International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Development of Automatic Evaluation System to Assess the Speech Therapy given to Aphasia Patients

Darshika Kelin .AR[1], Dr. Jothi .S[2]

[1]PG Scholar, Dept of CSE, St.Joseph's College of Engineering, Chennai, India
[2]Assistant Professor, Dept of CSE, St.Joseph's College of Engineering, Chennai, India

---

Abstract - *Aphasia is a communication disorder that results in damage to the language parts of the brain and this is common in older adults particularly to those who had stroke. Aphasia may co-occur with speech disorder such as dysarthria or aphaxia of speech that also results in brain damage. The research investigates about aphasia by brain imaging techniques which helps to define brain function, severity of brain damage and predict the severity of aphasia. In depth testing of language ability of the patients with various aphasic syndromes is helping to design effective treatment. In this paper, the features used to train the classifier are: Pitch of voiced segment of the speech and the Mel-Frequency Cepstrum Coefficients (MFCC). Feature extraction using Mel- Frequency Cepstrum Coefficients (MFCC) for ASR in MATLAB.*

In the acoustic modeling phase, classification is done with the extracted features. In this paper, the classification is to predict the major aphasia type like broca's aphasia and wernicke's aphasia. The result examines the accuracy and the effectiveness using K- Nearest Neighbor (KNN) and the Recurrent Neural Network (RNN) classifiers, and also investigated the speech therapy that should be given to the patient based on the severity. KNN's main disadvantage of becoming significantly slower as the volume of data increases makes it an impractical choice in environments where predictions need to be made rapidly. So we introduced Long Short-Term Memory network that defines the LSTM network and state the number of features in the input layer and the number of classes in the fully connected layer. TrainNetwork validates the network every 50 iterations by predicting the response of the validation data and calculating the validation loss and accuracy. Finally, we have achieved good performance using sequence training.

***Key Words*: broca's aphasia; wernicke's aphasia; K-Nearest Neighbour; Recurrent Neural Network; Long Short-Term Memory Network.**

1. INTRODUCTION

Speech is acoustic signal of the speaker's that contains information of idea of the speaker. Aphasia is a disorder that affects the languages part of the human brain. Aphasia persists as a disability in 21% - 38% of stokes survivors. The number of person with aphasia (PWA) in India is above two million. Accordingly to the recent reviews the predominant way is Speech Language Therapy (SLT) to the PWA. In this paper we classify only the major aphasia types like broca's aphasia in which the patient may understand the speech and know what they want to say but they frequently speak in short phrases that are produced with great effort. Another type is wernicke's aphasia in which the patient may speak long complete sentences that have no meaning, adding unnecessary words even creating made up words. Therapy-based interventions are group programs, training conversations, computer-based instructions constraint-induced therapy. Aphasia is a long-standing disorder that results in isolating individuals from family, friends and loss of autonomy among others. Being able to detect and treat aphasia is PWA's speech which would provide the useful information to the Speech Language Pathologist (SLP) for treatment process. In addition, leading to computer-based activities for in-home practice for PWAs. It could increases the PWAs awareness of errors and improves the self monitoring skills. Traditionally, HMM based modeling is derived in the speech recognition system that results in mismatch of the training and the testing data so much more effort has been spent in improving the recognition system. To improve the efficiency and accuracy in this paper we review the KNN and the LSTM and also investigated the development and assessment for the ASR technologies.

2. AUTOMATIC SPEECH RECOGNITION

Automatic Speech Recognition is the technique that allows the computer to capture the speech spoken by human through microphone. Automatic speech recognition is a difficult problem where the first papers date from about 1950. For the last few decades ASR has been profited with the flow of internet among researchers in the field of speech processing research area. A number of techniques like word-template matching, dynamic-time-wrapped, linear-time scaled, linguistically motivated approaches (finding the phonemes, assembling words, assembling sentences) and Hidden Markov Model were used, and still HMMs giving the best performance. ASR provides the accuracy and evaluation for the speech input from the PWAs in speech rehabilitation therapies. In addition, ASR

