Pakistan Journal of Scientific Research, ISSN: 0552-9050
vol. 1, No. 2 (13-18)

Voice Reminder Assistant based on Speech Recognition and Speaker Identification using Kaldi

Adnan Ali*, Javeria Naveed, and Meer Hamza

***Abstract*— A user often wants to do some tasks which don't need a specific time to complete but are important, due to busy routines and possible memory loss problems user forgets to accomplish such tasks or delay them to manage later. The task may be about conveying some information to a person when the user meets him, sort things in his wardrobe, or locker, etc. A user can meet the person anywhere or anytime so he cannot set a reminder for a specific time. Such tasks can be accomplished if the user has a reminder system, which reminds him about the tasks at runtime. In this paper, we present a reminder system based on speech recognition & speaker identification; personal voice- controlled aid for the user to remind about the tasks when some specific keyword/phrase is spoken by the user or user reaches a specific location so that the user can accomplish the important time independent tasks without being time-bound. The specified keyword can be anything, person name, place name, and any phrase. The user can speak to the assistant to remind about some task without un- locking his mobile or opening the application for a specific keyword/phrase & location. When the specific keyword/phrase is spoken or user reaches the specific location the assistant also speaks and reminds the user about the task. Voice reminder assistant supports Text-Independent Speaker Identification unlike other voice assistants already available.**

***Index Terms*— Voice Reminder, Speech Recognition, Speaker Identification, Assistant.**

I. Introduction

ENERGY Voice Reminder Assistant is a reminder application which helps the user to remind about the task when some specific speech keyword/phrase is spoken by the user or user reaches a specific location. It is an Android application, which is the implementation of the idea being presented in this paper. In old reminder systems, the user needs to set the date and time to remind about the task, but Voice Reminder Assistant gives some additional features so that the user is not date & time bound to complete the task, Because some tasks are not needed to be completed for a specified date & time, but are important to complete for such type of task completion the reminder assistant is helpful moreover it is personal voice- controlled system the user can use it by speaking to it. The user can set all types of reminders, date & time as well as keyword/phrase & location based reminders. The assistant is based on speech recognition & speaker identification. Speech Recognition is defined as the conversion of voice signal into readable text. Speech Recognition is used to operate a device with our voice instead of giving input manually. Usually, people prefer it because it provides us easiness to use systems. In our android application, the user's voice is converted into text on which further processing is done to determine what the user has spoken, the user sets a keyword/phrase with the help of speech recognition it helps the assistant to verify that is it same keyword/phrase for which the reminder is saved by the user as well as to understand the user commands to set the reminder through his voice.

Speaker Identification is defined as the process or technique of identifying the person through the person's voice. Nowadays, many authentication systems are using Speaker Identification for authentication purposes. We are using Speaker Identification to identify the person for which the assistant should work and take command. During

Identification, the user's voice features are extracted and further processing is carried out to determine that the user is the claimed one or not. There are two types of Speaker Identification: Text-Independent Speaker Identification & Text-Dependent Speaker Identification [1]. Text-Dependent Speaker Identification is mostly used when we need to identify the user just for a single sentence and for Identification the user has to speak the same sentence if the user fails to speak the same sentence the user is not recognized successfully, in such type of Identification the user cannot use the system freely. In Text-Independent Speaker Identification the user is given the system to use freely because in this type of Identification the user can be identified by speaking any sentence. In Voice Reminder Assistant Text-Independent Speech Identification is used so that the user can easily use the application.

II. Literature Review

There are many voice-controlled personal assistants [2] a user

Department of Information Technology University of Gujrat , Sialkot sub Campus Sialkot, Pakistan (correspondence e-mail: mhadnana01@gmail.com)

Received: 5-03-2021, Revised: 15-04-2021, Accepted: 15-04-2021

can speak to them and take help from them to do different things such as making a call, sending a message or setting a reminder for specific date & time. These types of assistants are made by companies like Google, Amazon & Apple. These assistants can be communicated by calling their name & giving commands. The advantage of these voice-controlled assistants is that an illiterate person or a person having any disabilities can easily communicate with the system & can have his work done perfectly. These Assistants are based on speech recognition & keyword spotting system.

