EasyChair Preprint

№ 3165

A Novel Approach for Generating Text Summary of Three Participants Spoken Audio

Ramesh Kagalkar and Basavaraj Hunshal

April 13, 2020

A novel Approach for generating Automatic Text Summary of Three Participants Spoken Audio

Ramesh Kagalkar
Dept. of CSE
KLE College of Engg. and Techn.,
Chikodi, India.
rameshvtu10@gmail.com

Basavaraj M. Hunshal
Dept. of CSE
KLE College of Engg. and Techn.,
Chikodi, India.
basavarajhunshal@klecet.edu.in

***Abstract*— Data Communication plays a vital role in transferring the information among the people. Auditory is a channel to communicate with each other. The aim of the paper is on computing framework for audio recognition which is further transferred to information and make it available wherever it is necessary by summarizing it. The sections used in this framework are classified into two modules called Training and Testing. Feature extraction of audio is performed under Training phase which further stored in the database. The Testing phase performs the features matching and classification producing the text description of the input audio file. The concept of topic modeling is applied on the text generated and its summary is derived as an output. The database stored for the system is of 100 audio files. The framework establishes a good technique of retrieving the text and generating its summary. The average accuracy rate of the designed system is about 90.06 % with three participant's consideration.**

Keywords— Text Summarization, Topic Modeling Procedure, Automatic Speech Recognition (ASR), Mel Frequency Cepstral Coefficient (MFCC), Support Vector Machine (SVM), Latent Dirichlet Allocation (LDA)

I. INTRODUCTION

Listening to the ever-growing amount of spoken audio sequentially is too slow. To access the content of computer media accurately and efficiently is a challenging task, because spoken audio combine the information from multiple levels such as phonetic, acoustic, syntactic, etc. Factors that affect the speech signals include room acoustics, channel and microphone characteristics and also the background noise. Although humans have developed mechanisms to deal with these factors, most of them are still very challenging for machines. These factors make the analysis of audio content a topic of on-going research. Hence easy browsing and retrieving of transmission content has become important in speech recognition. This paper proposed the idea of automatic speech recognition and generating its textual description. Mel Frequency Cepstral Coefficient (MFCC) algorithm is used for features extraction of the audio files. Extraction is easily performed by identifying all the linguistic content and discarding the other stuff like background noise, emotions etc. Further the Support Vector Machine (SVM) classification algorithm is used which performs the feature matching of the input audio file with the trained audio file features. The algorithm compares the input file features with all the trained audio files and concludes the matching with one of them which are nearly equivalent to the features of input file given. Finally the textual output is derived from the spoken content in an easy way. Topic identification from the derived text performed using Latent Dirichlet Allocation algorithm is also used for extracting a summary of audio content. The remainder of this article is organized as follows. In Section 2 we first provide the background knowledge regarding spoken content retrieval. The proposed methodology of automatic speech recognition (ASR) is explained in Section 3. The algorithms used are discussed in section 4. The database used for the system and the result of the proposed work is carried out in Section 5. Finally, the concluding remarks and the prospects for this area are derived in Section 6.

II. LITERATURE SURVEY

The related work offers the fundamental details required to understand audio content retrieval and a survey conducted on associated subjects, methodologies used, and also the issues identified have been mentioned.

[1] Discussed about extractive summary spotlights on perceive the vital information that is isolated and assembled to shape Text summarization. The phases included are: recognition of speech, extraction of feature from speech transcript, categorization, abstraction, and performance calculation. Auditory Features are separated from unrefined verbal communication signs are represented as audio / prosodic features.

Issues discussed by the author about the abstraction of a spoken document are Identify expressions, Human varieties, Semantic equality. Other issue to speech abstraction automation is how to tackle with detection results, together with word problems. The important issue discussed in speech recognition is about the transcripts generated by the recognition module may not be linguistically correct, and it is due to the recognition errors. Results of Text summarization are normally evaluated by ROUGE.

[2] Proposes an imaginative diagram based content Abstraction system for conventional single and multi document outline. It meant to lessen unique content reports to a short replacement summary which carries the most considerable particulars of the audio record. The summarizer has two Well-established semantic based content representation techniques, Semantic Role Labeling (SRL) and Explicit Semantic Analysis (ESA) just as the continually developing aggregate human information in Wikipedia. Proposed SRL-ESA based model for content summarization, as far as anyone is concerned, it is a principal study that combines Semantic Role Labeling and Wikipedia based express semantic analysis for text synopsis. It progresses the development of report similarity charts for graph-based content abstraction. The produced outcomes found significant execution upgrades in theory quality delineating the intensity of the job based representation of semantics and its mapping as being generated common ideas encoded in Wikipedia.

[3] Paper introduces a text mining context as well as its utilization for deep analysis of information conveyed by Politicians. In particular, they manage a professional

framework based discovery of the speech elements of enormous dataset of US Presidents' talks, extending from Washington to ongoing president Trump. The methodological commitment of the paper is in two phases.

