Covenant Journal of Informatics & Communication Technology. Vol. 10 No. 2, December, 2022
ISSN: *print* 2354 – 3566 *electronics* 2354 – 3507 DOI: An

Open Access Journal AvailableOnline

A Review of Accent-Based Automatic Speech Recognition Models for E-Learning Environment

[1]Omojokun Gabriel Aju, [2]Veronica Ijebusomma Osubor

[1]Department of Computer Science, Adekunle Ajasin University, Akungba-Akoko,Nigeria
[*]Corresponding Author: omojokun.aju@aaua.edu.ng

[2]Department of Computer Science, University of Benin, Benin City,Nigeria
viosubor@uniben.edu

Received: 23.11. 2022 Accepted: 08.12.2022

Publication: December 2022

Abstract— The adoption of electronics learning (e-learning) as a method of disseminating knowledge in the global educational system is growing at a rapid rate, and has created a shift in the knowledge acquisition methods from the conventional classrooms and tutors to the distributed e-learning technique that enables access to various learning resources much more conveniently and flexibly. However, notwithstanding the adaptive advantages of learner-centric contents of e-learning programmes, the distributed e-learning environment has unconsciously adopted few international languages as the languages of communication among the participants despite the various accents (mother language influence) among these participants. Adjusting to and accommodating these various accents has brought about the introduction of accents-based automatic speech recognition into the e-learning to resolve the effects of the accent differences. This paper reviews over 50 research papers to determine the development so far made in the design and implementation of accents-based automatic recognition models for the purpose of e-learning between year 2001 and 2021. The analysis of the review shows that 50% of the models reviewed adopted English language, 46.50% adopted the major Chinese and Indian languages and 3.50% adopted Swedish language as the mode of communication. It is therefore discovered that majority of the ASR models are centred on the European, American and Asian accents, while unconsciously excluding the various accents peculiarities associated with the less technologically resourced continents.

Keywords — Accents, Speech, Automatic Speech Recognition, E-learning, Feature Extraction Method.

1. Introduction

The Internet has become a major source of resources to the scholars and learners to acquire and share information with the advantage of integrating ICT applications into the teaching and learning processes (Akorful, 2014). The approach of incorporating multimedia technologies and Internet to education to increase learning quality by way of

easing access to learning facilities and services as well as distance exchanges of ideas and collaboration describes the concept of e-learning (Algahtani, 2011). An aautomatic speech recognition (ASR) technique is the process of converting speech signals into corresponding textual representations which has essentially been adopted in data analysis applications that process multimedia (audio/video) contents, such as speaker detection, and in applications that use voice in human-machine interfaces, such as intelligent personal assistants, interactive voice response (IVR) systems and voice search, among others (Toledano et al, 2018). The automatic speech recognition (ASR) has become an integral part of the e-learning technology and therefore assisting in the internationalization of the e-learning activities, the major problem however, is that a generalized ASR could not take care of the needs of all the learning participants (the trainers and the trainees) from different accents background within the e-learning environment.

This problem arises as a result of the e-learning environment majorly using international dominant languages, such as English, French, Mandarin and Chinese languages as media of communications, whereas the majority of these international dominant languages speakers are non-native speakers of the languages, therefore, as more people speak the dominant languages as a second language, there are increasing number of such languages' accents. Adjusting to and accommodating these various accents has become a major challenge for effective communication in the e-learning environment as accent differences have proved to represent a serious degradation that has been shown to affect the comprehension level and learning attitude of listeners (Suzanne, 2015).

Several accents-based automatic speech recognition models (Tan et al, 2021; Tawaqal and Suyanto, 2021; Ying et al, 2019; Upadhyay and Lui, 2018; David and Viktor, 2017; Sivaranjani and Bharathi, 2016; Zhenhao, 2015; Mannepalli et al, 2015 and Zhao et al, 2014) have been developed for e-learning platforms in the developed countries, particularly in United Kingdom, USA, China, India and Australia. However, while accents-based automatic speech recognition is constantly improving and new tools and methods are under constant development, most works in the field have focused on North American, Australian, European and Asian speakers, such as Chinese English speakers, Japanese English speakers and Indian English speakers where access to digital resources are concentrated (Sruti and Sanjib, 2018). This unequal availability of automatic speech recognition technologies hinders the technological growth and participation of the less digitally advanced regions with numerous accents differences in the e-learning education and trainings.

David and Viktor (2017) suggested that producing and making available an automatic speech recognition system that can identify as much as possible accents from different speakers on various international languages will increase the dissemination of information and acquisition of knowledge while reducing the barriers to e-learning education methods. Tan et al (2021) asserts that one way of improving this speech recognition divide is to do more research on the

portability of speech and language technologies for multilingual applications, especially for under-resourced languages and accents. This work therefore reviews the existing accents-based automatic speech recognition models in the e-learning environment. This review paper is structured in line with the research papers' components structuring guideline of Misra (2021).

2. E-Learning Technology

The old correspondence programmes were the earliest distance learning programmes of making learning activities available irrespective of the location of the learners, interestingly, the advent of computing technology changed the mode of delivering educational programmes over a long distance by opening new possibilities of learning activities for all academic programmes, be it undergraduate or graduate levels, employees training, research activities or any other type of academic offerings. The introduction of variety of technologies has not only promoted the delivery of learning through plain text, it has also enabled the creation of virtual classrooms via video conferencing (Deepali et al., 2016).

