Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)
IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1

AI based Speech Language Therapy using Speech Quality Parameters for Aphasia Person: A Comprehensive Review

Jothi K R
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
jothi.kr@vit.ac.in

Sivaraju S S
Department of EEE,
RVS College of Engineering and Technology, Coimbatore, India.
sssivaraju@gmail.com

Priyanka Jayant Yawalkar
School of Computer Science and Engineering, *Vellore Institute of Technology* Vellore, India
. priyanka.yawalkar2019@vitstudent.ac.in

***Abstract* — Aphasia is a communication disorder that hinders the ability to express and communicate with society. Aphasia individual can face difficulties in sentence formation, opinions and trouble with pronunciation. In this survey paper, various techniques have been reviewed based on previous researches including diagnosis and treatment. Automatic diagnosis can evaluate the severity level of aphasia disorder. The important feature of this disorder is that it will not affect their intelligence. This motivated to come up with a speech assessment system which will evaluate speech based on acoustic features. Further, speech will classify into Aphasia quotient score (0 - 100). This system will analyse aphasic speech and provides feedback to the patients. This will help them to improve their speaking and reading skills.**

Keywords— Aphasia, Cerebrum, Speech Language Pathologists, Automatic Speech Assessment, Speech Quality, Aphasia Quotient

I. INTRODUCTION

According to survey done by NIDCD in the United States, more than 1 million cases and in India approx. 3 million cases affected by aphasia disorder. After every 240 people, 1 person may have this disorder. According to GBD (Global Burden of Disease) study, 6.5 million people died and 25.7 million individuals survived with stroke. Mostly, this disorder experience by children's and age above 60. Aphasia are of two types: fluent speech and non-fluent speech. Non fluent kind of aphasia produce slow, effortful and halting speech. Aphasia with fluent speech can produce sentence effortlessly but they speak with grammatical errors and also makes unstructured sentences. It depends on severity of aphasia. Aphasia disorder is caused due to stroke, injury in brain, brain tumor or progressive neurological diseases. The major common cause is stroke. Due to which person may survive with aphasia. The rate of Aphasia disorder is very high compared to other diseases like cerebral palsy and Parkinson disease. Speech produced by person with aphasia are difficult to recognize because of repeated words, long pauses, omission of words, framing of unstructured sentences etc. These difficulties hinder the person's ability in writing, Speaking, reading and listening skills. Person with Aphasia mostly experience's social isolation because of communication problems, unable to express feelings and thoughts and frustration. In rehabilitation process, language is very complex and crucial function of human 's cerebrum that are supported by network of the neurons. So, Various patterns of injuries to the brain may result in aphasia syndrome.

After diagnosis of severity level, person with aphasia need treatment therapy that must be given by pathologists. The treatment given by SPL are very costly and it should be scheduled at consistent period. But it is difficult for many people to take therapy consistently because of financial condition, less resources of treatment, travelling cost and some health-related problems due to aphasia. The major part is that this disorder doesn't affects their intelligence. Person with aphasia may also have innovative thoughts that they wanted to share with society.

This motivated to come up with a speech assessment system which will assess the speech quality of PWAs. Role of machine learning is to classify aphasia severity using different algorithms which gives best results. Model learns from data and based on past history and classifies new data. Deep learning models classifies large dataset based on neural network. Deep learning models have capability to extracts feature automatically. In this survey paper, different Machine Learning and Deep Learning techniques have reviewed for Automatic Speech assessment system which will evaluate Aphasia person's speech based on phone level parameters. Phone level and word level features will be used for evaluation. Based on these parameters, speech will be classified into Aphasia quotient score (0 - 100). This system will analyse aphasic speech and provides feedback to the patients

II. BACKGROUND RELATED WORK

Previous research work has been done on automatic diagnosing an aphasia disorder by Kohlschein et al., (2017). Automatic detection and diagnosis of aphasia disorder person can give benefits to medical field for easy and fast diagnostics. In this paper, Aachen Aphasia database have used. It contains transcripts and recordings. To preprocess the dataset, overlap detection and speaker diarization were applied on spontaneous spoken recordings. [1].

Kohlschein.et al, (2018). From 20 years, University hospital takes Aachen Aphasia Test (AAT) sessions and collected aphasia speech dataset. The Aachen Aphasia Test (AAT) test conducted in Germany and collects dataset named as UKA AAT DB. Aachen Aphasia Test is a chain of tests that contains talk with patients, two hours of course for read and write and results were assessed by therapist for 6 hours [2]. In this paper, dataset had collected, cleaning and processing techniques applied over data. To classify the aphasia types, LSTM model were trained with 30 Aachen Aphasia Test (AAT) for every group [2].

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

Aphasic speech corpus (GREECAD) had developed by Varlokosta et al. (2016), in Greek. The aim of this paper was to collect Greek Aphasic dataset with transcription and annotation of spoken. These speech samples were annotated and transcribed using ELAN toolkit [3].

Upgraded Traditional treatment for aphasia involves One to one therapy with trained SLPs. Due to Lack of feedback, Le et al., (2014) were developed an intelligent system capable of automatic feedback to patients. First, Le et al., take data from patients by using mobile applications. There are four criteria for evaluation: Clarity, Fluidity, Efforts and Prosody. Le et al., mRMR used feature selection algorithm to avoid overfitting [4].

Le et al. (2017), presented a study of detecting neologistics and phonemic paraphasia's from AphasiaBank scripted dataset. AphasiaBank dataset contains spontaneous speech approximately of 126 hr. Le al. [5] proposed an ASR system using languages model for automatic speech transcription. Acoustics model (DBLSTM-RNN) is used for detecting phonetic and word boundaries through force alignment with transcripts. Set of features were considered for analysis speech qualities are Goodness of pronunciations, phone edit distance, duration measures etc. to classify paraphasia, decision tree, support vector machine and logistic regression algorithm used [5].