Keyword Spotting Using Contextual Speech Recognition [3] is used by the Google assistant for identifying the user commands which are given to assistant for processing. The assistant uses the keyword “Hey Google” & “Ok Google” to determine that the user has given the command to Google assistant or not. The Google assistant uses Text-Dependent

Speaker Identification if the user has trained the model for the assistant. Otherwise, any user can use the assistant. The procedure which is used by the Google assistant for processing commands is that the assistant waits for the trigger keyword (“Ok Google” or “Hey Google”) in the speech content that is spoken by the user & When the speaker speaks the trigger keyword the speech content after the keyword is sent to the server for converting speech into text on which further working is done.

Nowadays human to system interaction systems using speech recognition are increasing day by day. Speech Recognition is challenging because the environment, speaking style of a user or the context of a speech always varies. There are many Speech Recognition toolkits [4] which help us in building an ASR system such as HTK, CMU Sphinx & Kaldi. Hidden Markov Model Tool Kit (HTK) is used to build Hidden Markov Models and can also be used in the designing of the speech recognition system. HTK provides scripts for acoustic modeling, which can be changed for any other recognition applications. CMU Sphinx is available in java & C. the users can train a speech recognition model as well as acoustic models with the help of it. Kaldi is a speech recognition toolkit written in C++. Kaldi uses deep neural networks & Gaussian mixture models and standard Gaussian mixture models. Kaldi supports speech & speaker identification and Dairization. Speech Recognition process includes Acoustic Model, Dictionary Model & Language Model. The SRI Language Modeling Toolkit [5] (SRILM for short) is an open source software toolkit for statistical language modeling and related tasks. It is a toolkit for building and evaluating statistical language models (LMs). It supports N-gram statistics based LM, including the standard backoff models, with an array of standard smoothing algorithms.

III. METHODOLOGIES

The voice signal carries information with it which consists of the speech utterances the person is speaking as well as the identity of the speaker. So with the help of a person's voice, we can determine what the person is speaking and who the speaker is. The toolkit used for speech recognition and speaker identification systems building is Kaldi, all the development related to speech recognition & speaker identification has been done using it.

A. Speech Recognition

We are using an HMM (Hidden Markov Model) for training our system for speech recognition. Hidden Markov Model is a

statistical model having finite probabilistic hidden state transitions. In Kaldi for preparing Speech Recognition System following steps are done:

- Data Preparation

For training, a Speech Recognition System in Kaldi data preparation is very important. In data preparation, we needed audio data set with the help of which we wanted to train the system. We have used data Librespeech [6] data which is freely available at openslr (Open Speech & Language Resources).

Table 1. Librespeech Data Subset Information

subset	hours	per-speaker minutes	female speakers	male speakers	total speakers
train-clean-360	363.6	25	439	482	921

For data preparation, some files need to be created under a specific location in Kaldi. Some files for referencing the audio files, acoustic model, dictionary model & language model are needed. In Fig. 1 the files which needed to be created & and under which location has been shown.