1) Developing a methodology which is based on Text mining for the dataset development by tilling a scraping of web routine for gathering speeches.
2) Identifying the hidden structure of Speech information by creating Rank size routine upon each and every talks being the expressions of every discourse ranked as far as their frequencies.

The technique permits us to seek the changes into the speech structures without being definitely influenced by the changes in uses and implication of wordings.

[4] The literature plan is to consequently create 3 pictures. At first an interesting concept word will be presented, also the label mists identified with this idea will be imagined, and the last picture will introduce the mentality to this idea in reports (positive or negative). These three pictures make chronicles outlining the idea portrayal given in the arrangement of records. The Mechanism for identifying the story events are, the small chunk of data that becomes part of the content of any story.

- The procedure which is used to show story movements to the user.
- The collection of factors to change the story presentation of the events in content or all together.

The case study showed us the improved efficiency we can attain when we use classifier 1. Classifier 2 has only 489 negative statements and 111 positive statements as well as some negative words – 4783 and positive words – 2006.

[5] It aims for making Text Summarization utilizing Bidirectional gated repetitive unit strategy. The Bidirectional Gated Recurrent Unit (BiGRU) RNN architecture is used, so that it may correlate the surrounding words.

The result visualized as, from scenario-1 model there are 0.42 results effectively obtained words that become the important part of the input text as a desecrate form of result, not in a sentence form. However, both summarization results have low organization clarity between the statements. It has affected by most of the phrases in generated summary and it is repeated as well as statements constructed are with improper grammar. The models have missed punctuations like dot and comma. Therefore, the generated summary may be of poor quality.

[6] The main aim of proposed work is generation of Abstractive meeting summarization which covers significant focuses conversed in the discussion.

Methodologies discussed about the progressive versatile segmental encoder-decoder systems with various revising steps has been proposed in the paper for abstractive meeting summarization. The first define some essential ideas and phrasings to give a fundamental perspective of the abstractive meeting summarization problem. Then present the details of the proposed hierarchical adaptive segmental encoder-decoder network with numerous revising steps. When compared with summary obtained by HASMR(ml) method and generated summary by HASMR(rl) method is more fluent and expected. HASMR (rl) 48.64 17.45 22.13.

[7] The paper recommends the keyword in context measures just not improve the quality of summary producing in initial stage but it also proposes elite world algorithm to improve the key words in context.

The author has proposed Elite-Word Algorithm and the automatic summarization systems works with stereotyped method for handling. As an outcome, operations such as domain-specific content selection and content extraction are used to make short summary by retaining the originality.

The algorithm proposed will perform good to communicate as much statements in the article with predefined limit.

[8] The paper proposes an automated procedure for text analysis that works based on Fuzzy rules over extracted variety of features to yield the most significant data in the evaluated texts.

The presented technique summaries text by exploring correlated features to minimize dimensionality, and subsequently the number of Fuzzy rules are utilized for text summarization.

The approach proposed was compared with existing methods like naive baseline Score, Model and Sentence, using ROUGE measures. The results prove that the proposed method gives best f-measure (with 95% CI) in comparison with existing methods.

[9] The paper insights mainly on obtaining the text summary from only one text document using a modified Cuckoo Search (CS) Algorithm. For taking care of optimization issues in different areas the Cuckoo algorithm can be used. This can be used with sentiment score for summarizing the text record. The exploratory analysis uses benchmark database. Summarization can be done by applying various advances- Pre-processing methods comprises of division of sentences, tokenization, stop word expulsion, and stemming subtasks.

Then, the Guass CS and Fitness Function algorithms are used for extracting the Summarized yield of Text Document. The result of cuckoo search based model for Text Summarization (CSTS) has been noted as 43.11 and 13.98 for ROUGE-1 and ROUGE-2 respectively.

[10] The paper depicts a summarization framework built to auto-translate Korean speech into an English summary text. A Cross-Lingual Speech-to Text Summarization Method has been proposed by the author by categorizing two stages, in first stage Speech to Text is changed over by using few tools and used Google Translator for automatically recording a Korean transcript and translating it into English owing from its accessibility and elevated level of performance. Then the sentence grouping is done based on semantic comparability between two sentences. At last, the Multi-sentence compression technique is used for Summarizing Text.

The experimental results exhibit that our proposed strategy accomplishes a more significant level of comparison other gauge strategies.

[11] In the projected summarization approach, anomalies of text files are removed using a programming tool which the author has created called as KUSH that applies text summarization through a unique algorithm.

The aim is to design a robust, inventive framework to accomplish great Text summaries. In order to understand that the aim, author presented a novel document summarization method using graphs theory and entropy. The proposed Karcı Summarization technique has been made into two fundamental stages. To begin with, missing words that have no uniqueness (pronouns, prepositions, conjunctions and so on.) are excluded from the dataset. The text pre-processing tool that we developed, named KUSH, conducts certain normalizations.

The proposed approach provides text summary of around 100-words, yields considerably preferable outcomes compared to other approaches; and based on summary of 200-words, proposed methods gives still good outcomes in comparison to other competitive methods. For summary of 400-words, the ROUGE-1 metric values were the most higher over all terms, whereas the highest noteworthy qualities were required in ROUGE-L and ROUGE-SU metric values. The Luhn method achieved the second-best performance.