Rouvier et al., (2016), proposed speaker embedding method using deep learning for the Speaker Diariazation. These features were collected from hidden layers of DNN. The model was trained to recognize speaker identity. ETAPE French Broadcast news dataset was used for this experiment. Diariazation error rate used as a metric for speaker diariazation [6]. Metrics were compared between I vector PLDA, speaker cosine, embedding PLDA for performance of speaker embeddings. By identifying various identities with very a smaller number of hidden units, highly discriminative and compact features were produced. The speaker embedding technique can be adapted in ASR system [6].

Fraser et al., presented an experiment on automatic diagnosis of Primary Progressive Aphasia and its types such as Progressive Non-Fluent Aphasia and Semantic Dementia. Dataset were collected from Speech Language Pathology Department, University of Toronto [7]. Features divided in text features from transcripts and acoustic features from audio file. Feature set were observed using mRMR and p-value methods. Most of the feature selection methods have overviewed by Gerazov et al.,[22]. For classification of types, random forest, naïve Bayes and SVM classifier used with 81 features. Correctly diagnosis rate was across 87.4% and its subtypes up to 75.6% [7].

Qin et al, 2016, presented an ASR system to facilitate linguistic and acoustical analysis of aphasic speech. Dataset were taken from AphasiaBank. Acoustic modeling was implemented using DNN-HMM and GMM-HMM model. SAT (Speaker Adaptive Training) were applied on training and testing audio dataset by fMLLR (feature-space maximum likelihood linear regression) transform in GMM-HMM model. The *Qin et al,* shown reasonable level of accuracy for language and quantifying speech impairments.

Sak et al., (2015) presented an improved performance of the LSTM-RNN model as an acoustic model for huge vocabulary. Sak et al. investigated effect of word acoustics trained model over huge amount of training dataset ranging from 8000 to 90,000 word [9]. Training dataset contained almost 3 million audio files with duration of average 4 seconds. CTC Approach with RNN were used for labeling sequences where association between inputs and target were unknown. LSTM-RNN model shown good performance that out performs DNN [9].

Le et al, 2015, developed an automatic feedback system to the patients for assessing speech quality *Le et al.* collaborated with UMAP for speech assessment. Triphone crossword acoustic model with 2 gaussian mixtures applied on speech dataset. In [10], aphasic and non-aphasic speech were compared to analyze pitch and duration of utterances.

Le et al., (2016), LVCSR (Large Vocabulary Continuous Speech Recognition) established based on English language using deep neural network model [11]. Decoding and training methods were compared between AphasiaBank and UMAP. Gaussian acoustic models were trained with HMM model using tied state parameters. HMM-DNN model trained with and without i-vectors with minimum threshold 0.05%. MFCC and LDA methods applied with i-vector model and at the end decoding method applied [11].

From the past research, it is obvious to explore the specific pathology features. [12] shown results on PWA speech valuation using end to end method. It has higher performance than convectional automatic assessment method.

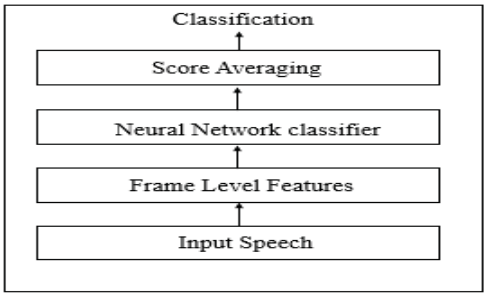

Figure 1: Qin et al., recommended frame work of automatic speech assessment.

Liu et al. (2018) described derived feature from Phone Posterior Probabilities produced by an Automatic Speech Recognition System. Phone Level distortions between impaired and unimpaired speakers can be measured from Kullback-Leibler divergence. Acoustic model architecture in this paper were combination of TDNN and BLSTM. CanPEV, CUSENT, AphasiaBank Corpora used for acoustic model. Each speaker had given Aphasia quotient score in 100 scale in a dataset. Frame level targets obtained from CD-GMM-HMM for corpus where the dataset is trained to generate State level phone Alignment. Conventional Acoustic Features were evaluated under eGeMAPS technique in OpenSMILE toolkit. For Aphasic Speech Assessment, polynomial SVM classifier with 5-fold cross validation were used. [13].

Silva et al., (2017) introduced semi-Automatic aphasia classification and diagnosis using pattern matching and feature extraction technique of audio processing. DSP (Digital signal processing) used for repetition tasks. Three diagnostic components taken into consideration such as word repetition, confrontation naming and comprehension analysis. For feature extraction, methods listed as LPCC (Linear Predictive Cepstral Coefficient), PLP (Perceptual

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

Linear Prediction) and MFCC (Mel-Frequency Cepstral Coefficient) that are commonly used for feature extraction from audio signals. For pattern matching, Dynamic time warping method is widely used which determines the aphasia speech signal compered original signal [14].

Teodoro et al., proposed a virtual clinician's system to help aphasia individual to practice self-initiated speeches in day to day communication. Trials splited in three categories: post-treatment, treatment and pre-treatment. The data were collected from recorded session between clinician and aphasia individual [15]. Discourse, sentence productivity, Lexical content and grammar content were taken into consideration. Expected outcomes from this experiment were efficiency, cost reduction, developing a greater database and maximizing the use of language skills in day to day life [15].

Tan et al., (2016), Investigation is done on voice automatic assessment using Cantonese speech utterances. Deep Neural Network-Hidden Markov Model (DNN-HMM) model were trained for ASR system with unimpaired speech data. Features were extracted using Linear discriminant Analysis (LDA) into 40 dimensions [18]. Using these features DNN-HMM model was trained based upon GMM-HMM system. Orthographic transcription of utterances used for training the bi-gram syllable. SER (Syllable Error Rate) and phone error rate (PER) attained by the DNN-HMM and GMM-HMM system are 7.82% and 10.51% [18].