According to Organization for Economic Cooperation and Development (2005), e-learning is defined as the use of information and communication technologies in diverse processes of education to support and enhance learning, including the usage of information and communication technology as a complement to traditional classrooms, online learning or mixing the two modes. Algahtani (2011) divided e-learning into two basic types of computer-based and the internet-based e-learning. The computer-based e-learning involves the use of a full range of computer resources (hardware and software) instead of the traditional methods by providing interactive software as a support tool within the class or as a tool for self-learning outside the class. The internet-based e-learning is considered as a further improvement of the computer-based learning, in which the learning resources are made available on the internet with links to other related knowledge sources.

The development of multimedia and information technologies, as well as the use of internet as a new technique of teaching has made radical changes to the traditional process of teaching (Deepali et al, 2016), while Huang and Chiu (2014) observed that the development in information technology has presented more choices for today's education. Agendas of schools and educational institutions have recognized e-learning as having the prospect to transform people, knowledge, skills and performance (Mohammed et al, 2014). The colleges, universities and other institutions of higher learning have continue to advance online course capability in the rapid developing cyber education market since the conception of e-learning technology (Love & Fry, 2006).

In the recent years in Nigeria, just like their counterparts in America, Asia and Europe, institutions of higher learning have increasingly invested in course management applications to provide virtual learning environments to enhance students learning and to assist in the administration of the course itself, while the university lecturers and other instructors are currently being encouraged to find ways to help their students improve their learning skills both inside

and outside of the classroom (Joy & Peter, 2014; Taiwo & Downe, 2013).

The e-learning programme of several Nigerian universities like Federal University of Technology, Akure (FUTA); National Open University of Nigeria (NOUN), Covenant University, Baze University and University of Nigeria, Nsukka (UNN) have supported and enhanced the learning processes through technological syncretism by blending the traditional educational structures and practices with the new technologies leading to the development of effective pedagogical practices and the enhancement of learning (Olabode et al, 2015).

3. Speech and Accents

Speech is considered as the primary and the most convenient means of communication between people which is defined as human vocal communication using language (Nesdale& Rooney, 1996). It contains a number of information cues that includes the spoken message content (Jain et al, 2007), and the speakers' related information like identity, language, gender, age and emotion (Benzeghiba et al, 2007; Bocklet et al, 2008; Kinnunen & Li, 2010; Koolagudi & Rao, 2012).

Language defines a set of rules for generating speech (Fromkin et al, 2013), and around 6,900 known living spoken languages exist in the world with most languages belonging to a certain language family (Suzanne, 2015). A spoken language considerably varies in terms of its regional dialects and accents. Dialect refers to linguistic variations of a language while accent refers to different ways of pronouncing a language within a community (Nerbonne, 2003). Macmillan English Dictionary defines the term accent as a way of saying words that shows what country, region or social class someone comes from.

Following this general definition, O'Grady et al (2005) defines an accent in terms of phonetics in which particular type of pronunciation determined by the phonetic characteristics of the speaker's mother tongue is transferred to his/her use of another language. Accents are therefore caused by the influence of one's first language (mother tongue) on the second language which is induced by variations in different word pronunciation and grammatical structures into the second language (Flege et al, 2003; Grosjean, 2020; Mai & Hoffmann, 2014). Interestingly, these variations are not random across speakers of a given language; this is because the source of these variations is the speakers' original mother tongue. For example, a Hausa speaker in Nigeria might have problems in pronouncing the /p/ consonant at the beginning of the English word "power", they often pronounce it as /f/.

Munro and Derwing (2006) observed that one of the reasons why so much attention is given to accent-related issues is because of the fact that there is a growing awareness, among researchers, teachers and learners of the key role of pronunciation in communication. According to the voice principle, students learn more when the narration in a multimedia lesson is spoken by a native voice rather than a non-native voice (Munro and Derwing, 2006). Flege (1984) argued that, the recognition of accents is related to acoustic differences between native and non-native speakers' segmental

articulations and suprasegmental levels. Segmental articulation is concerned with segment-related problems like saying "tree" instead of "three" and segments such as vowel and consonant allophones. Such parameters of intonation as prosody, rhythm, tempo, melody and more importantly the juncture in a language are unique to each language and make each language different from one another to a greater extent. Crystal (2003a) explains juncture as a boundary or transition point in phonological sequence which includes syllable, foot, morpheme and word boundaries. They are believed to play a role in certain phonological generalizations. The non-native speakers cannot apply the rules of juncture in their speech and that is why, they fall short of uttering the words in the flow of speech and they sound foreign-accented.

Accent differences and bad pronunciation have also shown to lead to social and professional discrimination between non-native and native teachers (Derwing et al, 2002). For example, in some countries where immigration has resulted to mixed society, universities and other educational institutions tend to prefer hiring native-speaking lecturers rather than non-native ones, resulting to the non-native lecturers losing their jobs (Lam, 2014). Nesdale and Rooney (1996) reported that speakers were often stereotyped based solely on their accents by their listeners, the researchers' experimental study revealed that once the listeners discovered that the speaker's accent is different from theirs, they categorized the speaker as having lower status regardless of the degree of the speaker's intelligence. Al-Alman et al (2013) also observed that sharing the same accent between the speaker and listeners produce an advantage as against the speaker and listeners having different accents, as the listeners experience higher comprehension rate when they share the same accent with the speaker. The study investigated the effects of accent of speakers in a multimedia tutorial on the participants' learning and attitudes toward the speakers. The above problems arise due to the e-learning environment majorly using international dominant languages as means of communication, however, it has been observed that majority of these languages' users are non-native speakers of the languages (Flege et al, 2003).

