SPEECH ASSIST

Amitha Khan Irshad[1], Arathy Krishnan[2], Avinash A.S[3], Siddharth Sanjay[4], Remya Annie Eapen[5], Lani Rachel Mathew[6]

[1-6]Department of Electronics, Mar Baselios College of Engineering and Technology, Kerala'.
*[7]**K. Gopakumar**, Department of Electronics, TKM College of Engineering, Kerala*

Abstract-- This paper aims at producing a real-time system that reconstructs partially formed words of persons with disability in speaking. RNN (Recurrent Neural Network) is used here to detect the partially spoken words and the obtained text is converted to audio signal. We created an Automatic Speech Recognition (ASR) system that can better classify the words spoken by people with defects like stuttering. Also we included joint subband coding system which provided a higher speech intelligibility performance than other Voice conversion approaches in most test conditions. It implies that the Voice conversion system could potentially be used as one of the electronic assistive technologies to improve the speech quality for a dysarthric speaker.

I INTRODUCTION

Dysarthria is a condition in which the muscles we use for speech are weak or you have difficulty controlling them. Dysarthria often is characterized by slurred or slow speech that can be difficult to understand. It contributes to failure of the voice motor control process, triggering sluggish or uncoordinated articulatory motion, resulting in incomprehensible speech. Depending on the type of dysarthria and its extent, various therapies have been suggested to enhance the intelligibility of dysarthric expression. Speech assistive system is used to help people with speaking disabilities. This system reconstructs the partially formed words of a person with speaking disability. In order to recognize the word an RNN is used. An RNN is one type of artificial neural network. In it the connection between the node of the RNN form a directed graph along a sequence. Audio information plays a rather important role in the increasing digital content that is available today; resulting in a need for methodologies that automatically analyze such content. Speaker Identification is one of the vital field of research based upon Voice Signals. Its other notable fields are: Speech Recognition, Speech-to-Text Conversion, and vice versa, etc. Mel Frequency Cepstral Coefficient (MFCC) is considered a key factor in performing Speaker Identification. But, there are other features lists available as an alternate to MFCC; like-Linear Predictor Coefficient (LPC), Spectrum Sub-band Centroid (SSC), Rhythm, Turbulence, Line Spectral Frequency (LPF), Chroma Factor, etc. Gaussian Mixture Model (GMM) is the most popular model for training on our data [2]. The training task can also be executed on other significant models; viz. Hidden Markov Model (HMM). Recently, most of the model training phase for a speaker identification project is executed using Deep learning; especially, Artificial Neural Networks (ANN). In this project, we are mainly focused on implementing MFCC and GMM in pair to achieve our target.

Another method we used here is joint subband coding, here we use electronic assistive technologies (EAT) used to improve the capability of a dysarthric user to communicate with a human. One of the effective EAT methods giving for dysarthria speaker is voice conversion(vc) it convert the dysarthric speech to normal speech.

II RELATED WORK

End-to-end dysarthric speech recognition using multiple databases by Yuki Takashima, Tetsuya Takiguchi, Yasuo Ariki,In this paper they used an end-to - end ASR program based on a LAS model for a Japanese individual with an articulation disorder. The amount of speech data obtained from these speakers is very small, due to their athetoid symptoms. They used additional speech data from a physically unimpaired person and foreign individuals with articulation disabilities to tackle this issue[1].

An Automatic Diagnosis and Assessment of Dysarthric Speech using Speech Disorder Specific Prosodic Features. This work suggests an automated diagnosis and assessment of dysarthria, unlike the traditional method. The aim of this paper is to diagnose the extent of dysarthria and classify it. Common prosodic traits of speech disorder are identified using genetic algorithms. Help vector machines are used to diagnose and test dysarthric expression.

III SPEECH RECOGNITION USING NEURAL NETWORK

All Automatic Speech Recognition(ASR) system works with a neural network. This network is trained to detect words spoken. Speech to Text(STT) and Text to Speech(TTS) here are neural networks, both of them works in different ways. STT converts audio input to text and TTS converts text to Audio output. The audio of the person with disability in speaking is taken as the input. From this audio input feature extraction is extracted[1].

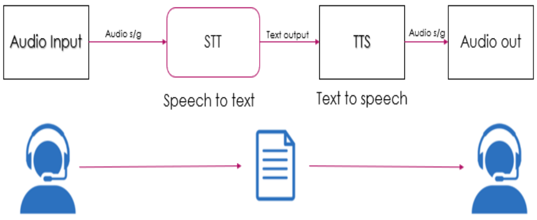


Fig:1 basic block diagram

IV PROPOSED METHOD USING RECURRENT NEURAL NETWORK

The features extracted from the speech samples for diagnosis are Mel Frequency Cepstral Coefficient (MFCCs), peak amplitude, , fundamental frequency . The above mentioned features are combined to form a complex feature vector. The same features are extracted from the speech samples of dysarthria speakers and controlled speakers. Hence, the speech features have been selected, then the selected samples are trained and tested using the recurrent neural network(RNN). The main significance of the methodology is to diagnose and classify the dysarthric speech by using speech disorder specific features.