R. Gupta et al. (2016), challenges of speech production disorders have an effect on fluency, phonation, intonation and aeromechanical components of respiration were involved [19]. Feature design and machine learning algorithms can be used to capture feature patterns. These techniques can offer combination of data driven technique and domain-knowledge technique for analysis of pathological speech. Depending on linguistic knowledge and pathological knowledge, typical prosody computational model can be developed [19].

Guozhen et al. (2015), Describes system for identification of severity of Parkinson's disease. Study describes classification into five levels based on severity and corresponding type as fluency, voice and articulation. Feature extraction were performed with OpenSMILE version 2.0 Four-fold cross validation method were used in weka tool to produce folds on training dataset [20]. *Guozhen et al.,* two stages were described for detecting Parkinson disease from acoustic signal. Unified Parkinson's Disease Rating Scale (UPDRS) contains impaired utterances with severity labels. more features were explored. The improvement is done on features that derived from syllable detection and i- vector features [20].

Qin et al.2018, introduced a method to extract textual features from erroneous automatic speech recognition (ASR) output were developed which was based on word embedding technique [21]. Also, group of Supra-segmental duration features were derived from the syllable level time alignments that were generated. The derived text and duration features and its combination were assessed in classification as well as in prediction of speech assessment score. Here, Cantonese dataset were used from AphasiaBank. DNN-HMM architecture used for acoustic model. 440 dimensional features derived using fMLLR and that transformed from Mel frequency cepstral coefficient (MFCC). Syllable-level transcription. Syllable bigram is a language model trained with transcriptions of all training acoustic speech dataset. Classification algorithms are SVM, binary decision tree and random forest used to classify High-AQ and Low-AQ groups. Overall syllable error rate calculated for aphasic speaker were 48.08% [21].

Balaji V et al. (2019), gave detailed analysis that were carried out from audio waveforms of with and without speech disorder and extracts features from audios. The comparison was carried out based on factors such as duration, PSD, frequency and time etc. Statistical model determines phoneme using MFCC extracted features.

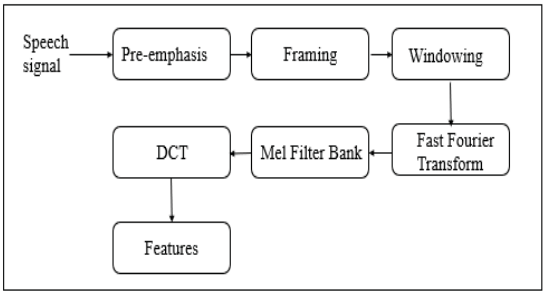

Figure 2: Step by Step MFCC extraction process by [23].

Speech signals are generally temporal signals and need to analyze in sequence of windows (10-20 ms). Also, standard frequency ranges for children from 200-400Hz, females from 150-300Hz, and males from 50-200Hz. For this system, SVM, HMM and hybrid SVM-HMM models recommended. In [23], three approaches were suggested for ASR system such as acoustic phonetic approach, pattern recognition approach and artificial intelligence approach. Mainly five blocks are required as language modelling, acoustic modeling, feature extraction, pronunciation modelling and decoder. Dataset were collected from university of Illions (UASPEECH). The acoustic model used here is GMM-HMM model. This traditional model also used by Xiong et al., (2019) for speech analysis of phoneme features in ASR system [16]. Feifei et al.,2018, LSTM model were used to generate articulatory features to improve accuracy [17]. 16KHZ sampled audio files with 16 bits per sample done using single channel with PCM. Observation from graphs shows that lower the recognition accuracy, higher the rate of disability [23].

A novel feature extraction method was using UAspeech introduced by combining PCA method with traditional Mel frequency cepstral coefficients (MFCC) [24]. This method was tested with HMM model by two measures i.e. accuracy of recognition and time complexity. Further, et al., on UAspeech dataset, BLSTM model were applied which improves 6% accuracy over traditional method [28].

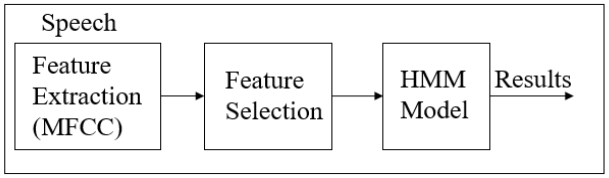

Figure 3: proposed model of ASR by Kalamani et al., 2014 [37].

The procedure of converting audio signal into feature matrix presentation are called as feature extraction of audio signal. Linear Discriminative Analysis (LDA), Principle

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

Component Analysis (PCA), Linear Prediction Coefficients (LPC), Reflection Coefficients (RC), Wavelet, Independent Component Analysis (ICA), MFCC, Linear Predictive Coding (LPC), Filter bank analysis Cepstral analysis are the techniques for feature extraction and widely applied in speech recognition. MFCC technique follows steps as pre-emphasis, apply frame_blocking with overlap samples, windowing method, FFT, Mel-frequency filter bank, cepstrum and delta applied over it. Experimental results of MFCC and PCA achieved in the range of 77.6% to 86.4% as number of mixture increases, accuracy decreases [24].

Qin et al (2019)., introduces an approach towards out of vocabulary words by classifying phone posteriograms into strongly and weakly constraint speech recognizer [25]. Posteriograms explore impaired speech characteristics such as voice abnormality, change in speaker rate. In this paper, Convolutional Neural Network model were used for classification of posteriograms and for prediction of severity of aphasia. All the experiments were done on Cantonese database. Syllable bigram is a language model used to decode the transcription. The overall Syllable Rate Error were about 16.73%. using CNN model binary classification done between severe and mild aphasia attained F1 score up to 89% [25]. CNN model further used by seedahmed et al., (2020) for mapping of relation between three different lucidity parameters and severity of aphasia. Mandarin spoken dataset used in [22]. Lucidity speech features such as fluency, tone and articulation etc.