4. Literature Review

Automatic computational recognition of human speech has long been of interest to researchers both in academia and industries. For this reason, the field has both a long history and a great deal of current activities. Research into ASR dated back to the early 1950s when the first ASR system was made for single-speaker digit recognition by Bell Laboratories. The system worked by locating formant frequencies, which are manifested as major regions of energy concentration in the power spectrum of each utterance and matching them with patterns (Davis et al, 1952). In 1962, IBM designed its "Shoebox machine" which was able to understand 16 words spoken in English (Tappert et al, 1973).

Over the next decades, speech recognition for vocabulary size increased from about a few hundred of words into several thousand words by using Hidden Markov Model (HMM) technique which has the capability of recognizing large number of words. However, whether speech recognition systems at the time could recognize a thousand words or thousands of words was insignificant, the systems

could only worked with discrete words with high level of words error rate (WER) couple with the inability of the systems to recognize various accents from different speakers. Ever since, different researchers had worked on the accents-based speech recognition systems which have today become part of the e-learning environment. The review of such related works is carried out here.

Frederick et al (1975) created a voice-activated typewriter that converts a spoken utterance into a sequence of letters that could be displayed on a screen. They built speakers-dependent speech recognition system, equipped with a language model and a large vocabulary. The goal was to provide automated telecommunication services to the public. To achieve this, systems were needed that could successfully process input from speakers with different regional accents, without the need for individual speaker training. The utterances are drawn from a corpus of two million words with the vocabulary of roughly 10000 words. The work was able to work on large vocabularies with language model to assist the French-Canadian and English-Canadian speakers. In 1979,

Diller (1979) developed a word verification system for unbounded speakers by using pre-stored English lexical entries. The work was tested for continuous speech recognition. In the system, the duration of the spoken words, the words' formant differences, the change in spectral shape between two adjacent frames and energy which was taken from a band of frequencies were the feature parameters considered. The system recorded 40% accuracy result, although, the system only considered single accent of the speakers for the feature parameters in the work.

Rathi (2001) presented an Accent-dependent HMM-Based Speech Recognizer for American, Australian and British accents on British English language using a universal hybrid system that is trained with data from American, Australian and British accented speakers' speeches. The work used training database that consists of digit string lengths range from 1 to 16 digits that were spoken by 700 speakers (350 male and 350 female) for a total of 2561 valid strings, the testing database has 300 speakers (107 male and 193 female). Zhirong and Tanja (2003) developed non-native spontaneous speech recognition through polyphone decision tree for the Chinese English speakers in China. The work had dataset of recorded audio file of 300 speakers which were fed into the system. The results showed a very low accuracy of the tested speech. The model was seen as one of the early accents-based speech recognition invention with little algorithms available resulting in the poor performance of the model at 34% recognition rate. In the same 2003,

Katagiri (2003) developed an automatic speech recognition model of normal speech based on Artificial Neural Networks. A small size of English vocabulary containing few words spoken from Indian English speakers was used for the system's training and testing. The spectral features of the spoken words were extracted per frame using Cepstral analysis and imported to a feed-forward neural network which uses a back propagation with momentum training algorithm. Shengmin et al (2004) developed a Chinese-English Bilingual Modelling for Cross-Language Speech Recognition using hierarchical phone clustering algorithm.

The work used phone modelling for cross-language speech recognition using both Chinese and English as communication languages. The training corpus consists of a Chinese speech database of DB863 and an English speech database of Wall Street Journal (WSJ0). DB863 is a continuous speech recognition corpus of 54 hours male speeches and 57 hours female speeches with total of 166 speakers.

Yanli et al (2005) presented an accent detection and speech recognition system for shanghai-accented mandarin language (the most spoken language in the world) to improve the education system of the shanghai-accented speakers whose mother tongue is Wu language with mandarin as second and official language. The work used Gaussian Mixture Model (GMM) for classification of the accent-based speech and Mel Frequency Cepstral Coefficients (MFCC) as features extraction vector for both training and testing. It used spontaneous speech data collected from 50 male and 50 female speakers of shanghai-accented mandarin language.

Deshpande et al (2005) developed an accent classification of speech on English language using American accent as the standard native accent and Indian accent as the non-native accents. The model used Gaussian Mixture Model (GMM) for the accent classification. The authors used data corpus that was collected in a quiet setting using a head mounted microphone with 76 files from 40 male speakers of native American accent group and 72 files from 36 male speakers of native Indian accent group. Huang and Wu (2007) developed a Hidden Markov Model based system for the generation of phonetic units for mixed language speech recognition based on acoustic and contextual analysis. The authors used characteristics of multilingual phonetic units to create a language model for two different languages of English and Mandarin. Universal phone models were generated for English and Mandarin by understanding phone characteristics from these languages independently. The work made use of prepared 600 recording sheets of a bilingual corpus collection with every sheet containing 80 reading sentences.

Dimitra et al (2010) presented automatic speech recognition of multiple accented English data to investigate the effect of multiple accents on an English broadcast news recognition system based on HMM-adapted accent dependent models in conjunction with a GMM accent classifier. The model made use of a multi-accented English corpus from six (6) different English-speaking regions of US, Great Britain, Australia, North Africa, Middle East and India for training and testing purposes with the proportion male speakers in the range of 40-70%. Furthermore, the study by Maheswari et al. (2010) produced a hybrid model of neural network for speakers' independent word recognition using a combined framework of statistical and neural network-based classifiers of Radial Basics Function and the Pattern Matching method. The model was trained with British English words consisting of 50 words spoken by 20 male speakers and 20 female speakers of India origin. The model made use of the British English language as the communication medium, reflecting the international acceptability and usage of the language in the educational environment.