provides the real time feedback response to their mistakes. ASR system is highly influenced by the feature extraction method chosen, since the classification stage will have to classify efficiently the input speech signal based on these extracted features. There are many applications in ASR system such as voice dialing, home automation system, text to speech conversion, speech to text conversion, telephone communication, lip synchronization etc. Recently, speech processing becomes the novel approach of security. Automatic Speech Recognition (ASR) only takes acoustic information contained in speech signal. In noisy environment, it gives very less accuracy. Speech processing is processed in three different levels. Considering the signal level processing the anatomy of human auditory system and processing the signals in the form of chunks called frames. In phoneme level, speech phonemes are collected and processed. Phoneme is the fundamental unit of speech. Word processing is the third level that paved way on linguistic entry of speech. Data collection and analysis, feature extraction, modeling and testing are the phases of the speech processing.

3. FEATURE EXTRACTION TECHNIQUES

The purpose of extraction is to find a set of properties of an utterance that have acoustic correlates in the speech signal, that is, the parameters that can be estimated through processing of the signal waveform. Such parameters are termed as features.

The following properties are required for a good feature extractor:

- Compact features to enable the real time analysis,
- Minimize the loss of discriminant information.

The features were extracted from the speech recording. The features used to train for classifier are: Pitch of the voiced segment of the speech and the Mel-Frequency Cepstrum Coefficients (MFCC). Pitch and Mel-Frequency Cepstrum Coefficients (MFCC) are extracted from speech signals recorded for 10 speakers. These features are used to train a K-Nearest Neighbor (KNN) classifier. Then, new speech signals that need to be classified go through the same feature extraction. Pitch and Mel-Frequency Cepstrum Coefficients (MFCC) are extracted from speech signals recorded for 10 speakers. Once you isolate a region of voiced speech, you can characterize it by estimating the pitch. Apply pitch detection to the word "two" to see how pitch changes over time. This is known as the pitch contour, and is characteristic to a speaker.

An ASR system has recognized what it has learned during its training. However, the
system is able to recognize even those words, not present in the training corpus for which sub-word units of the new word are known to the system and the new word exists in the system dictionary. The feature extraction techniques are Linear Discriminate Analysis (LDA), Linear Predictive Coding (LPC), Mel-frequency Cepstrum Coefficients (MFCC), RASTA-PLP (Relative Spectra Filtering of log domain coefficients).There are many other techniques used for extraction and most commonly MFCC, PLP and LPC are used algorithms in the field of speech processing. In this paper, pitch and MFCC features are extracted from each frame. Collected the samples and framing is done with 30ms with the overlay of 75%.

3.1 Pitch Detection Algorithm

Pitch detection for clean speech is mostly considered a solved problem. Pitch detection with noise and multi-pitch tracking remain difficult problems. There are many algorithms that have been extensively reported on in the literature with known trade-offs between computational cost and robustness to different types of noise.

Pitch is also used to indicate and analyze pathologies and diagnose physical defects. In MIR, pitch is used to categorize music, for query-by-humming systems, and as a primary feature in song identification systems.

Usually, a Pitch Detection Algorithm (PDA) estimates the pitch for a given time instant. The pitch estimate is then validated or corrected within a pitch tracking system. Pitch tracking systems enforce continuity of pitch estimates over time.

3.2 Mel- Frequency Cepstrum Coefficient

In this paper, we present MFCC for feature extraction. The proficiency of this phase is important for the further phase since it affects the behavior of modeling process.

Various signal processing operations such as sampling, framing, windowing and Mel cepstrum analysis are performed on the input signal, at different stages of the MFCC algorithm.

a. Sampling

The dataset is gathered from the aphasiabank, where it is a large-scale database basically available for the research works. In this paper, we present he dataset contains recordings of male and female subjects speaking words and numbers. Speech files are recorded in ‘wave’ format, with the following specifications: Fs = Sample rate in Hertz = 8000 and n = Number of bits per sample = 16. The raw files is converted into free lossless audio code (flac) using helper function in MATLAB. The speech files are divided into subdirectories depends on the labels corresponding to the speakers. When recording music or many types of acoustic events, audio waveforms are

typically sampled at 44.1 kHz (CD), 48 kHz, 88.2 kHz, or 96 kHz. Sampling rates higher than about 50 kHz to 60 kHz cannot supply more usable information for human listeners.

