D. Londhe et.al-2016	Chhattisgarhi	[15]	MFCC	HMM ANN and SVM	Isolated words	SD SI	HMM-62 ANN-78 SVM-82
Ahmed Ali1 et al.2016	Arabic	[16]	MFCC	RNN with n-gram models.	Aljazeera Arabic TV channel from March 2005 to December 2015.	SI	Precision 83 Recall 70
Yoo, S., Song, I., and Bengio, Y. (2019).	English	[17]	standard Kaldi recipe s	unidirectional RNN with LSTM	Libri Speech corpus 2.commercial datasets from Speech Ocean	SI	Not available
Sreeraj V.V et.al. -2017	Malayalam	[18]	MFCC and TEO	SVM	Malayalam dialect database	SI	MFCC-65, TEO-73 The combined system-78
Zhang, Q et. al. -2018	1) Chinese 2) Arabic	[19]	MFCC bottleneck feature extraction Autoencoder based feature extraction	DNN	1.DS four Chinese dialects. 2.Pan-Arabic corpus. 3.Multi-Genre Broadcast challenge corpus.	SI	Using traditional BNF the accuracy decreases from 95.5 to 90.1. By using the proposed BNF the average accuracy is 97.8. Using all the proposed autoencoder features the accuracy is 98.1.
Rao, K et. al. -2017	English	[22]		RNN, CTC, Multi Dialect Hierarchical Model	1. Google voice 2.English(British Indian and Australian accents)	SI	The grapheme-based system performs less by 0.7-10.3. The multi-dialect grapheme based model performs 9.5-16.7 better.

Table 1. Summary of significant works in DID

Sl.No	Citation	Result contributions.
1	Bhuvaneshwari Jolad, et. al.	The authors throw light on various methods for ASR systems with several approaches to feature extraction and classification. The authors likewise portrayed different methodologies followed by researchers in the speech recognition system in detail.
2	Sinha S. et. al.	The authors examine the influence of dialectal variations in the acoustic signals, by examining formant frequencies, pitch, pitch slope, duration, and intensity of vowel sounds. The authors used support vector machines along with dialect-specific Gaussian mixture models for the classification of dialects in the utterances.
3	Etman, A et. al.	In their work describes a detailed study of the ADI and proposed that temporal and prosodic features of acoustic signals are incredibly significant in DID since they comprise more detailed information about the different speech patterns.
4	Kumar Y et. al.	The authors presented a systematic review on ASR. The paper discusses various feature extraction techniques used in ASR.
5	Tan, Z et. al.	The authors propose methods to improve dialect independent speech for the Mandarin English code-switching.
6	Y. Gao et. al.	The authors proposed a novel method for language modeling by using Generative Adversarial Networks (GAN) to generate code switched text data.
7	Patil, U. G. et. al.	The authors audited important works held in ASR in recent years. The paper also discusses various research gaps to be filled in speech recognition.
8	Baran Uslu et. al.	The authors proposed a method for DID in the Turkish language. They focused on the prosodic features of the language for the study. The paper reveals that the neural network was used for the training of the system. They used only three features of the speech signal: pitch, jitter, and shimmer.

Table 2. Summary of ASR and DID in literature

5. Conclusion

An overwhelming measure of exploration has been directed in ASR and DID during recent decades. Several approaches were adopted by the researchers in experimenting with ASR and DID systems. Through an elaborative study of the dialect identification problem, we have seen the earlier studies were carried out using the conventional statistical and probabilistic approach with GMMs and HMMs. In recent decades studies related to ASR and DID have switched to ANNs, especially Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs). From our study, we found that neural networks have been widely used in state-of-the-art speech systems which showed better performance and lesser WER when compared with the traditional HMM and GMM. We have also observed that both temporal and prosodic features of the audio data contain more detailed information about different accents of a word, and it can be effectively used to work with ADI. It is also found that the ASR systems exhibit lower performance when evaluated on utterances with multiple dialects.Summary of ASR and DID in literature is given in Table 2.

References

[1] J. K. Chambers and P. Trudgill, *Dialectology,* Cambridge University Press, Cambridge, 1998.