Ming-liang and Biao (2012) developed a

Chinese speech identification system using Spectral Clustering-Gaussian Mixture Model (SC-GMM). The work adopts Expectation Maximization (EM) algorithm to train Gaussian Mixture Model (GMM) parameters while applying spectral Clustering to initialize GMM parameters. The authors used a Chinese language database and speakers from the four regions of North-China, Wu, Guangdong and Fujian with each having an average of 20 male and female speakers. Speech materials include self-introduction (consists of age, sex, occupation and contact address), introduction of hometown special products, tourist places, people and culture.

In 2013, an automatic speech recognition system for Bangla language based on Hidden Markov Model(HMM) and Mel Cepstrum Frequency Coefficients (MFCCs) feature extraction vector was developed (Akkas et al, 2013). The authors used a medium sized Bangla speech corpus consisting of 100 sentences spoken by 30 male and 30 female speakers from different accents background for the model. The model testing was done using a test speech corpus of 10 speakers from different accents background. In the work of Yating et al (2013), a speech recognition system on Uyghur accent spoken language, an official language of the Xinjiang Uyghur Autonomous Region in China was developed. Uyghur language is widely use in both social and official spheres, as well as in print, radio and television in China, it is also a common language by other ethnic minorities in Xinjiang. The work was based on Hidden Markov Model Toolkit (HTK) while using Mel Frequency Cepstral Coefficient (MFCC) for parametric features extraction. The model was trained with over 1500 utterances from different speakers selected from three different Chinese regions of Northwest, Southern and Eastern accents.

The research by Zhao et al (2014) produced an Acoustic feature of mandarin monophthongs by Tibetan speakers of Chinese people using experimental phonetic approach to examine the influence of accent on the speaking of Mandarin which learners acquired as a second language, especially in the case of the Chinese speakers of minority groups of Tibetans. The work used the speeches from both major Chinese speakers and the Tibetan speakers who were volunteers from Tianjin University with no history of speech, language or hearing disorders with Mandarin proficiency test scored 87% and above as case studies. The results produced an accuracy of 85% for the major Chinese speakers and accuracy of 68% for the minority Tibetan speakers.

Yang and Yu (2015) presented an acoustic analysis of Tibetan Lhasa language by analysing the acoustic characteristics of the Tibetan Lhasa language vowels using the theory of acoustic phonetics based on vowel formant frequency, acoustic vowel chart and fundamental frequency time. The work analysed the effects of different Tibetan accents on the speech recognition of Lhasa language using the fundamental frequency, vowel duration and pitch length as features. The model provides high accuracy for different accents of Tibetan speakers for Lhasa language, leading to an increase in the use of the language in the e-learning environment in China to educate the local Tibetan.

Harpreet and Rekha (2015) developed a Speech Recognition System for Punjabi Language of the South Asia using Hidden Markov Model (HMM) as accents classifier and Mel Cepstrum Frequency Coefficients (MFCCs) for feature extraction. The system was trained using the speech dataset from 97 speakers of Punjabi language from different region of the Southern Asia. The system testing was carried out with the speeches from major and minor speakers from India, Sri Lanka and Pakistan.

Kishori and Ratnadeep (2015) presented an Automatic Speech Recognition system for Marathi language, spoken in western and central India using Artificial Neural Network. The work made use of corpus containing 100 words of names of medicinal plants in Marathi language uttered by 100 speakers of age ranging from 20 to 50 years from Aurangabad region with each speaker given 300 speech samples and making the total number of speech samples become 30,000. The speech features were extracted using Discrete Wavelet Transforms (DWT) while the feature vector set obtained are classified using Artificial Neural Networks (ANN). Mannepalli et al (2015) developed a MFCC-GMM accent-based speech recognition system for Telugu speech signals. Telugu is an Indian language which is widely spoken in Southern part of India with different accents of Coastal Andhra (CA), Rayalaseema (RS) and Telangana (TG). The model used Gaussian Mixture Model (GMM) for classification of the speech based on accents and Mel Frequency Cepstral Coefficients (MFCC) as features extraction vector for both training and test samples. The authors used dataset samples of speeches from 117 native speakers of different accents of Telugu language from Telangana, coastal Andhra and Rayalaseema for both training and testing. The model increases the acceptability of Telugu language as medium of communication in the educational sector in the southern part of India.

Zhenhao (2015) developed an improved accent classification by combining phonetic vowels with acoustic features using phonetic characteristics and enhanced acoustic features for accent classification of seven major types of accented speech from the Foreign Accented English (FAE) of Arabic, French, Brazilian Portuguese, Mandarin, Russian, German and Hindi from Linguistic Data Consortium (LDC) with catalogue number LDC2007S08 which is one of the most comprehensive accented English speech database currently available that contains 4925 sentences of 23 types of accents. The work analysed acoustic characteristics of vowels in each accent based on Gaussian Mixture Model-Universal Background Model (GMM-UBM). The features are further optimized by Principle Component Analysis(PCA) and Heteroscedastic Linear Discriminant Analysis (HLDA).

Furthermore, Ville (2015) presented a state sponsored project of an Automatic Speech Recognition system for English language using Deep Neural Network termed “Foreign Accent Recognition Project”. A corpus of English language dataset with six accents of Russian, Hindi, Cantonese, Korean, Vietnamese, Japanese and Thai was used. The study conducted 5.57 hours of training and 0.52 hours of validation while using deep neural

network as the accents classifier and the Mel Frequency Cepstral Coefficient (MFCC) as the speech features extraction set.