[12] The main objective of paper is to summarise the multi-document information into proper formate of summarised text with retained meaning. The author has proposed a statistical feature based abstraction technique which uses fuzzy model. Cosine similarity is proposed to avoid the redundancy and achieved significant performance.

After survey of all the above papers it is concluded that, using MFCC algorithm the features can be easily obtained with more precision compared to other methodologies discussed at above. Hence the survey is conducted on the verity of techniques and algorithm used for converting speech to text summary. At last we have combined and proposed a framework with techniques of speech to text recognition and ultimately generating summarized text.

III. SYSTEM OVERVIEW

The Audio file or live recording can be provided as input to the system and the system then mines the information according to the need of user from archived audio or speech. The extracted things from audio are audio signals. The user is allowed to upload its stuff which is in audio format. MFCC is used for feature extraction. The framework is proposed under of two phases importantly, 1) Training phase and 2) Training phase in addition to these, another phase used is Topic modeling.

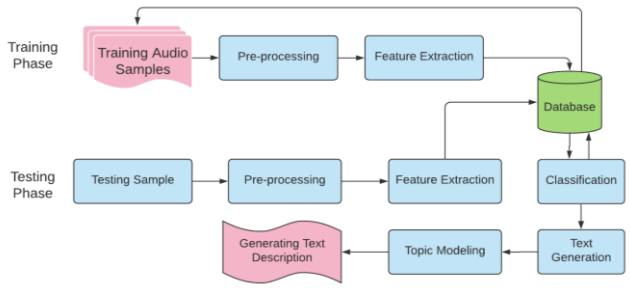

Fig. 3.1 Block diagram of system.

1) Training phase

In this phase system is trained by storing the samples into the database with its features and description which is required during audio testing. First, since audio is nothing but signals in wave form the audio is pre-processed and sliced into frames known as features. Audio features are important to train the system. More the number of features, more the time requires to process.

2) Testing phase

This phase checks the extracted features of audio and tries to match it with the feature pre-stored in the data repository. Here, the audio is analyzed and separated according to features and separated features are supplementary stored in the byte array. The .csv file can be used for storing the byte array. Afterwards, a speech or voice is taken as input. It compares the features with the trained data repository. For classification of the features SVM classifier is made used and finally the category of the audio domain will be predicted. At end the text is produced by evaluating the auditory content. Further noise is removed; the features present in Signals are extracted to detect objects.

3) Topic Modeling phase

This phase accepts the text content as input and gets the significant themes from it and further infers another content. The technique utilized here is the diverse meta-word extraction method. The most frequently appeared words are counted and output is determined in content. The generated output is the abstraction of the audio content given as input to system. Finally the output is the description of the given spoken matter.

IV. SYSTEM IMPLEMENTATION

The System is implemented in two phases by using various needed algorithms. The system predominantly separated into two phases First is Training phase and second is Testing phase, as per the phases the algorithms are been derived. At last, for topic modelling LDA algorithm is used. The algorithm for Training phase is represented in flowchart as below in fig. 4.1.

4.1 Training phase

It is used to pre-process the audio files and features are stored in the database. MFCC algorithm is used for feature mining process. Features are mined and its values are dumped into the database. The fig 4.1 shows the extracted features using MFCC algorithm.

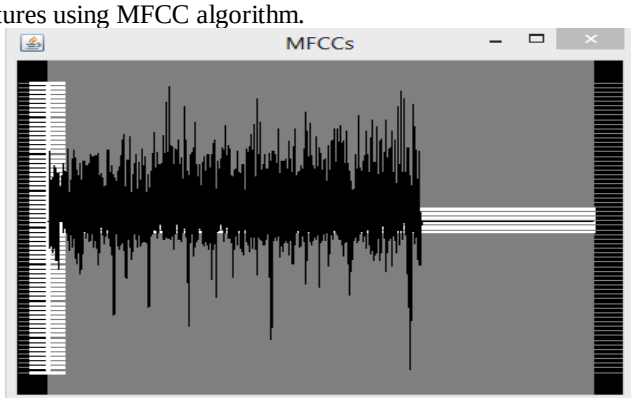

Fig. 4.1 Feature extractions by MFCC

- Flowchart for Training phase:

Main important techniques used in the Flowchart / algorithm of training phase are as below,

- Firstly MFCC is used to boost the amount of energy in the high frequencies. Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition.[3]
- The Hamming window: It is a decrease shaped by utilizing a raised cosine with non-zero endpoints, streamlined to limit the closest side projection. Number of points in the output window. On the off chance that zero or less, an empty array is returned. At the point when True, creates a symmetric window, for use in filter

design. The Hamming window has been utilized and it is increased by each frame to keep the congruity of the first and the last points in the Frame.

- Fast Fourier Transform (FFT): It is used for evaluating the frequency spectrum of speech FFT converts each frame of N samples from the time domain into the frequency domain. It also minimizes the calculation number. Discrete Fourier Transform (DFT) is normally computed via FFT algorithm.