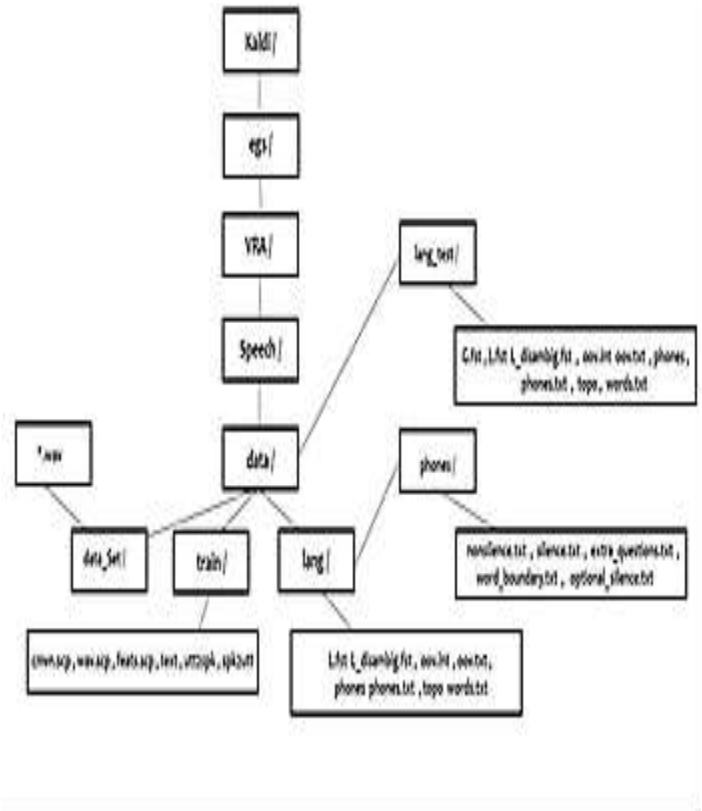

Fig. 1: Important Files & their locations

The files *.wav in the data_Set is the audio data set on which the speech recognition system is trained, files in train tells us how to communicate with the single audio file and features or any type of information related to it, files in the phones are useful as acoustic data and used to build the acoustic model, files in lang are for dictionary model, files in the lang_test are useful for making language model. Acoustic model & dictionary model has been prepared by us, for preparing language model SRILM has been used which is a toolkit for preparing language model. Acoustic Model is a collection of files which consists of phonemes sound written in text & any other sounds such as noise or silences produced. The acoustic model will be helpful in converting the audio signal into phonemes representation. Dictionary Model is a file also known as pronunciation dictionary which consists of isolated word and the corresponding phonemes which build up that word. The dictionary model is used to convert phonemes into isolated words. Language Model consists of sentence structure & long text sentences which are used to convert the isolated words into the meaningful sentence according to the language rules for which it is being used.

- Training of System

For Training the Speech Recognition system in HMM (Hidden Markov Model) the data preparation must be completely done. In Kaldi, we have trained our system as Mono-phone HMM. Mono-phone means that a single sound produced by the human will have only one corresponding phone. All the audio data set which was kept for training is used for training the speech recognition system. During training Acoustic Model, Dictionary Model & Language Model is trained for training data. For training Acoustic model acoustic data is used which contains phonemes. For training dictionary model maximum words with corresponding phonemes listed above them are used. For training language model a large text data is used to train the speech recognition system to recognize the sentence sequence. For training in Kaldi Mono-phone system steps/train_mono.sh file is used the file takes 3 parameters

1. Path to the reference directory for data training
2. Path to the Acoustic data, Dictionary data & Text data for Language model.
3. Path to the directory to store results of training.

The reference directory contains wav.scp, utt2spk, spk2utt which have information about the utterance-id, speaker-id, and path to the location of audios present in the data-set is present. When the system is trained the alignments are generated to know the best possible path for recognition. For alignments steps/align_si.sh file is used that take 4 parameters

1. Path to the reference directory for data training.
2. Path to the Acoustic Model, Dictionary Model & Language Model.
3. Path to the directory to the trained model.
4. Path to the directory to store results of alignments. After the system is trained and alignments are also generated then the next step is the graph compilation. The alignments generated are further compiled into a graph which is used for decision making later. For graph compilation utils/mkgraph.sh file is used that take 3 parameters

1. Path to the Acoustic Model, Dictionary Model & Language Model.
2. Path to the directory to the trained model.
3. Path to the directory to store compiled graph.

- Decoding

When the system is trained the last step we do is decoding it is the step in which we do testing of our trained ASR system. For testing, we convert the voice signal into the text at this stage & calculate the word error rate. For decoding in Kaldi steps/decode.sh is used the file takes 3 parameters

1. Path to the decoding graph
2. Path to the data to decode
3. Path to the directory to store results of decoding

For example, an audio signal is converted into the phonemes 'DH' 'AH0' 'N' 'AY1' ' N' 'B' 'OY2' 'Z' with the help of Acoustic Model. After that, the phonemes are converted into the isolated words "The" "Nine" "Boys" with the help of a dictionary model now for giving the isolated word the actual sentence shape language model is used which will convert it into the meaningful sentence "The nine boys". The working of speech recognition using the Acoustic Model, Dictionary Model & Language Model is shown in Fig. 2.