b. Framing

Speech is non-stationary in nature if the frame size is too long the signal properties may change too much across the window, affecting the time resolution. If the frame size is too short, resolution of narrow-band components will be sacrificed, affecting the frequency resolution adversely. A speech signal typically is stationary in windows of 20ms. Therefore the signal is divided into frames of 20 ms which corresponds to n samples:

$n = t_{st}\, f_s$

Overlapping frames are used to capture information that may occur at the frame boundaries. Number of frames is obtained by dividing the total number of samples in the input speech file. We obtained the framing size as 30ms with the overlap of 75%.

c. Windowing

Due to the lack of continuity at the beginning and end of the frame are likely to introduce unaccepted effects in the frequency response. Hence, each row is multiplied by window function. A window alters the signal, reducing it to nearly zero at the beginning and the end. We use Hamming window as, it introduces the least amount of distortion.

d. Extraction

The most popular and the widespread method to extract features is deriving Mel-Frequency Cepstrum Coefficient (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which depends on the human ear scale. MFCCs being examined as frequency domain features are much more accurate when comparing with time domain features. From the observed literatures the human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000Mels.

The Mels frequency can be calculated using the following formula:

$$\text{Mel}\,(f) = 2595 * \log 10\,(1 + f / 700).$$

A step by step implementation of the MFCC is shown in Figure 1.1. The speech samples is sampled to 16000 Hz for analysis purpose. Framing of N sample is done with the adjacent frame which is being separated by M. And at last the log Mel spectrum was converted into time. Then the output is so called Mel Frequency Cestrum Coefficients (MFCC). Hence the speech signal is dynamic and changes over time and here it is assumed that speech signals are stationary on short time scales and their processing is done in windows of 20-40 ms. This uses a 30 ms window with a 75% overlap.

Block diagram of Feature extraction using MFCC

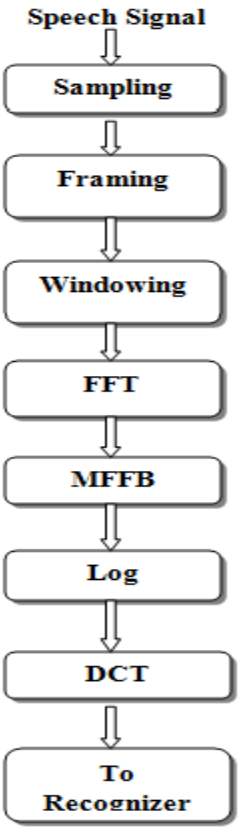

Figure 1.1 Derivation of MFCC

In MATLAB splitlabeldatastore method is used to split the datastore in two or more datastores. 80% of the data for each label is used for training, and the remaining 20% is used for testing. The countEachLabel method of audioDatastore is used to count the number of audio files per label.

Pitch and MFCC features are extracted from each frame using HelperCompute Pitch and MFCC which performs the following actions on the data read from each audio file:

1. Collect the samples into frames of 30 ms with an overlap of 75%.

2. Compute the pitch and 13 MFCCs (with the first MFCC coefficient replaced by log-energy of the audio signal) for the entire file.
3. Keep the pitch and MFCC information pertaining to the voiced frames only.
4. Get the directory name for the file. This corresponds to the name of the speaker and will be used as a label for training the classifier.

HelperCompute Pitch and MFCC returns a table containing the filename, pitch, MFCCs, and label (speaker name) as columns for each 30 ms frame.