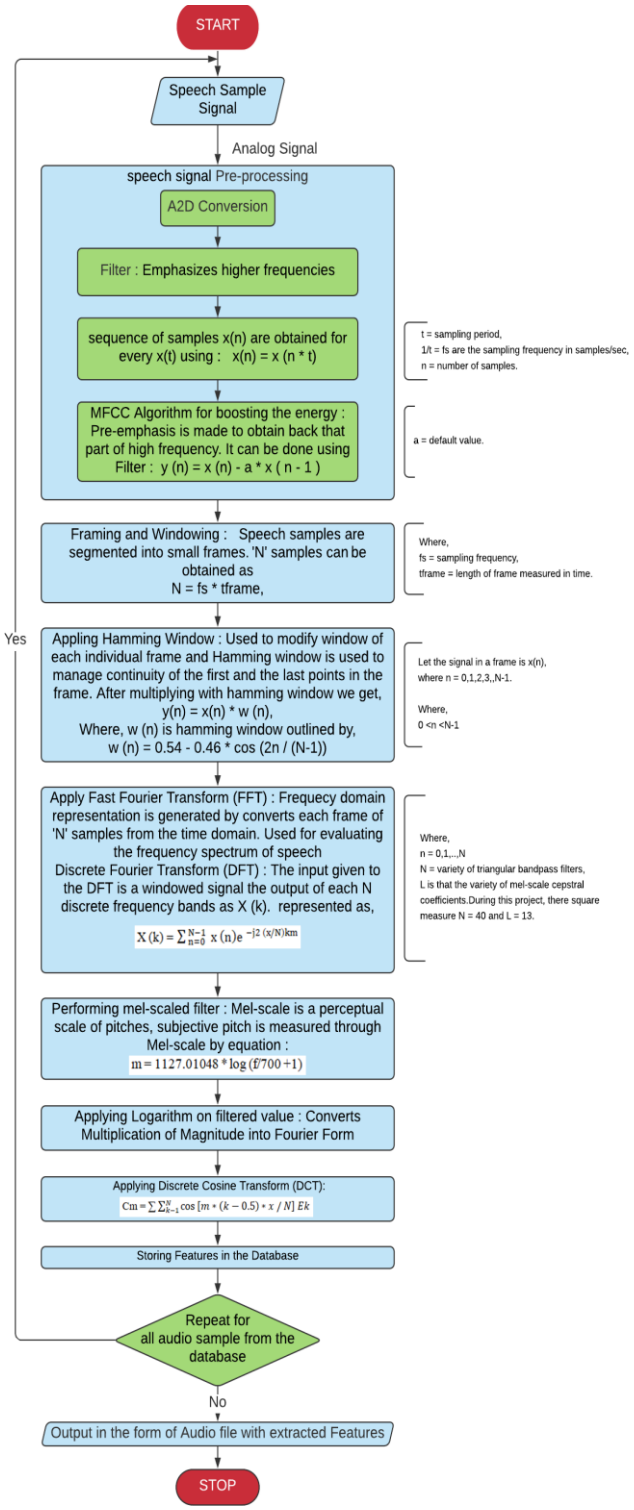

Fig. 4.2: Flowchart for Training phase

- Discrete Cosine Transformation: The DCT articulates a chain of finitely many data points in terms of a sum of cosine functions oscillating at various frequencies. A DCT is a Fourier related transform much like the DFT used to only float values. Right here we observe this at the strength acquired from 20 triangular band pass filters.

Input: spoken audio files.

Output: Features of each audio file are extracted and stored in the database.

4.2 Testing Phase

In testing phase an algorithm is used to test the features of the audio file and classify them using the SVM algorithm. In testing phase the audio file is tested with the trained audio kept in the database of the system. When the features of the audio approximately match with the other audio file, algorithm predicts the domain of the audio file and generates the text description.

By leaving last few steps remaining steps are same as testing phase. The complete algorithm is as follows.

Algorithm: Testing phase

Input: Sample Spoken Audio.

Output: Generating grammatically correct text description,.

Start

Step 1: Accept the sample Spoken Audio as input.

Step 2: Performing pre-processing of spoken audio.

Initially the analog form of signal is converted to digital signal. Then the digital signal is sliced into word samples and isolated word sample is passed through a filter which emphasizes higher frequencies. The sample sequence i.e., $x(n)$ is obtained from the continuous time signal $x(t)$ as,

$$x(n) = x(n * t)$$

Where,

t = sampling period,

$\frac{1}{t} = fs$ Is the sampling frequency in samples/sec,

n = number of samples.

To boost the energy level at the high frequencies MFCC algorithm is used. If you visualize the spectrum for speech segments more energy is present at low frequencies compared to higher frequencies.

Through pre-emphasis high frequency part is given back, it can be done by following filter,

$$y(n) = x(n) - a * x(n-1)$$

Where,

aa = constant value.

Step 3: Framing and Windowing.