Adam et al. (2020), presented an acoustic analysis of prosody depending on data that were obtained from Palestinian with broca's aphasia and fluent speakers [26]. Measures for acoustics are word speaking duration, phrase lengthening and syllable duration etc. The research in [26] can contribute to other languages. The results of this study address in four measures such as rate, melodic shape of utterances, duration and pitch values and range etc.

Takashima et al. (2020), introduced a novel approach for automatic speech recognition by using 2-step model-adaption method. Japanese dysarthric speech dataset used in this paper. Input features were extracted using Mel Frequesncy Cepstral Coefficient. Time Delay Neural network (TDNN) with ReLU as activation function used as acoustic model [27]. Out of order vocabulary is ratio of number of words that are not present in lexicon were up to 0.58%. Other models can also be implemented on the same dataset such as adversial training, KL divergence [27].

Zhang et al., proposed a novel way for pronunciation quality evaluation by discrimination technique [29]. GOP tries to differentiate various quality of pronunciation but unable to map probabilities into scores. Segmentation of dataset into phones done by forced alignment by using AM. Phone wise score performance of TPDA and GOP method were compared based on scoring difference and correlation [29].

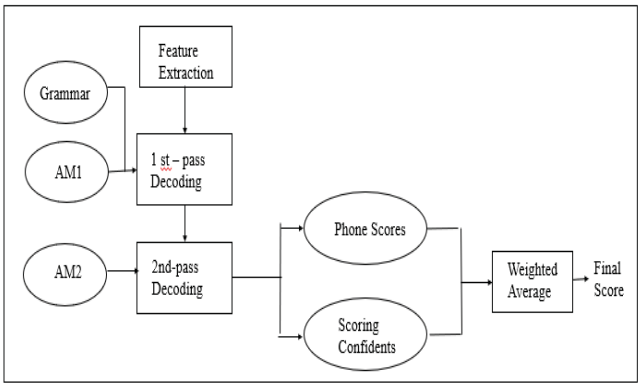

Figure 4: TPDA system flow proposed by Zhang et al.[29]

M. Shahin et al. (2020), focused on outstanding approach used to find the 3 common pronunciation errors made by children [31]. In this paper, main challenges are mentioned during speech processing such as speaker diariazation, inaccurate acoustic model, distortion and unreliability in speeches etc. In this paper, binary classifier was trained using UA speech dataset. The model achieves accuracy up to 87% [31].

Tan et al. (2015), is focused on diction that includes speech rhythm, stress and intonation of words and sentences. This research say that appropriate intonation is essential for conveying the emotions [32]. Tan et al., finds meaningful representation of intonation from acoustic signals. The research is working on four individual persons, two are from anomic & two are from transcortical sensory with different age, gender etc [32].

In this research, Lee et al., used deep belief network (DBN) to find word level mispronunciation. This method works by coordinate a non-native speech with one native speech. TIMIT corpus have used for native English target language. And non-native audio used from CU-CHLOE corpus. Lee et al., finds difference that characterize the amount of mis-alignment from alignment path. In this paper the system works in different DBN training 1) pre-training 2) Fine-Training. After the result of this experiment Lee et al., find that result is improved by 10.4% and system works remain steady. In future if there is larger data set the improve system performance [33]. Main focused of [30] was to enhance pronunciation of speech disordered person. Using DBN model approach, achieved accuracy were 69 to 70.51% [30].

Wilson et al., (2018), The primary aim is that to describe a quick aphasia battery that is to support research into neuroplasticity after language region of brain [34]. There are 3 main features 1. Time efficiency,2. Battery must be psychometrically sound,3. It should be yielding a multi-dimensional characteristic of impaired in past there is no such kind of battery that are used for English language aphasia that are meet these 3 criteria. This research has given described the manufacturing and quality validation of the QAB. [34] bridges the gap between comprehensive battery that are taken lot of time and rapid screening instruments that are provide limited data.

Henry et.al. (2018), reviewed an automatic evaluation of primary progressive aphasia which includes aphasia batteries which was specifically designed for PPA and for stroke as well as for particular cognitive linguistics function.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

speech language pathologists are major aspect of rehabilitative and diagnostic process. Individual with apraxia and dysarthria have difficulty of word repetition due to phonological short-term memory. The aimed of the paper was to evaluate clinical variant and primary progressive aphasia diagnosis and to keep track of recovery progress over period of time [36].

Jokel et al. (2014), reviewed three specific form of primary progressive aphasia were recognized as logopenic, non-fluent/agrammatic and semantic etc. difference between these variants are by grammatical correction, fluency, rate of connected speech as well as status of semantics etc. Repost of PPA severity based on word retrieval classifies as mild, severe and sometimes varied [38].

Christensen et al. (2014), investigated different techniques by evaluating acoustic closeness between individuals and their rankings. American UA Speech corpus is collection of English speeches of dysarthric individuals of about 18 hrs. recorded from 15 speakers [39]. Very rarely large data available to train speaker dependent models. From previous research work, AMI meeting dataset used for training speaker independent models. Utterance can be selected based on likelihood target scores of trained models. Whole acoustic model depends on HMM (Hidden Markov Model) that were trained by using max. likelihood criteria. triphones with GMM (Gaussian Mixture model), state clustered with 16 components were used. Assessment of word Recognition on UA Speech have improved up to 11.5% compared to system that uses whole data and also shows good result against CMLLR-SAT. 60% improvement in the accuracy of intelligible system [39].