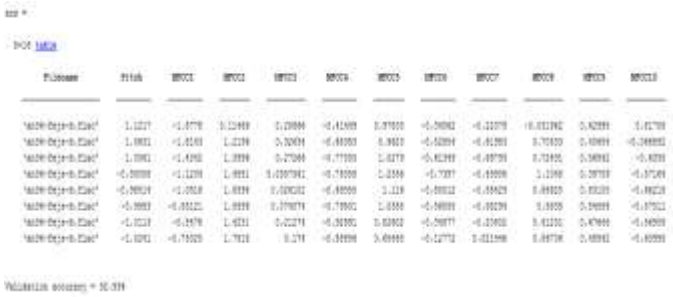

Figure 1.2 Screenshot of Pitch and MFCC Values with accuracy

MFCC11	MFCC12	MFCC13	Label
-0.29036	-0.47	-0.04532	'fejs'
-0.60354	-0.53907	-0.51924	'fejs'
-0.60484	-0.62735	-1.0637	'fejs'
-0.36655	-0.67672	-0.87127	'fejs'
-0.84113	-0.66903	-0.60151	'fejs'
-1.1255	-0.64767	-0.8494	'fejs'
-1.3199	-0.43764	-0.38468	'fejs'
-1.8518	-0.50805	-0.26085	'fejs'

Figure 1.3 MFCC values and labels

4. ACOUSTIC MODELING

Acoustic models are used to connect the observed features of the signals with the expected phonetics of the hypothesis sentence. The most representative implementation of this process is probabilistic, making use of hidden Markov models (HMM). Classification or recognition is the process of comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity between them. There are two possible approaches to match the patterns available during training and testing. The first one is based on the distance between the acoustic units of training corpus and the acoustic unit of recognition and is known as dynamic time warping (DTW). The second model is based on the maximization of the occurrence probability between training and recognition units and is implemented by HMMs. In this paper, we present the classification method like KNN and RNN training network. These methods will be discussed following with their implementation.

i. K- Nearest Neighbor Algorithm

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time. KNN's main disadvantage of becoming significantly slower as the volume of data increases makes it an impractical choice in environments where predictions need to be made rapidly. Moreover, there are faster algorithms that can produce more accurate classification results. However, provided you have sufficient computing resources to speedily handle the data you are using to make predictions, KNN can still be useful in solving problems that have solutions that depend on identifying similar objects. Here the features were collected from all ten speakers, and we train a classifier based on them. K-nearest neighbor is a classification technique naturally suited for multi-class classification. The hyperparameters for the nearest neighbor classifier include the number of nearest neighbors, the distance metric used to compute distance to the neighbors, and the weight of the distance metric. The hyperparameters are selected to optimize validation accuracy and performance on the test set. Finally trained the classifier and found the cross-validation accuracy.