The speech samples are segmented into small frames. Speech signal is thought to stay stationary in times of approximately 20ms. 'N' samples can be obtained as,

$$N = fs * tframe$$

Where,
fs = sampling frequency,
tframe = length of frame measured in time.

Step 4: Apply hamming window.

Windowing is a concept to build window of each one of frame, to reduce the discontinuity of the signal at the beginning and the end of each frame. The Hamming window is used, which is multiplied by each frame, it keep the continuity of the first and the last points in the frame. Let the signal in a frame is *x(n)*, where *n = 0,1,2,3,…,N-1*. After multiplying with hamming window we get,

$$y(n) = x(n) * w(n)$$

Where, *w(n)* is hamming window outlined by,

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

Where, 0<n<N-1

Step 5: Apply Fast Fourier Transform (FFT).

It is used for evaluating the frequency spectrum of speech FFT converts each frame of N samples from the time domain into the frequency domain. It also minimizes the calculation number. Discrete Fourier Transform (DFT) is normally computed via.FFT algorithm. The input given to DFT is a windowed signal the output for each N discrete frequency bands, as X(k) which represents the magnitude phrase of that frequency. It can be represented by;

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2(x/N)km}$$

Where, n = 0, 1, .., N

N = variety of triangular bandpass filters,

'L' is that the variety of Mel-scale cepstral coefficients.

During this project, there square measure N = 40 and L = 13.

Step 6: Performing mel-scaled filter.

The absolute values of DFT have been calculated. Mel-scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The speech signal consists of tones, with different frequencies and each tone with frequency f, its subjective pitch is measured through Mel-scale by the equation;

$$m = 1127.01048 * \log(f/700 + 1)$$

This equation shows the relationship between both the frequencies in hertz Mel scale frequency. For Mel-scaling a set of 20 triangular band pass filters or filter bank are used. Therefore the results of the FFT will be the information about the amount of energy at each frequency band.

Step 7: Performing logarithm on filtered value.

Step converts the multiplication of the magnitude in the Fourier transform into addition. Simply by using the command 'log' applied on the Mel-filtered speech segment, which gives the natural logarithm of the element.

Step 8: Apply Discrete Cosine Transform (DCT).

DCT expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. A DCT is a Fourier related transform similar to the DFT using only real numbers.

Here we apply this on the energy obtained from 20 triangular band pass filters. It can be represented as;

$$Cm = \sum \sum_{k-1}^{N} \cos[m * (k - 0.5) * x / N] Ek$$

Step 9: Classification using SVM (Computed features are compared with data features) Classification is performed and the features values are stored in a binary array by using,

$$y = f(x, \alpha)$$

Where,
'α' are the parameters of the function ,
Training set (x 1, y 1),..., (x m, y m) ,
f(xi) = Nonlinear discriminant function, where value of function is calculated as,

$$f(x_i) = w^T . \emptyset(x_i) + b$$

Where,

$\emptyset(x_i)$ is a nonlinear function and its maps the vector 'x_i' into a feature space of higher dimensionality.

Step 10: Sentence generation

Step 11: Apply LDA algorithm to generate summary of text

Stop

4.3 LDA algorithm

LDA is an unsupervised learning algorithm that perspectives reports as groups of words (ordered or unordered). LDA works by first making a key presumption: the manner in which a document was created was by picking a lot of topics and afterward for every topic picking a lot of words.

The flowchart shows execution of LDA procedure to build the summarized document in grammatically correct manner.