Liu et al. (2019), presented an automatic speech recognition for Cantonese aphasia individuals. The DNN-HMM model in system is trained using spontaneous utterances from native speakers. DNN-HMM model for speech recognition were described by Mirsamadi et al.,[48]. The posterior extracted features majorly used for pattern recognition. SVM classifier used for classification of aphasia severity. Accuracy of classification based on two classes were 90.3%. for high severe disorder classification output shows mostly inconsistent result [40].

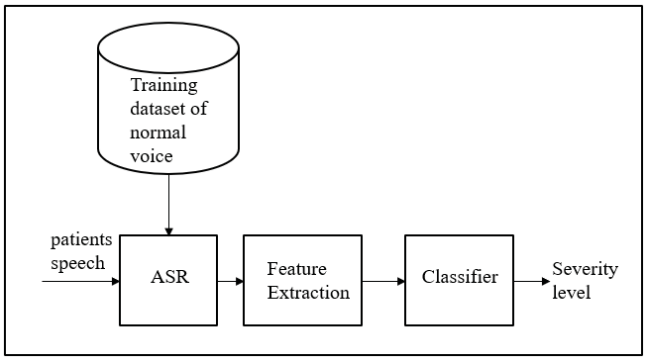

Figure 5: Automatic speech classification workflow [35]

Above workflow proposed by Liu et al., (2017). Normal speech dataset and patient's dataset given to ASR system. Output of ASR system provided to feature extraction. Extracted features were trained using DNN-HMM model to predict severity level [35].

Previous research shown that vocal tract have non-laryngeal tumors which can have negative impact on pronounciation in speech. These can affect some changes in spectral energy, flux, sharpness. The set of acoustic features were extracted using open SMILE toolkit. In this paper, *Huang* et al., proposed a model for automatic detection of pathological voices by using Asymmetric Sparse Kernel Partial Least Square Classifier (ASKPLSC). Extracted spectral features majorly used in ASR system and methods for this process used here is MFCC, RASTA etc. For this novel method, NKICCRT speech dataset was used. Accuracy of this approach achieved to 74% by comparing with UA systemwhich was 68.9% accuracy [42].

Huang et al.(2016), presented every pathologial utterance using three different types of audio signals representation such as Gaussian distribution, cross-correlation matrix and linear subspace etc. For this experiment, NKI CCRT dataset were used. Four different classifiers were used such as logistic regression, kernal Partial Least Squares, Kernal SVM and k-nearest neighbour. Comparison between proposed method and other baseline system study were performed. The accuracy of the combined classifiers obtained upto 78% compared to baseline accuracy of 68.9% [43].

Kim et al.(2015), proposed automated diagnosis system for rate of severity classification of parkinsons disease. Main aimed of [44], was to make combined system for unsupervised clustering method and to detect distortion in PD speech to enhance accuracy of prediction. Supervised algorithm such as Genetic algo.,KNN and SVM used for classifiaction of parkinson's disease described by Abhishek M.S et al., (2020). Accuracy of prediction were upto 85% to 95% [47]. The system includes segmental features, Non-Linear time series, spoken rhythm and i-vector etc. The feature set named as ComPaRE contains LLD(Low level descriptor) that is speech quality, spectral, prosody and speech features. If Root Mean Squared enrgy of frame is greater than 10% of 0.9 quantile in utterance then keep voice otherwise remove it. All the features were extracted from OpenSMILE tool. Kaldi toolkit have used for transcript(phone-sequences) prediction of dataset. Kaldi toolkit were describe in [41]. The acoustic model were trained using Mexican-spanish broadcast news dataset to get the properties of native spandish person's speech. [44].

Kim et al.(2015), proposed a novel appropach to capture non-typical variationin prosody, pronounciation and voice quality in pathological speeches. Also proposed posterior smoothening scheme using post classification and performed feature level fusion. NKI CCRT and TORGO dataset used to test the performmace of the model. Kim et al., evaluated sentence level feature set of individual dataset and fusion of feature level set on both dataset. Proposed smoothnening technique enhanced the prediction of classes over test datasetv[45].

Wang et al. (2019), improved the English word pronunciation system using deep neural network-hidden markov models (DNN-HMMs). This system is based on HTW platform. In result they compared the system scoring with the score given by teacher, the system enters this scoring automatically inside it and compared with manually enter scoring system[46].

Gupta et al. (2015), trained MOE model using enhanced EM (Expectation maximization) depending on annealing gradient ascent procedure. This algorithm increases the performance of intelligible system. Dataset use

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

for this system is NKI CCRT speech data. Jitter, pitch, intensity, voice probability are the features used for classification. Statistical functions like standard deviation, minimum, maximum, mean, quartile, range have used here. This paper aimed to every expert that can map with pathological criteria [49].

Assessment of pathological speeches are done by trained pathologists. The tests are costly and inconsistent. To make reliable and consistent assessment, Berisha et al., proposed a novel data driven approach to existed problem. Berisha et al., have proposed a cost function for analyzing speeches from experiments, this system was recognized acoustic features that mapped with original dataset [50]. 5 aspects of speech characteristics were considered such as voice quality, pronunciation, prosody, nasality and severity. Future work of this paper was to make systemrobust for more noisy data.

III. COMPARATIVE BASED GAP ANALYSIS

Table I. Based on Feature extraction, acoustic and language models.