Sivaranjani and Bharathi (2016) produced a continuous speech recognition system for Tamil Language based on the Hidden Markov Model for the acoustic modelling and MFCC extractive technique. The work made use of language model that provides context to distinguish between words and phrases that sound similar and dictionary model that captures the pronunciation of similar words or phrases differently by different speakers. A dataset (speeches) from 20 speakers of native and non-native Tamil language spoken by the people of India and Sri Lanka was used. Desai and Vishvjit (2016) presented Neural Network based Gujarati Speech Recognition for the Gujarati speakers of the state of Gujarati in India for local educational purpose. The work followed Artificial Neural Network and MFCC extractive vectors. It made use of dataset collected by in-ear microphone recorded up to 2.3kHz-4kHz Noise hum - 90dB around 200Hz from different speakers in India. The system has being in use at the state of Gujarati and entire India for e-learning purpose.

David and Viktor (2017) developed a Swedish language classification system using deep learning techniques to identify the accents of speakers and appropriately recognize the speech from such speakers. The work used the SweDia2000 database, a research database containing Swedish speech recordings. The data consists of interviews with people at over one hundred different locations in Sweden and Swedish-speaking parts of Finland and Norway with an average of 12 persons from each location area. The model was able to improve the Swedish language recognition from different accents to promote national and international educational activities in Sweden.

Sruti and Sanjib (2018) presented Speech recognition model with reference to Assamese language using novel fusion technique by combining the Hidden Markov Model (HMM), Vector Quantization and I-vector techniques together for speech classification while using the Mel Frequency Cepstral Coefficient (MFCC) as the speech features extraction set. The aim of the work was to develop an automatic speech recognition system for the Assamese language, spoken by the Assam tribe of India for educational purpose as a way of promoting the usage of the local language among the Indian population. The authors used training data from 150 Indian Assamese speakers with 1500 speech (vocabularies) samples while the test data consists of speech utterances for each unique word spoken by both male and female speakers of Assamese language.

Upadhyay and Lui (2018) presented a foreign English accent classification system using Deep Neural Networks for foreign accented English speeches. The authors made use of spoken English corpus consisting of 30 speakers from 6 different countries of China, India, France, Germany, Turkey and Spain with distinct accent (5 speakers from each country) to train and test the system. The work used MFCC as feature extraction vectors and Deep Neural network (DNN) as speech classifier. The dataset was developed using the birth place and current position of the speakers. Ying et al (2019) produced Chinese accent-based

Sichuan dialect (Language) using Mel frequency cepstrum coefficients (MFCC) as the speech features extractor and Hidden Markow Model-Long Short-Term Memory (HMM-LSTM) as the speech's processor. A Sichuan dialect dataset of characters and their common phonemes sequences was created for the system to identify polyphone and special pronunciation vocabularies in Sichuan.

Tan et al (2021) worked on the accented English speech recognition using Mel frequency cepstrum coefficients (MFCC) as the speech features extractor and Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) as the speech classifier. A dataset of 160 hours accented English data from auxiliary Librispeech dataset were used for training and a test time augmentation and embedding fusion scheme was used to improve the system performance. Tawaqal and Suyanto (2021) developed an accented Indonesia language for five accents of Javanese, Sundanese, Banjar, Bugis and Malay using Mel frequency cepstrum coefficients (MFCC) as the speech pre-processor and a deep recurrent neural network (DRNN) as the speech classifier. The work used dataset developed by recording 500 speakers using a wireless microphone and Adobe Audition software. Each accent has 100 speakers: 50 males and 50 females. They also have five age categories. Early Adolescence: 12-16 years, Late Adolescence: 17-25 years, Early Adults: 26 - 35 years, Late Adult: 36 - 45, and Seniors: more than 46 years (all the categories have 10 people in each accent and gender).

Table 1: The Survey of Accent–Based Speech Recognition Models

S/N	Authors	Language	Accent Covered	Feature Extraction Method	Technique Used
1	Frederick et al, 1975	English	Canadian	Statistical Analysis	Statistical
2	Diller (1979)	British English	Britain	Statistical Analysis	GMM
3	Rathi (2001)	British English	Australian, British and American	MFCC	HMM
4	Zhirong and Tanja (2003)	English	China	Polyphone Decision Tree	HMM
5	Katagiri (2003)	English	India	Cepstral Analysis	ANN
6	Shengmin et al, (2004)	English	China	Hierarchical Clustering Algorithm	HMM
7	Yanli et al, (2005)	Mandarin	Shanghai	MFCC	GMM
8	Deshpande et al, (2005)	American English	American, Indian	MFCC	GMM
9	Huang and Wu (2007)	English and Mandarin	Taiwan	MFCC	HMM