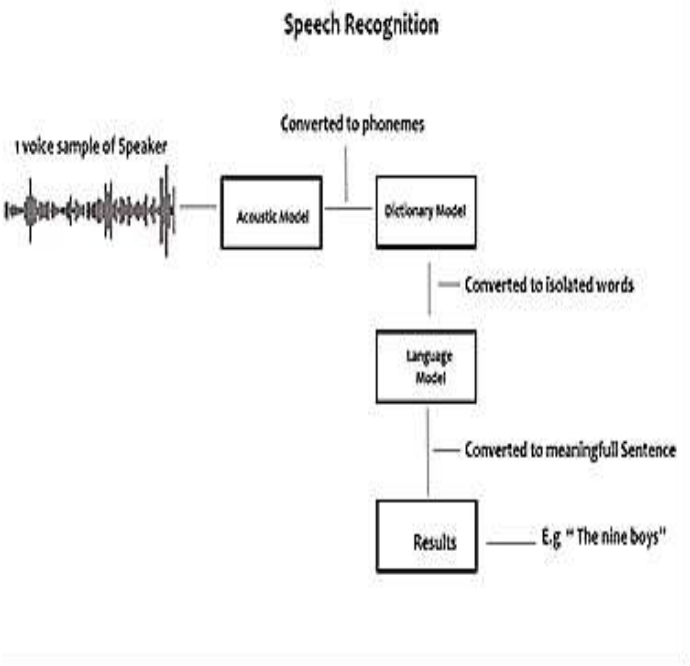

Fig. 2: Working of Speech Recognition System.

The word error rate for our ASR system has been calculated as 7.8% on the average calculation

B. Speaker Identification

Speaker Identification is defined as the process or technique of identifying the person through the person's voice. Nowadays many authentication systems are using speaker Identification for authentication purposes. We are using Speaker Identification to identify the person for which the assistant should work and take commands.

For identifying a person the process is carried in two steps

- Training
- Testing

Figure 3 explains the training process, the Person gives his multiple voice samples and from multiple voice samples the voice features which are specific to that person only are extracted and saved for identifying the user later

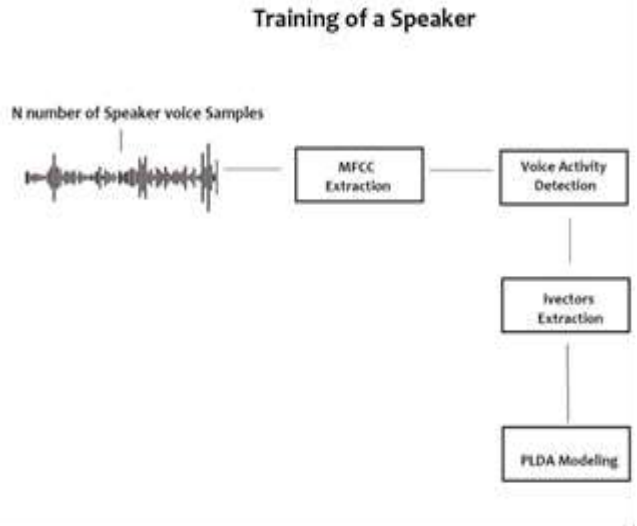

Fig. 3: Training of Speaker for enrollment of User.

Figure 4 explains the training process, the person single voice sample is taken and voice features are extracted and compared with the already saved featured if the maximum similarity is found the user is considered as the claimed user.

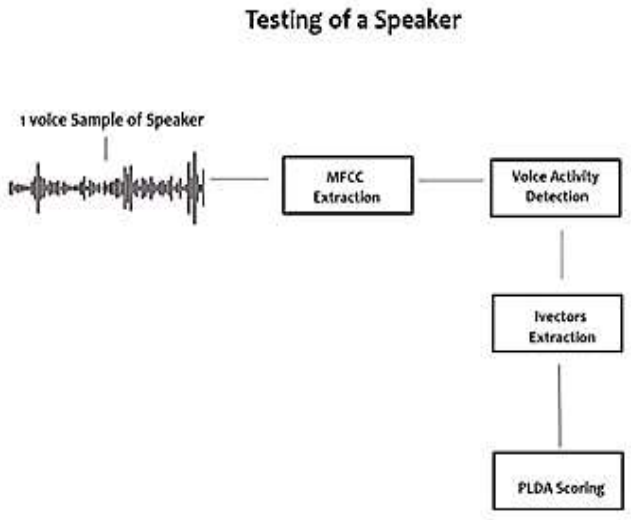

Fig. 4: Testing of Speaker for Identifying User.