[2] H. Singh and A.K. Bathla, A survey on speech recognition, *Int. J. Adv. Res. Comput. Eng. Technol.,* 2(6), (2013), 2186–2189.

[3] Y. Zhang, Speech recognition using deep learning algorithms, *Stanford Univ.,* Stanford, CA,USA,Tech.Rep., (2013), 1–5.

[4] M. Liu, B. Xu, T. Hunng, Y. Deng and C. Li, Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling, *In: Proceedings acoustics, speech, and signal processing,* 2(2000), 1025–1028.

[5] M. A. Anusuya, S. K. Katti, Speech Recognition by Machine: A Review, *International Journal of Computer Science and Information Security*, 6(3), (2009).

[6] A. P. Singh, R. Nath, and S. Kumar, A Survey: Speech Recognition Approaches and Techniques, *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gorakhpur, (2018), 1–4.

[7] Bhuvaneshwari Jolad and R. Rajashri Khanai, An Art of Speech Recognition: A Review, *2019 2nd International Conference on Signal Processing and Communication (ICSPC)*.

[8] M. Levent and JHL. Hansen, Language accent classification in American English, *Speech Commun.,* 18(4), (1996), 353–367.

[9] PA. Torres-Carrasquillo, TP. Gleason and DA. Reynolds, Dialect identification using Gaussian mixture models, *In: ODYSSEY 04-The speaker and language recognition workshop*, (2004), 297–300.

[10] A. Hanani, MJ. Russell and MJ. Carey, Human and computer recognition of regional accents and ethnic groups from British English speech, *Comput Speech Lang.,* 27(1), (2013), 59–74.

[11] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet and C. Wellekens, Automatic speech recognition and speech variability: A review, *Speech Communication*, 49(10-11), (2007), 763–786.

[12] I. Kardava, J. Antidze and N. Gulua, Solving the problem of the accents for speech recognition systems, *International Journal of Signal Processing Systems*, 4(3), (2016), 235–238.

[13] F. Biadsy, Automatic dialect and accent recognition and its application to speech recognition, Ph.D. thesis, Graduate School Arts Sci., Columbia Univ., New York City, NY, USA, (2011), 1–171.

[14] A. Pedro, Torres-Carrasquillo, Douglas Sturim, A. Douglas, Reynolds and Alan McCree, Eigen-channel Compensation and Discriminatively Trained Gaussian Mixture Models for Dialect and Accent Recognition, MIT Lincoln Laboratory, *Information Systems Technology Group, Lexington*, MA, USA.

[15] N. D. Londhe, M. K. Ahirwal and P. Lodha, Machine Learning Paradigms for Speech Recognition of an Indian Dialect, *International Conference on Communication and Signal Processing,* 2016, India, IEEE.

[16] Ahmed Ali1, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals and Yifan Zhang, *The mgb-2 challenge:arabic multi-dialect broadcast media recognition*, 2016.

[17] S. Yoo, I. Song and Y. Bengio, A Highly Adaptive Acoustic Model for Accurate Multi-dialect Speech Recognition, *ICASSP 2019-IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[18] V. V. Sreeraj and R. Rajan, Automatic dialect recognition using feature fusion, *2017 IEEE International Conference on Trends in Electronics and Informatics*, 2017.

[19] Q. Zhang and J. H. L. Hansen, Language/Dialect Recognition Based on Unsupervised Deep Learning, *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 26(5), (2018), 873–882.

[20] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, Speech Recognition Using Deep Neural Networks: a Systematic Review, *IEEE Access*, 2019.

[21] G. Hintonet al, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE SignalProcess. Mag.,* 29(6), (2012), 82–97.

[22] K. Rao, and H. Sak, Multi-accent speech recognition with hierarchical grapheme-based models, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[23] S. Lokesh, and M. R. Devi, *Speech recognition sys-*

tem using enhanced mel frequency cepstral coefficient with windowing and framing method, Cluster Computing, Springer, 2017.

[24] Yishan Jiao1, Ming Tu, Visar Berisha and Julie Liss, Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features, *INTERSPEECH,* 2016.