10	Dimitra et al, (2010)	American English	US, Great Britain, Australia, North Africa, Middle East and India	Acoustic Model Adaptation	GMM-HMM
11	Maheswari et al, (2010)	British English	India	Radial Basis Function	ANN
12	Ming-liang and Biao (2012)	Chinese	Mandarin, Shanghai, Guangdong and Fujian	EM-MFCC	SC-GMM
13	Akkas et al, (2013)	Bangla	Indian	MFCC	HMM
14	Yating et al, (2013)	Uyghur	China	MFCC	HMM-TK
15	Ferrer et al, (2014)	English	Britain, Japan	MFCC	GMM
16	Zhao et al, (2014)	Mandarin	China, Tibetan	MFCC	HMM
17	Yang and Yu (2015)	Lhasa	Tibetan (China)	MFCC	HMM
18	Harpreet and Rekha (2015)	Punjabi	India, Sri Lanka and Pakistan	MFCC	HMM
19	Kishori and Ratnadeep (2015)	Marathi	Aurangabad (India)	DWT	ANN
20	Mannepalli et al, (2015)	Telugu	Coastal Andhra, Telangana and Rayalaseema	MFCC	GMM
21	Zhenhao (2015)	English	Arabic, French, Brazilian Portuguese, Mandarin, Russian, German and Hindi	PCA- HLDA	GMM-UBM
22	Ville (2015)	English	Russian, Hindi, Cantonese, Korean, Vietnamese, Japanese and Thai	MFCC	DNN
23	Sivaranjani and Bharathi (2016)	Tamil	India, Sri Lanka	MFCC	HMM
24	Desai and Vishvjit (2016)	Gujarati	India	MFCC RCCC	ANN
25	David and Viktor (2017)	Swedish	Sweden, Finland	MFCC	DNN
26	Sruti and Sanjib (2018)	Assamese	India	MFCC	HMM I-Vector

27	Upadhyay and Lui (2018)	English	China, India, France, Germany, Turkey and Spain	MFCC	DNN
28	Ying et al (2019)	Sichuan	Chinese	MFCC	HMM-LSTM
29	Tan et al (2021)	English (Librispeech corpus)	Chinese	MFCC	CNN-LSTM
30	Tawaqal and Suyanto (2021)	Indonesia language (Bahasa Indonesia)	Javanese, Sundanese, Banjar, Bugis, and Malay	MFCC	DRNN

Table 2: The Reviewed Models' Adopted Language, Accents, Features Extractors and Classification Techniques

Adopted Languages	Covered Accents	Features Extractor	Classification Technique
English, Mandarin, Chinese, Bangla, Uyghur, Lhasa, Punjabi, Marathi, Telugu, Tamil, Gujarati, Swedish, Assamese, Sichuan, Bahasa Indonesia.	Canadian, Britain, Australian, American, Chinese (General), Indian (General), Shanghai, Taiwan, North Africa, Mandarin, Guangdong, Fujian, Japan, Tibetan, Sri Lanka, Pakistan, Aurangabad (India), Coastal Andhra, Telangana, Rayalaseema, Arabic, French, Brazilian Portuguese, Russian, German, Hindi, Russian, Cantonese, Korean, Vietnamese, Japanese, Thai, Sweden, Finland, Turkey, Spain, Javanese, Sundanese, Banjar, Bugis, and Malay.	Statistical Analysis, Polyphone Decision Tree, Cepstral Analysis, Hierarchical Clustering Algorithm, MFCC, Acoustic Model Adaptation, Radial Basis Function, DWT, PCA-HLDA, EM-MFCC.	Statistical, HMM, GMM, ANN, GMM-HMM, SC-GMM, HMM-TK, GMM-UBM, HMM I-Vector, DNN, HMM-LSTM, CNN-LSTM, DRNN.

5. Conclusion

The adoption of e-learning as the act of disseminating knowledge and removing the access restrictions to learning is rapidly increasing, so also the effects of the various accents among the participants of the e-learning cannot be overemphasized. The introduction of the accents-based automatic speech recognition systems has been a great invention in reducing the degradation effects of the varying accents. However, the observation from this review is that, the current accents-based automatic speech recognition models do not accommodate the nuanced vocalized accent problems of less technologically resourced regions, such as Africa and Central America continents with various accents from different regions of the continents. The reviewed works made use of 15 different languages and 41 accents from American, European and Asian countries. About 50% of the reviewed models adopted English language, 46.50% and 3.50% adopted various Asian languages and Swedish language respectively. The review establishes the need for development of ASR models that takes into consideration the regional

peculiarities of the African continent and other less digital regions of the world for a more inclusive and globalized e-learning experience, as the current works so far seems to erroneously assume that the speakers from these regions are not much of a factor in e-learning or that problems associated with accents or phonetic aberrations do not apply to the speakers of these regions.

References

Akkas A, Manwar H, Mohammad NB (2013). Automatic Speech Recognition Technique for Bangla Words. International Journal of Advanced Science and Technology, Vol. 50, pp.51-60.

Akorful V (2014). The Role of E-Learning: The Advantages and Disadvantages of Its Adoption in Higher Education. International Journal of Education and Research, Vol. 2, No. 12, pp. 397-410.

Al-Alman A, Asassfeh S, Al-Shboul Y (2013). EFL Learners' Listening Comprehension and Awareness of Metacognitive Strategies: How Are They Related? International EducationStudies. Vol. 6, No. 9, pp. 31-39.

Algahtani AF (2011). Evaluating the Effectiveness of the E-learning Experience in SomeUniversities in Saudi Arabia from Male Students' Perceptions, Durham, Durham University Press. Available: http://etheses.dur.ac.uk/3215/

Benzeghiba M, De Mori R, Deroo O (2007). Automatic Speech recognition and Speech Variability: A Review. Speech Communication, Vol. 49, pp. 763–786.

Bocklet T, Maier A, Bauer JG et al (2008). Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2008), Las Vegas. pp. 1605–1608.

Crystal, D (2003a). English as a global language (2nd Edition.). Cambridge: Cambridge University Press.

David L, Viktor B (2017). Swedish Dialect Classification Using Artificial Neural Networks and Gaussian Mixture Models. Chalmers University of Technology, Division of Applied Mathematics and Statistics. August, 2017.

Davis K, Biddulph R, Balashek S (1952). Automatic Recognition of Spoken Digits. The Journal of the Acoustical Society of America, Vol. 24, No. 6, pp. 637–642.