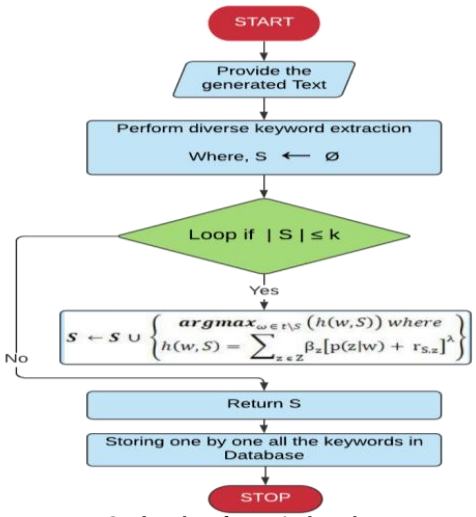

Fig. 4.3: Flowchart for LDA algorithm

V. DATABASE AND DESCRIPTIONS

Table 5.1: Dataset for the system

Sr. No.	Domain	Number of Audio samples	Description of Audio
1	Daycare	Daycare_Audio1	Children's are crying
		Daycare_Audio2	A woman is feeding to the child.
		Daycare_Audio3	Some children are playing.
		Daycare_Audio4	Some children are studying
		Daycare_Audio5	A maid is cleaning the room.
2	Garden	Garden_Audio1	This Audio is of garden and some people are walking on a track
		Garden_Audio2	There are trees and a gardener watering the plants.
		Garden_Audio3	Some children are playing in a play area.
		Garden_Audio4	An old man and woman is sitting on a lawn.
		Garden_Audio5	A boy is jogging on a track.
3	Bakery	Bakery_Audio1	A man ordered a list of bakery products
		Bakery_Audio2	A seller is selling bread.
		Bakery_Audio3	A boy is asking for a pizza bread
		Bakery_Audio4	A lady is selling a cake to a boy
4	Church	Church _Audio1	Marriage rituals in the church
		Church _Audio2	Morning prayers at the church
		Church _Audio3	A man and woman are praying
		Church _Audio4	Peoples are singing song
		Church _Audio5	A married couple is discussing about function
5	Library	Library_Audio1	Librarian is instructing the student
		Library_Audio2	Some students are arguing about the news in the newspaper
		Library_Audio3	A librarian is taking an entry of student.
		Library_Audio4	One boy is reading a book.
		Library_Audio5	Staff is talking on phone about the new arrival books

The dataset used for implementation contains approximately 500 audio from various domains. To evaluate text extraction process, some audio and its description are stored into the database to get best results shown in table 5.1. The database is made of text description of audio that are used as training. These audios are divided into various categories like shopping, street market, Water Park, playground, hospital, railway station, traffic signal, college, airport etc. It is maintained database of trained audio samples classified according to the domain. The table consists of domain name with the size of the file and the required processing. The audio content of the domain is also shown in the table.

All the above Audio Test Datasets are stored in the Database in .mp3 format. The description of the audio file is stored at last column it is required to identify the content of file. The maximum considered duration of an audio file is 10 sec. The database can be further upgraded to store bigdata files with the help of terabyte hard disk. Database is made by spoken audio that has been done using professional Digital Recorder: Head phone, In-built X/Y stereo mikes record. 4-channel real-time recording using built-in and external mikes, Digitally controlled, high-quality mike for improved audio quality, Large 1.9-Inch LCD screen and good user interface for easy operation, 24bit/96kHz Linear PCM recording for pristine recording, Recording/Playback Format: WAV, MP3, Storage capacity: Comes with 4GB mSD; Expandable to 32GB.

VI. RESULTS AND DISCUSSION

The audio file is taken as an input from the database. The database is storage for the trained audio files. When the input audio is provided they match the features of the audio using the MFCC and the SVM algorithm. The time required for deriving the final output is separately calculated depending on the step by procedure of the algorithm. The final output is derived in text. The algorithm keeps the track of the time taken for feature extraction, classification, topic modeling and text generation. The table 6.1 shows the results of the different audio files with their expected output and derived output. Up to 3number of participants conversation is recorded and stored.

In table 6.1, audio samples of 5 different domains are taken as an input. For experiments, we have taken time duration up to 10 second. The expected description of the daycare_Audio file is "Madam I am going to office. Leave your child here and pick him at 2 p.m. Ok. His dad will come to pick him. Do send him early. Yes Madam I will come surely", where the actual output derived is "Madam am going to office. Leave your child here and pick him p.m. His dad come pick him. send him early. Madam I will come surely". There is slightly difference in the output. The derived output appears somewhat grammatically incorrect as the system skips some or the other words from the conversation. Similarly for the rest of the audio file the output derived is shown in the table.

Table 6.1: Result analysis for 3 participants.

Sr. No.	Audio Sample	Duration of Audio (in sec)	Expected Result	Actual Result
1	Daycare_Audio	6	Madam I am going to office. Leave your child here and pick him at 2 p.m. Ok. His dad will come to pick him. Do send him early. Yes Madam I will come surely.	Madam am going to office. Leave your child here and pick him p.m. His dad come pick him. send him early. Madam I will come surely
2	Garden_Audio	10	Hi raj. You want to play with me. Yes we will play cricket. No we will play volleyball. Guys let have a toss for that.	Hi raj. want to play with me. Yes will play cricket. No we will play volleyball. Guy's lets have toss.
3	Bakery_Audio	7	Sir do you have Cake. Yes madam we have all fruits flavor in it. Son do tell which you want. I want pineapple cake. Ok you will get it.	Sir you have Cake. Yes madam we have all fruits flavor it. Son do tell which you want. I want pineapple cake. Ok will get it.
4	Church_Audio	9	Please come fast otherwise you will miss you prayer. Please wait for me else I wouldn't reach there fast. Do come soon church is far from here. Ok I will. Hey don't worry I will help you in reaching there.	Please come fast otherwise you will miss you prayer. Please wait for me else reach there fast. Do come soon church is far from here. Ok will. Hey don't worry I will help you in reaching there.
5	Library_Audio	9	Madam I want to return this book. Ok. Fill that register with the book name and today's date. Madam first issues me a book since I am late for my class.	Madam I want to return this book. Ok. Fill that register with the book name and today's date. Madam first issues me a book since I am late for my class.

The derived output is further pass to the LDA algorithm to identify the topics from the text and conclude a summary of the audio conversation into text. The summary derived is also in English language. The table 5.3 shows the topics identified by the LDA algorithm and finally the text description is derived.

Table 6.2: Generating text description.