Author name	**Feature extraction & Dataset used**	**Acoustic & Language model**	**Merits and Demerits**	**Limitation & Future work**
Schmitt *et al., [1]*	open SMILE toolkit Dataset: AAT	LSTM-RNN	Merit: system detects aphasia automatically within less time. Demerit: system unable to classify high level aphasia which contains non-fluent speeches.	Limited to spontaneous speech. Future work is to make use of CNN model.
Le et al.,2014 [4]	Features extracted from transcript Dataset: UMAP	Decision Tree, SVM, Logistic Regression, Naive Bayes, Random Forest	Merits: system provides feedback to the PWA about verbal practice. Demerits: Haven't applied model selection method.	Limitation: system gives better results with few classes of training data. Future work is to consider more phone level features
Fraser et al,[7]	Lu's L2 Syntactic Complexity Analyzer Dataset: Speech Pathology Dept., University Of Toronto	naïve Bayes, random forest and SVM	Merits: automatically detects Semantic Dementia and PPA Demerits: Lack of interaction between feature set and task of diagnostic.	Limitation: NB accuracy affected by mRMR i.e.29%. Future work is to analyze the features during classification process.
Qin et al, 2016 [8]	MFCC Dataset: Aphasia Bank	DNN-HMM and GMM-HMM	Merits: analyses linguistic and acoustical features Merits: recognizing non-fluent and fillers are challenging for system.	Limited to fluent speech Future work: better language model is required to improve performance
Le et al, 2015, [10]	- Dataset: UMAP	Triphone crossword with two GMM	Merits: alternative system for treatment therapy and cost efficient. Demerits: system is not robust when data contains noise.	Limited phone level features have taken. Future work is to improve rhythm features using Envelop based analytical methods.
Le et al., 2016, [11]	MFCC, LDA methods with i-vector Dataset: Aphasia Bank, UMAP	HMM-DNN, HMM-GMM	Merits: system provides Real Time feedback to PWA. Demerits: manual transcriptions are required.	Limitation is i-vector benefited to only high AQ level. Future work: More adaptation techniques depending on diagnoses.
Qin et al.,2018 [12]	MFCC Dataset: Aphasia Bank	GRU-RNN, CNN	Merits: saves manual tasks required for assessment. Demerit: more utterance labels required for accurate recognition.	Limitation: system failed while learning content from utterances. Future work is to improvise accurate recognition by adding more neural network.
Lee et al.,2016 [18]	LDA Dataset: CUSENT, CanPEV	DNN-HMM, syllable bi-gram	Merits: Approach can be used for other languages. Demerits: System l applicable for fluent speech only.	Research is Limited to Cantonese dataset only. Future work is to use of dysphonia features and phone alignments
Qin et al.,2018 [21]	MFCC Dataset: Aphasia Bank	BDT, SVM, RF, CBOW	Merits: ASR system is developed using text features. Demerits: More enhance methodology is needed.	Limitation of CBOW model is it cannot capture the order of word. Future work is to increase audio speech dataset
Balaji V et al., [23]	MFCC Dataset: UA SPEECH	SVM, HMM, hybrid SVM-HMM, GMM-HMM	Merits: maps audio signals between with without speech disorders. Demerits: word set vocabulary were already trained.	Limited dataset is used. Future work is to build system using open set vocabulary

Table. 1. Example comparison based on Feature extraction, acoustic and language models.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

TABLE II. Comparison between Acoustic features, language features and parameters for analysis

Author name	Acoustic Feature set	Language feature set	Parameters for Analysis
Le et al.,2014,[4]	Acoustic Feature: Filler, clear speech, non-speech, vague speech	language features: word level features	SER
Qin et al, 2016, [8]	total number of pauses and speech chunks, average duration of speech chunk and pauses	Number of speeches chunks, syllables and pauses Duration of speech chunk	Syllable error rate (SER), AQ
Le et al.,2015 [10]	duration of vague speech, clear speech and non-speech, total amount of voiced duration	Free form: silences, and speech regions, fillers	WER (Word Error Rate), UAR%
Le. et al., 2016, [12]	Phone level features	No language model has used	Phone error rate (PER)
Lee et al.,2016,[18]	Phone level features	Word level features	PER, SER, pattern matching rate
Qin et al.2018,[21]	Supra segmental duration features	Syllable level features	High-AQ, Low-AQ, SER

Table. 2. Example comparison based on Feature extraction, acoustic and language models.

IV. SUMMARY

In this paper, different techniques have reviewed for Automatic Speech assessment system that classifies severity level of aphasic speech. Dfferent datasets are available with and without transcriptions. Aphasia speech evaluates based on features such as prosody, clear, effect etc. Each features have many sub-features divided in syllable and phoneme level features. For feature extraction, various techniques used for acoustic features. The most common method used for feature extraction is MFCC technique. Both machine learning and deep learning models have used for classification. Recent models such Convolutional Neural Network, Recurrent Neural Networks, BLSTM and Hybrid models gives better results than traditional algorithms. Various parameters such as SER, PER, WER, UAR, AQ have considered for analysis. Results can be improved by adding bigger neural networks. More work is needed for non-spontaneous speeches to make accurate prediction. This System can be made availabe in different target languages.

V. CONCLUSION

In this paper, different techniques are reviewed for feature extraction, feature selection and classification for diagnosis of aphasia disorder and treatment therapy. Different approaches were used for classification of severity level of aphasia based on phone level and syllable level parameter. Various parameters such as syllable error rate, goodness of pronunciation and phone error rate etc., are considered for analysis.

From this survey, the future work is to explore more phone level and word level parameters. Most of the techniques limited to spontaneous speech. We will investigate work on spontaneous and non-spontaneous speeches. Automatic Speech assessment system can be used for other target languages.

REFERENCES

[1] C. Kohlschein, M. Schmitt, B. Schüller, S. Jeschke and C. J. Werner, "A machine learning based system for the automatic evaluation of aphasia speech," 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), Dalian, 2017, pp. 1-6, doi: 10.1109/HealthCom.2017.8210766.

[2] Kohlschein, Christian & Klischies, Daniel & Meisen, Tobias & Schuller, Björn & Werner, Cornelius. (2018). Automatic processing of clinical aphasia data collected during diagnosis sessions: challenges and prospects.