Deepali P, Wadhai VM, Thakare VM (2016). E-Learning System and Higher Education. International Journal of Computer Science and Mobile Computing, Vol.5, Issue.2, pp. 274-280.

Derwing TM, Rossiter MJ, Munro MJ (2002). Teaching Native Speakers to Listen to Foreign-Accented Speeches. Journal of Multilingual and Multicultural Development, Vol. 23, pp. 245–259.

Desai V, Vishvjit KT (2016). Neural Network based Gujarati Speech Recognition for Dataset Collected by in-ear Microphone. Proceedings of 6th International Conference onAdvances in Computing and Communications (ICACC, 2016), Cochin, ISSN: 1877-0509, pp.668-675.

Deshpande S, Chikkerur S, Govindaraju

V (2005). Accent Classification in Speech. Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), pp. 139–143

Diller T (1979). Phonetic word verification. Proceedings of IEEE International Conference in Acoustics, Speech and SignalProcessing (ICASSP'79). Vol. 4, pp. 256–261

Dimitra V, Lori L, Jean-Luc G (2010). Automatic Speech Recognition of Multiple Accented English Data. Proceedings of the 11th Annual International Conference of the International Speech Communication Association, Chiba, Japan. pp. 1652-1655.

Flege JE (1984). The detection of French accent by American listeners. Journal of the Acoustical Society of America, Vol. 76, pp. 692–707.

Flege J.E, Schirru C, MacKay IRA (2003). Interaction between the Native And Second Language Phonetic Subsystems. Journal of Speech Communication. Vol. 40, No. 4, pp. 467–491.

Frederick J, Lalit RB, Robert LM (1975). Design of a Linguistic Statistical Decoder forthe Recognition of Continuous Speech. IEEE Transactions on Information Theory, Vol. IT-21, No. 3, pp. 250-256

Fromkin V, Rodman R, Hyams N (2013). An Introduction to Language, 10th Edition. Boston, Cengage Learning Publishing.

Grosjean F (2010). Bilingual: Life and Reality. Cambridge, Harvard University Press.

Harpreet K, Rekha B (2015). Speech Recognition System for Punjabi Language. International Journal of Advanced Research inComputer Science and Software Engineering, Vol. 5, Issue 8, ISSN: 2277 128X, pp. 566-573.

Huang CL, Wu CH (2007). Generation of Phonetic Units for Mixed Language Speech Recognition Based on Acoustic and Contextual Analysis. IEEE Transactions on Computers. Vol. 56, pp. 1218-1225.

Huang YM, Chiu PS (2014). The effectiveness of a meaningful learning-based evaluation model for context-aware mobile learning. British Journal of Educational Technology. Vol. 86. DOI: 10.1111/bjet.12147

Jain AK, Flynn P, Ross AA (2007). Handbook of Biometrics. London, Springer.

Joy OE, Peter MO (2014). Prospects and Challenges of E-Learning with Nigerian Universities. Proceedings of 6th Annual International Conference on ICT for Africa, Yaounde, Cameroon. Vol. 6, pp. 301-312.

Katagiri S (2003). Speech Pattern Recognition Using Neural Networks. Pattern Recognition in Speech and Language Processing, CRC Press. pp. 115-147.

Kinnunen T, Li H (2010). An Overview of Text-Independent Speaker Recognition: From Features to Super

Vectors. Speech Communication, Vol. 52**,** pp.12–40.

Kishori RG, Ratnadeep RD (2015) Automatic Speech Recognition of Marathi isolated words using Neural Network. International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6, No. 5, (ISSN: 0975-9646), pp.4296- 4298

Koolagudi SG, Rao KS (2012). Emotion Recognition from Speech: A Review. International Journal of Speech Technology, Vol.15, pp. 99–117.

Lam VCN (2014). Effects of Speaker's Accent in a Multimedia Tutorial on Non-Native Students' Learning and Attitudes. Southern Illinois University, Carbondale. (PhD Thesis).

Love N, Fry N (2006). Accounting Students' Perceptions of a Virtual Learning Environment: Springboard or Safety Net? International Journal of Accounting Education, Vol. 15, No. 2. pp. 151- 166.

Maheswari NU, Kabilan AP, Venkatesh R (2010) A Hybrid model of Neural Network Approach for Speaker independent Word Recognition. International Journal of Computer Theory and Engineering, Vol.2, No.6, pp. 73-78.

Mai R, Hoffmann S (2014). Accents in Business Communication: An Integrative Model and Propositions for Future Research. Journal of Consumer Psychology, Vol. 24, No.1, pp.137-158.

Mannepalli K, Sastry PN, Suman M (2015). MFCC-GMM Based Accent Recognition System for Telugu Speech Signals. International Journal of Speech Technology, Vol. 19, Issue 1, pp. 87-93

Ming-liang G, Biao AZ (2012). Chinese Dialect Identification Using SC-GMM. Advanced Materials Research, Vols. 433-440, Switzerland, Trans Tech Publications, pp. 3292-3296.

Misra S (2021). A Step by-Step Guide for Choosing Project Topics and Writing Research Papers in ICT Related Disciplines. Communications in Computer and Information Science, Springer International Publishing, Vol., 1350, pp.727-744.