GMM (Gaussian Mixture Model) is used; GMM is a parametric model of the probability distribution continuous measurements which is used for feature extraction for biometric systems [7]. In our assistant, it is used to extract vocal-tract related spectral features. Steps included in the Speaker Identification training and testing are described below:

- MFCC Coefficients Extraction

The first step is the extraction of MFCC Coefficients. MFCC (Mel-Frequency Cepstral Coefficients) [8] is defined as the representation of components of the audio signal which are helpful in understanding the important features. For this Purpose, the voice signal is divided into small frames of equal length than from each frame coefficients are extracted. For each frame, the periodogram estimate of the power spectrum is calculated. The Mel filter bank is applied to the power spectra, and in each filter, the sum of the energy is each calculated. The logarithm of all filter bank energies is calculated. After that DCT of the log filter bank energies is calculated. 13 DCT coefficients are kept rest are discarded. These 13 coefficients are the MFCC for the frame and in the exact way described above each small frame is processed and we get n x 13 frame matrix where n is the number of frames and 13 is the MFCC coefficients. In Kaldi, we can compute MFCC as it provides us accurately computed values with the help of compute-mfcc-feats.cc program file.

- VAD (Voice Activity Detection)

When n x 13 Matrix is generated of the coefficients extracted are of voiced frames as well as unvoiced. To extract the user-specific the unvoiced frames coefficients are removed and the voiced frames coefficients are kept this step is known as VAD (Voice Activity Detection) [9]. To differentiate the voice frame & unvoiced frame the frames having more noise low log Mel energy are considered as unvoiced and frames having high log Mel energy are considered are voiced frames. In Kaldi, we can remove unvoiced frames with the help of compute-vad.cc program file which is included in Kaldi.

- I-vector Extraction

For I-vector Extraction GMM-UBM (Universal Background Model) based I-vector extractor is trained. With the help of which I-vector are extracted for creating a user-specific model. I-vector is known as Identity vectors which extract only speaker-specific characteristics from the given MFCC Coefficients and do not depend on the noise or phonetic contents [9]. For training a GMM-UBM I-vector extractor a large data-set of non-target persons is taken and extractor is trained. In Kaldi, we can extract I-vectors with the help of ivector-extract.cc program file which is included in Kaldi.

- PLDA Training

PLDA (Probabilistic Linear Discriminant Analysis) [10] is a classifier that is used to distribute the speakers with the help of

I-vectors by creating a probabilistic model which is later used to identify a speaker on the basis of maximum likelihood found during PLDA Scoring which is the performed during the testing of the speaker. The PLDA Training is done when the training of the user is done during registration/enrollment time. The probabilistic model that is prepared by the classifier consists of the sum of three things, speaker factor, inter- session factor & residual noise. PLDA Classifier between speakers & within speaker variability. In Kaldi, we can train the PLDA model with the help of ivector-compute-plda.cc program file which is included in Kaldi

- PLDA Scoring

PLDA Scoring [10] is done during the testing process when a single voice sample of the user is used to know the identity of the user. The already trained PLDA model is used to judge the identity of the user. To identify the user a threshold is determined if the likelihood is maximum & is greater the threshold the person is determined as the registered user. If there is no likelihood found the user is considered as not claimed. The likelihood is determined by matching the speaker I-vector extracted during training & test I-vector extracted during the testing. In Kaldi, we can extract I-vectors with the help of Ivector-plda-scoring.cc program file which is included in Kaldi.

The Speaker Training & testing steps are almost same the difference is that when the user is trained multiple utterances of the user are taken as voice samples PLDA model is trained but during testing, only one single utterance of the speaker is taken & instead of training PLDA model PLDA Scoring is done. During the registration process, only the speaker Identification model is trained for the user enrollment whereas the Speech recognition model is not trained every time

IV. System Working

Voice Reminder Assistant will work as an aide for the user to remind the task when some specific keyword/phrase is spoken by the person or the person reaches a specific location. The main features which differentiate Voice Reminder Assistant from other assistants such as "Google Assistant" and reminder systems.