V JOINT SUB-BAND CODING

This method provides a joint feature learning approach to improve a sub-band coding network-based voice conversion(VC) system. The results showed that the JSB coding system provided a higher speech intelligibility performance than other VC approaches. It implies that the JSBcoding system can be used as one of the electronic assistive technologies to improve the speech quality for a dysarthric speaker. The procedure that follow is first we extract the acoustic features of log-power spectra ($\boldsymbol{L}j$) and mel frequency cepstral coefficient (MFCC) ($\boldsymbol{M}j$) by the frame. We collected 2585 samples for hearing in noise test. We tested and trained it in the neural network. The results indicate that the JSBcoding system can improve the performance of dysarthria speech better than the other. Joint feature learning-based SBcoding system to improve speech intelligibility performance. It implies that joint feature learning is useful for a VC system to achieve better performance, in the limited data conditions.

$(\boldsymbol{F}j = [(\boldsymbol{L}j), (\boldsymbol{M}j)]$

$w(n)=0.54-0.46\cos(2n/(N-1))$

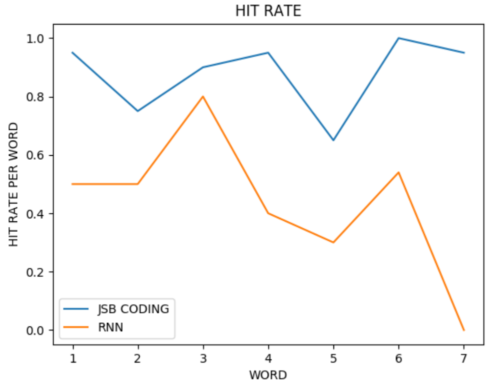
VI EXPERIMENTAL RESULT

Joint feature learning is useful for VC system to achieve better performance, in the limited data conditions. Genetic algorithm identifies speech prosodic specific features like Mel Frequency Cepstral Coefficients(MFCCs). Another main contribution is proposed system is gender and speaker independent hence making the proposed system more robust. The proposed JSB coding system could provide suitable and slight benefit for a dysarthric user in close- and open-set testing conditions. RNN is used here to detect the partially spoken words and the obtained text is converted to audio signal. This audio signal will be synthesized using another RNN to obtain similarity to the user's voice. MFCC technique is used for feature extraction. Training is required for making the network like the user's vocabulary and voice.

WORDS	HIT RATE PER WORD
Hello	20/20
Tablet	19/20
Music	19/20
Water	19/20
Chair	18/20
Food	15/20
Sleep	13/20

Table 1: Hit rate per word

Total number of word predicted right=123/140



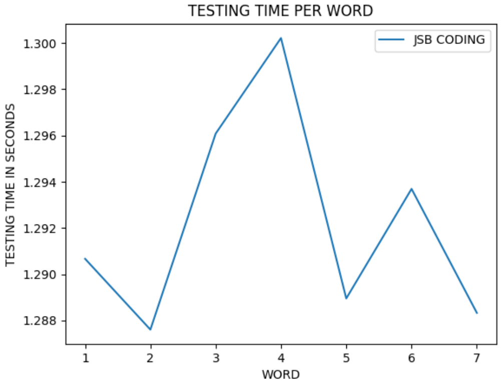


Fig 2: Testing time per word

Fig 3: Hit rate per word

VII CONCLUSION

Our system tries to find out the partially pronounced sound. Here recurrent neural network is used for the recognition of the word. The proposed JSB coding system could provide suitable and slight benefit for a dysarthric user in close- and open-set testing conditions. By using JSB coding method, 87.85% accuracy was obtained. The training time is 4 samples per second and training time of 2-3 seconds in real time.

VIII REFERENCES

[1]. G. Vyas, M. K. Dutta, J. Prinosil and P. Harár, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," 2016 39th International Conference on Telecommunications and Signal Processing (TSP), Vienna, 2016, pp. 515-518, doi: 10.1109/TSP.2016.7760933.

[2] Y. Takashima, T. Takiguchi and Y. Ariki, "End-to-end Dysarthric Speech Recognition Using Multiple Databases," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6395-6399,doi:10.1109/ICASSP.2019.8683803

[3]. George E. Dahl, Dong Yu, Li Deng, and Alex Acero, “Large vocabulary continuous speech recognition with context-dependent DBN-HMMs,” in ICASSP, 2011, pp. 4688–4691.

[4] Rudzicz F. “Adjusting dysarthtia speech signals to be more intelligible”, Journal in speech communication, vol 27, pp 1163-1177, 2013.

[5] M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for MFCC feature extraction," 2010 4th International Conference on Signal Processing and Communication Systems, Gold Coast, QLD, 2010, pp. 1-5.