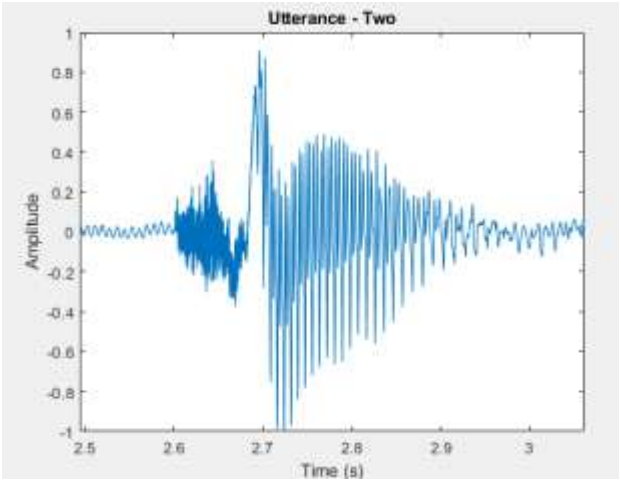

Figure 2.1 Time-domain representation of word two

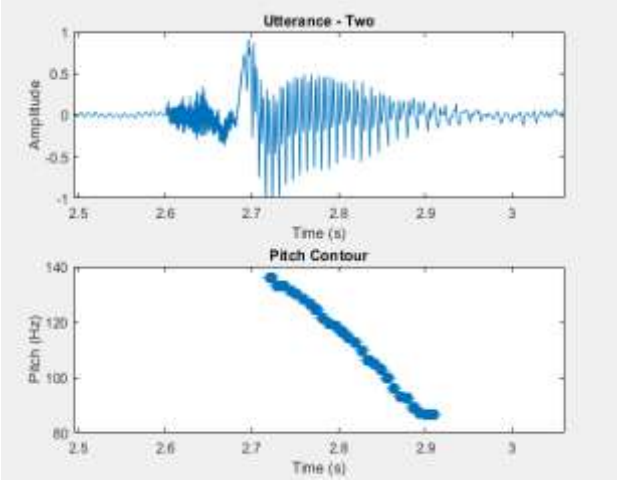

Figure 2.2 Pitch detection of word two and pitch contour

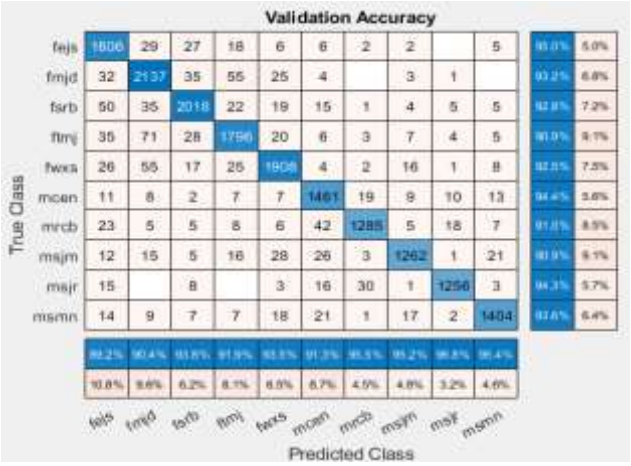

Figure 2.3 Validation Accuracy table

ii. Recurrent Neural Network Algorithm

Recurrent Neural Networks, or RNNs, were designed to work with sequence prediction problems. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behavior required in complex problem domains like machine translation, speech recognition etc. Now I have created sequence classification network using the Long short- term memory classification and the extracted features and labels gets loaded as the data for the classification phase. An LSTM network is a type of Recurrent Neural Network(RNN) that learns long-term dependencies between time steps of data. Now defined the LSTM network and state the number of features in the input layer and the number of classes in the fully connected layer. The network architecture can vary depending on the types and numbers of layers included. The types and number of layers included depends on the particular application or data. The next step is to set up the training options for the network. Use the training options function to define the global training parameters. To train a network, use the object returned by trainingOptions as an input argument to the train Network function. To perform network validation during training, specified validation data using the 'ValidationData' name-value pair argument of trainingOptions. TrainNetwork validates the network every 50 iterations by predicting the response of the validation data and calculating the validation loss and accuracy. Finally during this validation the accuracy rate as 95.41% and during the testing phase got the accuracy of 0.941.

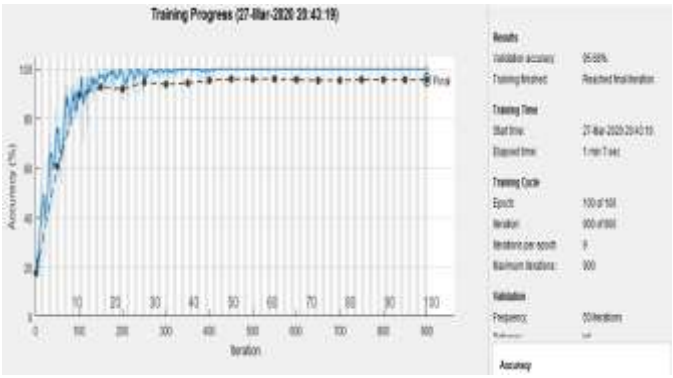

Figure 3.1 Validation of training

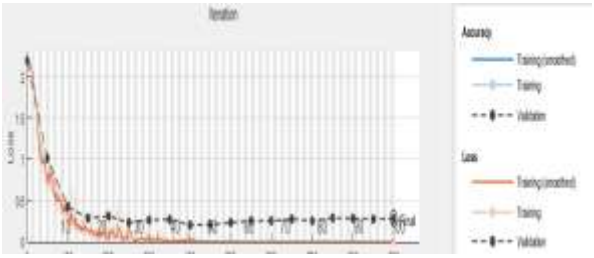

Figure 3.2 Loss

```
ans =

  5×1 cell array

    {12×20 double}
    {12×26 double}
    {12×22 double}
    {12×20 double}
    {12×21 double}


acc =

    0.9676
```