[3] Varlokosta, Spyridoula & Stamouli, Spyridoula & Karasimos, Athanasios & Markopoulos, Georgios & Kakavoulia, Maria & Nerantzini, Michaela & Pantoula, Katerina & Fyndanis, Valantis & Economou, Alexandra & Protopapas, Athanassios. (2016). A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications.

[4] D. Le, K. Licata, E. Mercado, C. Persad and E. M. Provost, "Automatic analysis of speech quality for aphasia treatment," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 4853-4857, doi: 10.1109/ICASSP.2014.6854524.

[5] Le, Duc & Licata, Keli & Mower Provost, Emily. (2017). Automatic Paraphasia Detection from Aphasic Speech: A Preliminary Study. 294-298. 10.21437/Interspeech.2017-626.

[6] M. Rouvier, P. Bousquet and B. Favre, "Speaker diarization through speaker embeddings," 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, 2015, pp. 2082-2086, doi: 10.1109/EUSIPCO.2015.7362751.

[7] Fraser, Kathleen & Rudzicz, Frank & Rochon, Elizabeth. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes.

[8] Y. Qin, T. Lee, A. P. H. Kong and S. P. Law, "Towards automatic assessment of aphasia speech using automatic speech recognition techniques," 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, 2016, pp. 1-4, doi: 10.1109/ISCSLP.2016.7918445.

[9] Sak, H. et al. "Fast and accurate recurrent neural network acoustic models for speech recognition." *INTERSPEECH* (2015).

[10] Le, Duc & Mower Provost, Emily. (2015). Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

[11] Le, Duc & Mower Provost, Emily. (2016). Improving Automatic Recognition of Aphasic Speech with AphasiaBank. 2681-2685. 10.21437/Interspeech.2016-213.

[12] Y. Qin, T. Lee, Y. Wu and A. P. H. Kong, "An End-to-End Approach to Automatic Speech Assessment for People with Aphasia," 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei City, Taiwan, 2018, pp. 66-70, doi: 10.1109/ISCSLP.2018.8706690.

[13] Y. Liu, Y. Qin, S. Feng, T. Lee and P. C. Ching, "Disordered Speech Assessment Using Kullback-Leibler Divergence Features with Multi-Task Acoustic Modeling," 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei City, Taiwan, 2018, pp. 61-65, doi: 10.1109/ISCSLP.2018.8706657.

[14] M. Khan, B. N. Silva, S. H. Ahmed, A. Ahmad, S. Din and H. Song, "You speak, we detect: Quantitative diagnosis of anomic and Wernicke's aphasia using digital signal processing techniques," 2017 IEEE International Conference on Communications (ICC), Paris, 2017, pp. 1-6, doi: 10.1109/ICC.2017.7996967.

[15] G. Teodoro, N. Martin, E. Keshner, J. Y. Shi and A. Rudnicky, "Virtual clinicians for the treatment of aphasia and speech disorders," 2013 International Conference on Virtual Rehabilitation (ICVR), Philadelphia, PA, 2013, pp. 158-159, doi: 10.1109/ICVR.2013.6662079.

[16] Xiong, Feifei & Barker, Jon & Christensen, Heidi. (2019). Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

Personalised Dysarthric Speech Recognition. 10.1109/ICASSP.2019.8683091.

[17] F. Xiong, J. Barker and H. Christensen, "Deep Learning of Articulatory-Based Representations and Applications for Improving Dysarthric Speech Recognition," Speech Communication; 13th ITG-Symposium, Oldenburg, Germany, 2018, pp. 1-5.

[18] Lee, Tan & Yuanyuan, Liu & Yeung, Yu & Law, Thomas & Lee, Kathy. (2016). Predicting Severity of Voice Disorder from DNN-HMM Acoustic Posteriors. 97-101. 10.21437/Interspeech.2016-1098.

[19] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 6470-6474, doi: 10.1109/ICASSP.2016.7472923.

[20] An, Guozhen & Brizan, David Guy & Ma, Min & Morales, Michelle & Syed, Ali & Rosenberg, Andrew. (2015). Automatic Recognition of Unified Parkinson's Disease Rating from Speech with Acoustic, i-Vector and Phonotactic Features.

[21] Y. Qin, T. Lee and A. P. Hin Kong, "Automatic Speech Assessment for Aphasic Patients Based on Syllable-Level Embedding and Supra-Segmental Duration Features," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 5994-5998, doi: 10.1109/ICASSP.2018.8461289.

[22] S. Mahmoud et al., "An Efficient Deep Learning Based Method for Speech Assessment of Mandarin-Speaking Aphasic Patients," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3011104.

[23] V. Balaji and G. Sadashivappa, "Waveform Analysis and Feature Extraction from Speech Data of Dysarthric Persons," 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2019, pp. 955-960, doi: 10.1109/SPIN.2019.8711768.

[24] H. Trang, Tran Hoang Loc and Huynh Bui Hoang Nam, "Proposed combination of PCA and MFCC feature extraction in speech recognition system," 2014 International Conference on Advanced Technologies for Communications (ATC 2014), Hanoi, 2014, pp. 697-702, doi: 10.1109/ATC.2014.7043477.

[25] Y. Qin, T. Lee and A. P. Hin Kong, "Combining Phone Posteriorgrams from Strong and Weak Recognizers for Automatic Speech Assessment of People with Aphasia," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6420-6424, doi: 10.1109/ICASSP.2019.8683835.

[26] Dergipark.org.tr. 2020. [online] Available at: <https://dergipark.org.tr/en/download/article-file/104812> [Accessed 14 September 2020].

[27] R. Takashima, T. Takiguchi and Y. Ariki, "Two-Step Acoustic Model Adaptation for Dysarthric Speech Recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6104-6108, doi: 10.1109/ICASSP40776.2020.9053725.