Mohammed Y, Raid AM, Haitham AE (2014). Adaptive E-Learning System Based on Learning Interactivity. International journal of Computer Science and Network Solutions, Vol. 2, No. 4, pp. 9-18

Munro MJ, Derwing T (2006). Modelling Perceptions of the Comprehensibility and Accentedness of L2 Speech: The Role of Speaking Rate. Studies in Second Language Acquisition, Vol. 23, pp. 451-468

Nerbonne J (2003). Linguistic Variation and Computation. Proceedings of the TenthConference on European Chapter of the Association for Computational Linguistics, (EACL, 2003), Budapest, Hungary. Vol. 1, pp. 3-10.

Nesdale D, Rooney R (1996). Evaluations and Stereotyping of Accented Speakers by Preadolescent Children. Journal of Language and Social Psychology, Vol. 15, No. 2, pp. 133-

154.

Organization for Economic Co-Operation and Development (OECD). (2005). E-learning in Tertiary Educatio: Policy Briefs. OECD Publication. http://www.oecd.org/dataoecd/27/35/35991871.pdf [Accessed: 8 June, 2022].

O'Grady W, Archibald J, Aronoff M, Rees-Miller J (2005). Contemporary Linguistics: An Introduction. (5th edition). Boston, Bedford/St. Martin's.

Olabode O, Fasoranbaku AO, Oluwadare AS (2015). Adoption of E-Learning Technology in Nigerian Tertiary Institution of Learning. British Journal of Applied Science and Technology, Vol. 10, No. 2, pp. 1-15.

Rathi C (2001). Accent-Independent Universal HMM-Based Speech Recognizer for American, Australian and British English. Proceedings of the 7th European Conference on Speech Communication and Technology (ECSCT, 2001), Aalborg, Denmark. Vol. 3.

Shengmin Y, Shuwu Z, Bo X (2004). "Chinese-English Bilingual Phone Modelling for Cross-Language Speech Recognition". In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 2004). Vol. 1, pp. 917–920.

Sivaranjani C, Bharathi B (2016). Syllable Based Continuous Speech Recognition for Tamil Language. International Journal of Advanced Engineering Technology, Vol. 7, Issue 1, (E-ISSN 0976-3945), pp. 01-04.

Sruti SB, Sanjib KK (2018). Speech Recognition with Reference to Assamese Language Using Novel Fusion Technique. International Journal of Speech Technology, Vol. 21, No. 2. pp. 251-263

Suzanne R (2015). The Global Extinction of Languages and Its Consequences for Cultural Diversity. University of Oxford. Switzerland, Springer International Publishing.

Taiwo AA, Downe GA (2013). The theory of user acceptance and use of technology (UTAUT): A Meta-Analytic Review of Empirical Findings. Journal of Theoretical and Applied Information Technology, Vol. 49, No. 1. E-ISSN: 1817-3195

Tan T, Lu Y, Ma R, Zhu S, Guo J, Qian Y (2021). AISpeech-SJTU ASR System for the Accented English Speech Recognition Challenge. Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2021). Toronto, Canada, pp. 346-352

Tappert CC, Dixon NR, Rabinowitz AS (1973). Application of Sequential Decoding for Converting Phonetic to Graphemic Representation in Automatic Recognition of Continuous Speech. IEEE Transaction Audio Electroacoustic, Vol. AU-21, pp. 225-228.

Tawaqal B, Suyanto S (2021). Recognizing Five Major Dialects in Indonesia Based on MFCC and DRNN. Journal of Physics, Vol.

1844 – 012003.

Toledano DT, Fernández-Gallego MP, Lozano-Diez A (2018). Multi-Resolution Speech Analysis for Automatic Speech Recognition Using Deep Neural Networks: Experiments on TIMIT. PLoS ONE 13(10): e0205355.

Upadhyay R, Lui S (2018). Foreign English Accent Classification Using Deep Belief Networks. Proceedings of the IEEE 12th International Conference on Semantic Computing (ICSC, 2018). Laguna Hills, California. pp. 290-293

Ville H (2015). Foreign Accent Recognition Project (FARP). MATINE-Finland Ministry of Defence, 2015/2500M-0036. Helsinki. ISBN 978-951-25-2758-8 (PDF).

Yang L, Yu H (2015) The Five Main Vowels Acoustic Analysis of Tibetan Lhasa Dialect. Proceedings of Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), pp. 1104–1108

Yanli Z, Richard S, Liang G et al (2005). Accent Detection and Speech Recognition for Shanghai-Accented Mandarin. In Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH, 2005), Lisbon, Portugal. pp. 217-220

Yating Y, Bo M, Xinyu T, Osman T (2013). Speech Recognition Research on Uyghur Accent Spoken Language. Proceedings of 2013 International Conference on Asian Language Processing, pp. 163-166.

Ying W, Zhang L, Deng H (2019). Sichuan Dialect Speech Recognition with Deep LSTM Network. Frontiers of Computer Science. Vol. 14, pp. 378-387

Zhao L, Feng H, Wang H, Dang J (2014). Acoustic Features of Mandarin Monophthongs by Tibetan Speakers. Proceedings of the IEEE International

Conference on Asian Language Processing (IALP 2014), pp. 147–150

Zhenhao G (2015). Improved Accent Classification Combining Phonetic Vowels with Acoustic Features. Proceedings of IEEE 8th International Congress on Image and Signal Processing (CISP 2015). pp. 1204–1209.

Zhirong W, Tanja S (2003). Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization. In the Proceedings of 8th European Conference on Speech Communication and Technology, Geneva, Vol. 9, (ISSN 1018-4074), pp.1449-1452.

Zhirong W, Tanja S (2003). Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization. In the Proceedings of 8th European Conference on Speech Communication and Technology, Geneva, Vol. 9, (ISSN 1018-4074), pp.1449-1452.