Sr. No.	Audio Sample	Actual Result	Topic Identified	Text Description
1	Daycare_Audio	Madam am going to office. Leave your child here and pick him p.m. His dad come pick him. send him early. Madam I will come surely	Madam, Office, Child, Dad, early	This is conversation related to business development between Employee and the boss.
2	Garden_Audio	Hi raj. want to play with me. Yes will play cricket. No we will play volleyball. Guy's lets have toss.	Raj, Play, Cricket, Volleyball toss	This conversation is between doctor and the patient. Doctor is giving medicine to the patient.
3	Bakery_Audio	Sir you have Cake. Yes madam we have all fruits flavor it. Son do tell which you want. I want pineapple cake. Ok will get it.	Sir, Cake, Madam, Fruit, Son, pineapple	This conversation is from the market where a lady is asking and bargaining for potatoes to the shopkeeper
4	Church_Audio	Please come fast otherwise you will miss you prayer. Please wait for me else reach there fast. Do come soon church is far from here. Ok will. Hey don't worry I will	Please, Prayer, Wait, Fast, church	A daughter is asking for a new dress to her mom while the mom is saying that she will get it on her birthday.

		help you in reaching there.		
5	Library_Audio	Madam I want to return this book. Ok. Fill that register with the book name and today's date. Madam first issues me a book since I am late for my class.	Madam, Book, Register, Name, Date, class	This conversation is between the student and the teacher where the teacher is asking the student to be ready for the surprise test and one of the student is saying no as they have their PT class.

In table the audio sample 1 is of Daycare domain where the actual output derived is "Madam am going to office. Leave your child here and pick him p.m. His dad come pick him. send him early. Madam I will come surely".

After the topic identification the text generated is "This is conversation related to business development between Employee and the boss", which is the summary of the business audio.

Table 6.3: Recognition rate for 3 participants.

Sr. No.	Audio Sample	Number of Participants	Processing Time (in sec)	Recognition rate
1	Daycare_Audio	3	30.474	Expected Word Count = 16 Actual Word Count = 13 Recognition Rate = 81.25 %
2	Garden_Audio	3	32.751	Expected Word Count = 18 Actual Word Count = 16 Recognition Rate =88.8 %
3	Bakery_Audio	3	63.497	Expected Word Count = 23 Actual Word Count = 20 Recognition Rate =86.95 %
4	Church_Audio	3	32.499	Expected Word Count = 30 Actual Word Count = 28 Recognition Rate =93.33%
5	Library_Audio	3	31.390	Expected Word Count = 18 Actual Word Count = 18 Recognition Rate =100%

The table 6.3 depicts the audio samples set out from table 6.1 with the actual word count verses expected. The number of participants considered in the dataset is 3. Recognition rate (in %) obtained is deliberated for each audio sample by calculating the actual and expected word count. The following formula can be use,

$$\text{Recognition Rate} = \frac{\text{Expected word count}}{\text{Actual word count}} * 100 \ \ \%$$

The system performance depends on processing time and recognition rate. The recognition rate can be considered as overall accuracy of the system. To find the average recognition rate of audio samples containing 3 participants the formula used is,

$$\text{Avg. recognition rate (RR)} = \frac{\sum_{k=1}^{n} \text{RR of Audio Samples}_k}{\textit{No.of Audio Samples}}$$

Where '*n*' is no. of samples and '*k*' is variable.

$$= \frac{81.25 + 88.8 + 86.95 + 93.33 + 100}{5} = 90.06\ \%$$

Therefore, the avg. recognition rate for all audio samples contains two participants is 90.06 %.

A) Result Analysis for Audio Samples

The time taken for feature extraction by MFCC, classification by SVM and topic modelling by LDA algorithm is discussed in the next section.

a) Processing Time Required for 3 Participants

The audio samples taken as an input contains 3 numbers of participants. The table 6.4 shows the processing time required for feature extraction, classification, topic modeling and text generation. The processing time of an audio sample is the sum of time required for feature extraction, classification, topic modeling and sentence generation of that sample.

Table 6.4: Processing time for audio sample of 3 participants.

Audio_Sample	Processing Time			
	Feature Extraction	Classification	Topic Modeling	Text Generation
Daycare_Audio	1.55	28.344	0.643	0.21
Garden_Audio	1.765	30.543	0.342	0.1
Bakery_Audio	1.876	29.487	0.987	0.34
Church_Audio	1.543	30.213	0.643	0.10
Library_Audio	1.113	29.543	0.534	0.2

In the table 6.4 the time for feature extraction, classification, topic modeling and sentence generation is shown as the overall processing time for particular audio. The graph is plotted showing the processing time for all the audio samples from the table 6.2.