The user will register into the reminder assistant by providing voice samples during registration process after registration user can avail the services of the assistant. The user needs to set a reminder for a specific keyword/phrase or location. When the claimed user speaks the exact keyword/phrase which was saved during reminder creation or reaches the location for which reminder is created the assistant will determine the condition for notifying the user to be true and notify the user by speaking to the user to do the task. For example, the user named "Ali" saved a reminder "Update the laptop drivers" for the keyword "Laptop" the assistant will notify the user by speaking "Hey Ali update the laptop drivers" if the claimed user "Ali" speaks the keyword "Laptop" or the user reaches the location "Sialkot" the assistant will notify the user about the task. For locating the position of user "Global Positioning System" is used.

The main feature of the voice reminder assistant is that it works for only that user who has registered in it using its mobile phone. If a user named "Ali" has registered into the application and signed into it. The assistant will not act on any other person voice it will only work for "Ali". Voice Reminder Assistant is a three-tier architecture system. The components included in the architecture are the following:

A. Server: Server consists of Speech Recognition Model and Speaker Identification Model when the user's voice is processed by the above models the server returns the speaker name and speech text to Android Application

B. Database: The User is authenticated by his email and password to sign in into the application, Every user has an email address and a password which he uses to log in to the application we have developed the assistant as an android application and the user has been provided the facility to use the application in another mobile without registering into the application again. The user can create the reminder manually or by using his voice, the reminders information and authentication credentials of the user are saved into the database.

C. Android Application: The Android Application uses the mobile microphone to record the voice of the user and sends it to the server. At the server after doing processing on the voice of the user the android application is returned the identified user name and the speech keyword/phrase spoken by the user , when android application receives the user name if it is the claimed user name the received speech text is verified that it to give a notification or create a reminder. Figure 5 explains the system architecture & how the components are interacting with each other.

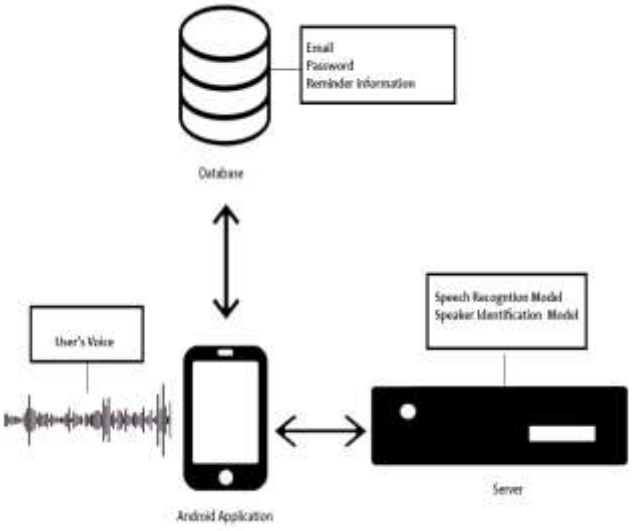

Fig. 3. Overview of Energy Policies of Pakistan.

V. Conclusion

Voice Reminder Assistant represented here is developed to give a reminder to the user on speech event and speaker

Identification. It is proposed to help the user to be not time bound and accomplish tasks easily. Currently, we have developed it in the English language it can be further developed into different languages which can be used by different people. Further GPS connectivity can be used further in different way in to the existing assistant, reminders can be created in which speech event will be detected for a specific location, for example, a user sets a reminder "give documents to the boss" that if he speaks "document" in his office location. As Speech based Applications have a wide scope nowadays many new concepts are being developed this assistant can be also developed further with many new modules and services.

REFERENCES

[1] Chowdhury, S., Mamun, N., Khan, A. A. S., & Ahmed, F, "Text dependent and independent speaker recognition using neural responses from the model of the auditory system," International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, Cox's Bazar, Bangladesh. Chowdhury, S., Mamun, N., Khan, A. A. S., & Ahmed, F: IEEE, 2017. pp. 871-874.

[2] Abhay Dekate, Chaitanya Kulkarni, Rohan Killedar, "Study of Voice Controlled Personal Assistant Device", vol. 42, no. 1, December 2016. [Online]. Available: https://www.ijcttjournal.org/2016/Volume42/number-1/IJCTT-V42P107.pdf . [Accessed Feb.1, 2019].