[25] K. Kumpf and R. W. King, Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks, *In Proc. Euro Speech,* 4(1997), 2323–2326.

[26] T. Chen, C. Huang, E. Chang and J. Wang, Automatic accent identification using gaussian mixture models, *In Automatic Speech Recognition and Understanding*, IEEE Workshop on. Madonna di Campiglio, Italy: IEEE, (2001), 343–346.

[27] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Juraf sky, R. Starr, and S. Y. Yoon, Accent detection and speech recognition for shanghai-accented mandarin, *In Interspeech, Lisbon, Portugal: Citeseer*, (2005), 217–220.

[28] H. Tang and A. A. Ghorbani, Accent classification using support vector machine and hidden Markov model, *In Advances in Artificial Intelligence*, Springer, (2003), 629–631.

[29] S. Deshpande, S. Chikkerur and V. Govindaraju, Accent classification in speech, *In Automatic Identification Advanced Technologies*, Fourth IEEE Workshop on. Buffalo, NY, USA: IEEE, (2005), 139–143.

[30] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen and T. N. Sainathet, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *Signal Processing Magazine,* IEEE, 29(6), (2012), 82–97.

[31] H. Zen and H. Sak, Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis, *In Acoustics, Speech and Signal Processing (I-CASSP)*, IEEE International Conference on Brisbane, Australia: IEEE, (2015), 4470–4474.

[32] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, An experimental study on speech enhancement based on deep neural networks, *Signal Processing Letters*, IEEE, 21(1), (2014), 65–68.

[33] Y. Jiao, M. Tu, V. Berisha, and J. Liss, Online speaking rate estimation using recurrent neural networks, *In acoustics, Speech and Signal Processing*, IEEE International Conference on Shanghai, China: IEEE, 2016.

[34] A. Rabiee and S. Setayeshi, Persian accents identification using an adaptive neural network, *In Second International Workshop on Education Technology and Computer Science,* Wuhan, China: IEEE, (2010), 7–10.

[35] S. Sinha, A. Jain and S. S Agrawal, Empirical analysis of linguistic and paralinguistic information for automatic dialect classification, 2017.

[36] A. Etman, and A. A. L. Beex, Language and Dialect Identification: A survey, *SAI Intelligent Systems Conference (IntelliSys),* 2015.

[37] A. A. Nti, *Studying dialects to understand Human Languages*, M.S. thesis Massachusetts Institute of Technology, 2009.

[38] Y. Kumar and N. Singh, A Comprehensive View of Automatic Speech Recognition System - A Systematic Literature Review, *2019 International Conference on Automation, Computational, and Technology Management (ICACTM),* 2019.

[39] Z. Tan, X. Fan, H. Zhu and E. Lin, Addressing Accent Mismatch In Mandarin-English Code-Switching Speech Recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[40] V. K. Muneer, K. P. Muhamed Basheer and Rababa Kareem Kollathodi, Smart device controlling through voice commands given in Malayalam language, *Malaya Journal of Matematik*, 5(1), (2019), 445–450.

[41] Gao, J. Feng, Y. Liu, L. Hou, X. Pan, and Y. Ma, Code-switching sentence generation by bert and generative adversarial networks, *Proc. Interspeech,* (2019), 3525–3529.

[42] U. G. Patil, S. D. Shirbahadurkar, and A. N. Paithane, Automatic speech recognition models: A characteristic and performance review, *2016 International Conference on Computing Communication Control and Automation (ICCUBEA),* 2016.

[43] Baran Uslu, Hakan Tora, Turkish Regional Dialect Recognition Using Acoustic Features of Voiced Segments, *International Journal of Signal Processing Systems,* 6(2), (2018).

[44] Haoye Lua, Haolong Zhang, Amit Nayak, A Deep Neural Network for Audio Classification with a Classifier Attention Mechanism, *arxiv.org,2006*, 2020.

★★★★★★★★★
ISSN(P):2319 – 3786
Malaya Journal of Matematik
ISSN(O):2321 – 5666
★★★★★★★★★