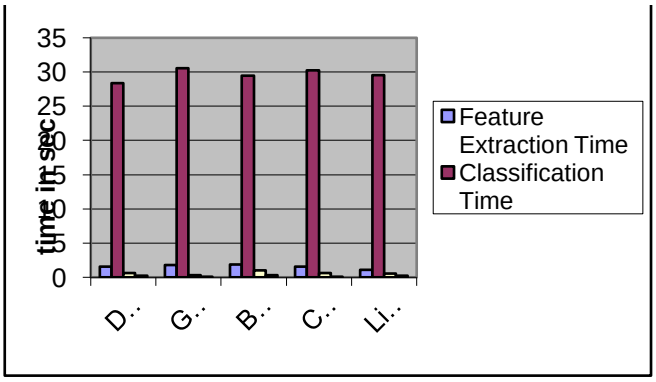

Fig. 6.1 Graph showing processing time

The figure 6.1 shows the feature extraction, classification, topic modeling and text generation of audio sample.

b) Recognition Rate for 3 Participants

The figure 6.2 shows the recognition rate for the 3 participants for each input audio sample. The table 5.4 shows the total processing time required for different 5 audio samples. The samples taken are related to daycare, garden, bakery, church and library domains.

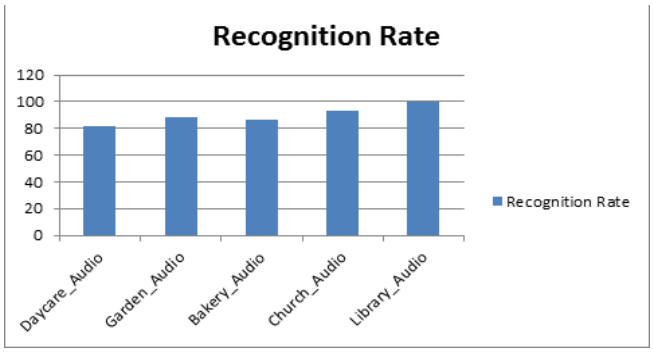

Fig. 6.2 Recognition rate for 3 participants

VII. CONCLUSION

The Proposed system has introduced natural language descriptions of long audios by SVM classification for complex audio containing up to 5 numbers of participants. The process uses feature extraction, classification, topic modeling and text generation. Each audio split into frames at some micro second intervals and the filtering, noise removal is applied on every features. Features are mined using MFCC algorithm and these features are used for comparison of testing with the training audio and give the description of audio. The performance of the implemented system can be calculated and gives the accuracy of 90.06 %. In future work, the system can be carried out for audios containing more number of participants to generate grammatically correct text description. The audio can be extended for long duration and also the system can be modified to support real time audio translation to produce grammatically correct text description.

REFERENCES

[1] "Speech Summarization for Tamil Language", A. NithyaKalyani and S. Jothilakshmi, Intelligent Speech Signal Processing, © 2019 Elsevier Inc.

[2] "SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis", Muhidin Mohamed, Mourad Oussalahb, eceived 10 August 2018; Received in revised form 20 February 2019; Accepted 12 April 2019, 0306-4573/ © 2019 Elsevier Ltd.

[3] "A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches", Valerio Ficcadenti, Roy Cerqueti, Marcel Ausloos, Elsevier, Expert Systems with Applications 123 (2019) 127–142.

[4] "Text Summarization For Storytelling: Formal Document Case", Piotr Janaszkiewicz, Justyna Krysińska, Marcin Prys, Magdalena Kieruzel, Tomasz Lipczyński, Przemysław Różewski, Scince direct, Elsevier Procedia, International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018, Belgrade, Serbia.

[5] "Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit", Rike Adeliaa, Suyanto Suyantoa,, Untari Novia Wisesty, Scince direct, Elsevier Procedia, 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12-13 September 2019.

[6] "Abstractive meeting summarization by hierarchical adaptive segmental network learning with multiple revising steps", Jiyuan Zheng a , Zhou Zhao a, Zehan Song a , Min Yang b , Jun Xiao a , Xiaohui Yan, Neurocomputing, Elsevier, November 1, 2019;21:57

[7] "Abstractive Text Summarization of Research Articles Based on Word Associations", G. Dineshnath and S. Saraswathi, Journal of Computational and Theoretical Nanoscience Vol. 16, 1–4, 2019

[8] "A text summarization method based on fuzzy rules and applicable to automated assessment, Fábio Bif Goularte, Silvia Modesto Nassar, Renato Fileto, Horacio Saggion, Elsevier, Expert Systems with Applications 115 (2019) 264–275.

[9] "Text Summarization Technique by Sentiment Analysis and Cuckoo Search Algorithm", S Mandal, GK Singh, A Pal - Computing in Engineering and Technology, 2020 – Springer.

[10] "Cross-Lingual Korean Speech-to-Text Summarization", HyoJeon Yoon1, Dinh Tuyen Hoang1, Ngoc Thanh Nguyen2, and Dosam Hwang, Springer Nature Switzerland AG 2019, ACIIDS 2019, LNAI 11431, pp. 198–206, 2019.

[11] "Karcı summarization: A simple and effective approach for automatic text summarization using Karcı entropy", Cengiz Harka, Ali Karcı, Elsevier, Information Processing and Management 57 (2020) 102187, Accepted 18 December 2019.

[12] Darshna Patel, Saurabh Shah, Hitesh Chhinkaniwala, "Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique" Expert Systems With Applications 134 (2019) 167–177, Elsevier.