[28] C. Bhat and H. Strik, "Automatic Assessment of Sentence-Level Dysarthria Intelligibility Using BLSTM," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 322-330, Feb. 2020, doi: 10.1109/JSTSP.2020.2967652.

[29] J. Zhang, F. Pan, B. Dong and Y. Yan, "A novel discriminative method for pronunciation quality assessment," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 8223-8226, doi: 10.1109/ICASSP.2013.6639268.

[30] Christensen, Heidi et al. "Learning speaker-specific pronunciations of disordered speech." *INTERSPEECH* (2013).

[31] M. Shahin, U. Zafar and B. Ahmed, "The Automatic Detection of Speech Disorders in Children: Challenges, Opportunities, and Preliminary Results," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 400-412, Feb. 2020, doi: 10.1109/JSTSP.2019.2959393.

[32] Tan, Lee & Kong, Anthony Pak-Hin & Wang-Kong, Lam. (2015). Measuring prosodic deficits in oral discourse by speakers with fluent aphasia. Frontiers in Psychology. 6. 10.3389/conf.fpsyg.2015.65.00047.

[33] A. Lee, Y. Zhang and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 8227-8231, doi: 10.1109/ICASSP.2013.6639269.

[34] Wilson SM, Eriksson DK, Schneck SM, Lucanie JM. A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function [published correction appears in PLoS One. 2018 Jun 15;13(6):e0199469]. *PLoS One*. 2018;13(2):e0192773. Published 2018 Feb 9. doi:10.1371/journal.pone.0192773

[35] Y. Liu, T. Lee, T. Law and K. Y. Lee, "Acoustical Assessment of Voice Disorder With Continuous Speech Using ASR Posterior Features," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 1047-1059, June 2019, doi: 10.1109/TASLP.2019.2905778.

[36] Henry ML, Grasso SM. Assessment of Individuals with Primary Progressive Aphasia. *Semin Speech Lang*. 2018;39(3):231-241. doi:10.1055/s-0038-1660782

[37] M. Kalamani, S. Valarmathy, C. Poonkuzhali and Catherine J N, "Feature selection algorithms for automatic speech recognition," 2014 International Conference on Computer Communication and Informatics, Coimbatore, 2014, pp. 1-7, doi: 10.1109/ICCCI.2014.6921797.

[38] Raiser, Theresa & Croot, Karen & Nickels, Lyndsey & Taylor, Cathleen & Ackl, Nibal & Wlasich, Elisabeth & Stenglein-Krapf, Gisela & Rominger, Axel & Danek, Adrian. (2014). WORD RETRIEVAL IN PRIMARY PROGRESSIVE APHASIA FOLLOWING LANGUAGE THERAPY. Alzheimer's and Dementia. 10. P916. 10.1016/j.jalz.2014.07.120.

[39] H. Christensen, I. Casanueva, S. Cunningham, P. Green and T. Hain, "Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data," 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, 2014, pp. 254-259, doi: 10.1109/SLT.2014.7078583.

[40] Y. Liu, T. Lee, T. Law and K. Y. Lee, "Acoustical Assessment of Voice Disorder With Continuous Speech Using ASR Posterior Features," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 1047-1059, June 2019, doi: 10.1109/TASLP.2019.2905778.

[41] Povey, Daniel & Ghoshal, Arnab & Boulianne, Gilles & Burget, Lukáš & Glembek, Ondrej & Goel, Nagendra & Hannemann, Mirko & Motlíček, Petr & Qian, Yanmin & Schwarz, Petr & Silovský, Jan & Stemmer, Georg & Vesel, Karel. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

[42] D. Huang, M. Dong and H. Li, "Intelligibility detection of pathological speech using asymmetric sparse kernel partial least squares classifier," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 3744-3748, doi: 10.1109/ICASSP.2014.6854301.

[43] D. Huang, M. Dong and H. Li, "Combining multiple kernel models for automatic intelligibility detection of pathological speech," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 6485-6489, doi: 10.1109/ICASSP.2016.7472926.

[44] Kim, Jangwon & Nasir, Md & Gupta, Rahul & Van Segbroeck, Maarten & Bone, Daniel & Black, Matthew & Skordilis, Zisis & Yang, Zhaojun & Georgiou, Panayiotis & Narayanan, Shrikanth. (2015). Automatic estimation of Parkinson's disease severity from diverse speech tasks.

[45] Kim, Jangwon et al. "Automatic intelligibility classification of sentence-level pathological speech." *Computer speech & language* 29 1 (2015): 132-144 .

[46] H. Wang, J. Xu, H. Ge and Y. Wang, "Design and Implementation of an English Pronunciation Scoring System for Pupils Based on DNN-HMM," 2019 10th International Conference on Information Technology in Medicine and Education (ITME), Qingdao, China, 2019, pp. 348-352, doi: 10.1109/ITME.2019.00085.

[47] *International Journal of Innovative Technology and Exploring Engineering*, 2020. Diagnosis of Parkinson's Disorder through Speech Data using Machine Learning Algorithms. 9(3), pp.69-72.

[48] Mirsamadi, Seyedmahdad & Hansen, John. (2015). A Study on Deep Neural Network Acoustic Model Adaptation for Robust Far-field Speech Recognition.

[49] R. Gupta, K. Audhkhasi and S. Narayanan, "A mixture of experts approach towards intelligibility classification of pathological speech," 2015 IEEE International Conference on Acoustics, Speech and Signal

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.

Processing (ICASSP), Brisbane, QLD, 2015, pp. 1986-1990, doi: 10.1109/ICASSP.2015.7178318.

[50] V. Berisha, J. Liss, S. Sandoval, R. Utianski and A. Spanias, "Modeling pathological speech perception from data with similarity labels," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 915-919, doi: 10.1109/ICASSP.2014.6853730.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on August 27,2021 at 17:02:35 UTC from IEEE Xplore. Restrictions apply.