Figure 3.3 Testing Accuracy

5. LINGUISTIC THERAPY

Language consideration plays an important role in the clinical therapy. Speech language Pathologist needs to familiarize themselves with predominant language level where the language breaks down in aphasia. Some of the language parameters are: word frequency, imageability, grammatical class, nameability, phonological features, morphological features, types of sentences, minimally differing word pairs for auditory and visual discrimination, syllabic patterns, phonologically balanced and pronounceable non-words, accent, and sonority.

As per the recent study, there is strong proof to show that SLT is helpful to PWA. The intensity, dose, and duration of therapy are censorious factors. Impairment-based therapy and overall communication methods are commonly reinforcing as the ultimate goals are the same. There is absolute lack of awareness about need, efficacy, and availability of therapy and rehabilitation among PWA and caregivers. Hence, the needs of PWA are generally addressed at an primary level, leading to low expectations. There is no aspiration to become self-dependent again. A change in social attitudes must come by transmit an awareness that individual goals for PWA are worth the time and resources. It is imperative to increase awareness among the public and professionals about the effectiveness of speech and language therapy. The advantages of early referral and interceding should be highlighted.

Newly discovered methods in SLT such as "constraint-induced therapy" and intensive language action therapy involving intense and prolonged therapy sessions of 3–4 hr every day for a short duration of around 15 days as well as forcing the patient to use only speech for a few hours every day are known to construct quantifiable profits.

Group therapy sessions can be convenient if:

(i) Patients with the nearly same contour are offered particular practice sessions,
(ii) Different groups are provided general-purpose learning and practice opportunities to enhance the overall communication skills, and
(iii) To provide a manifesto for social interactions.

6. CONCLUSION

Thus in this paper, the technique used in each stage of the speech recognition are discussed. In this paper, Pitch of the voiced segment were calculated .This overall review found that MFCC is best among the feature extraction techniques because it is noise robust. We evaluated the performance of KNN- based modeling for the noise robust recognition system, the time domain representation of a particular word two was examined and which is characterized by a fundamental frequency. Once the speech is isolated it is characterized by estimating the pitch. We have applied the pitch detection algorithm to see how the pitch changes over time, which is called as pitch contour. MFCC is calculated for every file and the processing is done in windows of 30ms with a 75% overlap. Here cross validation is done and found the validation accuracy of 92.93%. To improve the accuracy and robustness and to train a deep neural network that classifies the features using Long Short-Term Memory (LSTM) network . LSTM network is type of Recurrent Neural Network (RNN). The network was trained and the labels were predicted. The recognition accuracy found to be 95.68%. Thus in this paper, the robustness of the speech is obtained in RNN network. Further analysis will be as hybridization of K-Nearest Neighbor and the Long Short-term memory network to obtain recognition accuracy and effectiveness. And also IOT based system will be implemented in which the patients will gain their therapy easily from anywhere with the internet connection.

REFERENCES

[1] Moritz Einfalt Rainer Lienhart, Matthew Lee, Lyndon Kennedy, "Detecting Speech Impairments from Temporal Visual Facial Features of Aphasia Patients," IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 103-108, 2019.

[2] Christian Kohlschein, Daniel Klischie`s, Tobias Meisen, Bj¨ornW. Schuller, Cornelius J.Werner, "Automatic Processing of Clinical Aphasia Data collected during Diagnosis Sessions: Challenges and Prospects," Proceedings of the LREC Workshop "Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)", Dimitrios Kokkinakis (ed.) pp.11-18, 2018.

[3] Ying Qin, Tan Lee and Anthony Pak Hin Kong, "Automatic Speech Assessment For Aphasic Patients Based On Syllable-Level Embedding And Supra-Segmental Duration Features," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5994-5998, 2018.