[3] Michaely, A. H., Zhang, X., Simko, G., Parada, C., & Aleksic, P, "Keyword spotting for Google assistant using contextual speech recognition."IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 Michaely, A. H., Zhang, X., Simko, G., Parada, C., & Aleksic, P: IEEE, 2017. pp. 272-278

[4] Sahu, P. K., & Ganesh, D. S, "A study on automatic speech recognition toolkits" International Conference on Microwave, Optical and Communication Engineering , December 18-20, 2015, IIT Bhubaneswar, India. Sahu, P. K., & Ganesh, D. S: IEEE, 2015. pp. 365-368.

[5] Denver, CO. Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at sixteen: Update and outlook. In Proc. IEEE automatic speech recognition and understanding workshop. Waikoloa, HI

[6] Reynolds D. (2009) Gaussian Mixture Models. In: Li S.Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA

[7] Anayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: IEEE. pp 5206–5210

[8] Martinez, J., Perez, H., Escamilla, E., & Suzuki, M. M, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques". CONIELECOMP 22nd International Conference on Electrical Communications and Computers, 2012: IEEE, 2017. pp. 248-251.

[9] Jamil, Mohsin, Asim Waris, Syed Omer Gilani, Bilal A. Khawaja, Muhammad Nasir Khan, and Ali Raza. "Design of

[10] Robust Higher-Order Repetitive Controller Using Phase Lead Compensator." IEEE Access 8 (2020): 30603-30614.

[11] Raza A, Akhtar A, Jamil M, Abbas G, Gilani SO, Yuchao L, Khan MN, Izhar T, Dianguo X, Williams BW. A protection scheme for multi-terminal VSC-HVDC transmission systems. IEEE Access. 2017 Dec 25;6:3159-66.

[12] Bashir N, Jamil M, Waris A, Khan MN, Malik MH, Butt SI. Design and Development of Experimental Hardware in Loop Model for the Study of Vibration Induced in Tall Structure with Active Control. Indian Journal of Science and Technology. 2016 Jun;9:21.

[13] Jamil M, Arshad R, Rashid U, Ayaz Y, Khan MN. Design and analysis of repetitive controllers for grid connected inverter considering plant bandwidth for interfacing renewable energy sources. In2014 International Conference on Renewable Energy Research and Application (ICRERA) 2014 Oct 19 (pp. 468-473). IEEE.

[14] Khan MN, Jamil M, Gilani SO, Ahmad I, Uzair M, Omer H. Photo detector-based indoor positioning systems variants: A new look. Computers & Electrical Engineering. 2020 May 1;83:106607.

[15] Kashif H, Khan MN, Altalbe A. Hybrid Optical-Radio Transmission System Link Quality: Link Budget Analysis. IEEE Access. 2020 Mar 18;8:65983-92.

[16] Zafar K, Gilani SO, Waris A, Ahmed A, Jamil M, Khan MN, Sohail Kashif A. Skin Lesion Segmentation from Dermoscopic Images Using Convolutional Neural Network. Sensors. 2020 Jan;20(6):1601.

[17] Uzair M, D DONY RO, Jamil M, MAHMOOD KB, Khan MN. A no-reference framework for evaluating video quality streamed through wireless network. Turkish Journal of Electrical Engineering & Computer Sciences. 2019 Sep 18;27(5):3383-99.

[18] Khan MN, Gilani SO, Jamil M, Rafay A, Awais Q, Khawaja BA, Uzair M, Malik AW. Maximizing throughput of hybrid FSO-RF communication system: An algorithm. IEEE Access. 2018 May 25;6:30039-48.

[19] Khan MN, Jamil M, Hussain M. Adaptation of hybrid FSO/RF communication system using puncturing technique. Radioengineering. 2016 Dec 1;25(4):12-9.

[20] Khan MN, Jamil M. Adaptive hybrid free space optical/radio frequency communication system. Telecommunication Systems. 2017 May 1;65(1):117-26.

[21] Andreas Nautsch (2014). Speaker verification using i-vectors, Evaluation of text-independent speaker verification systems based on identity-vectors in short and variant duration scenarios (Master thesis, University of Applied Science, Hochschule Darmstadt). Retrieved February 25, 2019, from https://www.dasec.h-da.de/wp-content/uploads/2014/06/Masterthesis_nautsch.pdf

[22] A. Kanagasundaram, R. J. Vogt, D. B. Dean, and S. Sridharan, "PLDA based speaker recognition on short utterances," in Proc. of Odyssey: The Speaker and Language Recognition Workshop. ISCA, 2012.