[4] Vishal Passricha and Rajesh Kumar Aggarwal, "Convolutional Neural Networks for Raw Speech Recognition," chapter.2, 2018.

[5] Rezki Trianto1, Tzu-Chiang Tai2, and Jia-Ching Wang, "Fast-LSTM Acoustic Model for Distant Speech Recognition," IEEE International Conference on Consumer Electronics (ICCE), 2018.

[6] Ying Qin1, Tan Lee1, Yuzhong Wu1, and Anthony Pak Hin Kong, "An End-to-End Approach to Automatic Speech Assessment for People with Aphasia," IEEE 11th

International Symposium on Chinese Spoken Language Processing, pp. 66-70, 2018.

[7] Christian Kohlschein, Maximilian Schmitt, Bj¨orn Schuller, Sabina Jeschke and Cornelius J. Werner "A Machine Learning Based System for the Automatic Evaluation of Aphasia Speech", IEEE 19th International Conference on e-Health Networking, Applications and Services, 2017.

[8] Duc Le, Keli Licata, Emily Mower Provost, "Automatic Paraphasia Detection from Aphasic Speech: A Preliminary Study," INTERSPEECH Stockholm, Sweden, pp.294-298, 2017.

[9] Duc Le and Emily Mower Provost, "Improving Automatic Recognition of Aphasic Speech with AphasiaBank," University of Michigan Computer Science and Engineering, 2016.

[10] Tan Lee, Yuanyuan Liu, Pei-Wen Huang, Jen-Tzung Chien, Wang Kong Lam
Yu Ting Yeung, Thomas K.T. Law, Kathy Y.S. Lee, Anthony Pak-Hin Kong, Sam-Po Law, "Automatic Speech Recognition For Acoustical Analysis And Assessment Of Cantonese Pathological Voice And Speech," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6475-6479, 2016.

[11] Kartiki Gupta, Divya Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech Recognition System," IEEE 6th International Conference-Cloud System and Big Data Engineering, pp. 493-497 2016.

[12] Aman Ankit, Sonu Kumar Mishra, Rinaz Shaikh, Chandraketu Kumar Gupta, Prakhar Mathur and Soudamini Pawar, "A Survey Paper on Acoustic Speech Recognition Techniques," International Journal of Recent Advances in Engineering & Technology (IJRAET), 2347 - 2812, Vol.4, Issue -7, 2016.

[13] Duc Le, Keli Licata, Carol Persad, and Emily Mower Provost, "Automatic Assessment of Speech Intelligibility for Individuals With Aphasia," Member, IEEE, ACM Transactions on Audio, Speech, And Language Processing, Vol. 24, 2016

[14] Samina Khalid, Tehmina Khalil, Shamila Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning" Science and Information Conference, pp. 372-378, 2014.

[15] S. B. Magre, P. V. Janse, R. R. Deshmukh, "A Review on Feature Extraction and Noise Reduction Technique," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 2, 2014 .

[16] Duc Le, Keli Licata, Elizabeth Mercado, Carol Persad, and Emily Mower Provost, "Automatic Analysis Of Speech Quality For Aphasia Treatment," IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[17] Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, Lirong Dai, and Qingfeng Liu,"Fast Adaptation of Deep Neural Network Based on Discriminant Codes for Speech Recognition," IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, Vol. 22, NO. 12, 2014.

[18] Ms. Rupali S Chavan, Dr. Ganesh. S, "An Overview of Speech Recognition Using HMM, Sable International Journal of Computer Science and Mobile Computing," IJCSMC, Vol. 2, Issue. 6, pp.233 – 238, 2013.

[19] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition International Journal For Advance Research In Engineering And Technology," Vol. 1, Issue VI, 2013.

[20]Hinton, Geoffrey et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", *IEEE Signal Processing Magazine*, Vol. 29.6, pp. 82-97, 2012.

[21] N.A. Association, "Aphasia", http:// www.aphasia.org/

[22] https:// www.wikipedia.org/ - Wikipedia.