2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Instant Bi-Lingual Captions

Mahesh Dhumal, Hemant Kumar Kushwaha, Vaibhav Gupta, Sona R. Pawara
Department of Computer Engineering
STES's Sinhgad Academy of Engineering
Pune, Maharashtra, India
mahesh90010@gmail.com, hkushwaha345@gmail.com, vaibhavgupta710@gmail.com, srpawara.sae@sinhgad.edu

***Abstract* — With rapid development of the internet, the tremendous growth has been observed in video sharing platform. Millions of users have been using this platform as source of knowledge, entertainment, self-development, but language barrier has confined the reach of this platform to certain amount of population. Also the hearing-impaired people could not make satisfactory use of this platform. Subtitles play important role in such scenarios, with the help of which people can understand the content of the video. Various applications have been developed for subtitle generation to resolve issue of language barrier and hearing impairment. In this paper we have surveyed different methods for speech recognition and bilingual subtitle generation and proposed a system for instant bi-lingual subtitle generation based on MFCC and HMM.**

Keywords*— *Mel Frequency Cepstral Coefficients(MFCC), Support Vector Machine(SVM), Hidden Markov Model(HMM), Sentence Boundary Detection(SBD), Continuous Density Hidden Markov Model(CDHMM), Large Vocabulary Continuous Speech Recognition(LVCSR), Artificial Neural Network(ANN),Power-Normalized Cepstral Coefficients (PNCC)

I. INTRODUCTION

With exponential growth of Internet, right from 16 million users in 1995 which was 0.4% of total population to 4,156 million users in 2017 which is 54.4% of population. video sharing platform has also observed tremendous growth. Such vast resources of video sharing platform cannot be properly used by people due to language barrier or due to hearing-impairment. Subtitles provide them way to understand the content of video.

Current video sharing industry is mainly using manual subtitle generation system but this subtitle generation system makes the huge wastage of time as well as resources. Though Some portion of industry is adopting the automatic subtitle generation system, but this system can't be implemented for live videos as it takes the time nearly twice as that of the length of video. Also this system is primarily focused on subtitle generation in English which may not help entire the population. Number of frameworks based on advanced speech recognition techniques involving MFCC, PNCC etc have been proposed for automatic video subtitling. The conventional framework composed of MFCC and HMM works efficiently in low varying conditions but it fails in highly varying environment. There is another framework which can be used for speech recognition based on PNCC and HMM which is robust to noise , but it takes slightly higher computational cost as compare to traditional MFCC. To resolve such issues a system is needed which will significantly reduce the cost of subtitle generation and also the time required for processing. Such system will save the huge amount of resources of video sharing industry and will generate the subtitles in the language specified by user without any delay.

This paper is bases on the comparisons between various algorithms to understand merits and demerits of each and develop an application providing higher accuracy in results. The survey done in this paper composed of the researches done by various analysts, their methodologies, theoretical formulae and the conclusions they have arrived at.

II. LITERATURE SURVEY

For recognition of speech signal and generation of subtitle Su Myat Mon and Hla Myo Tun in 2015 [1], implemented a Speech-to-Text conversion system using MFCC for feature extraction and HMM as the recognizers. MFCC is used to extract features from the speech signals of isolated words. And, HMM method is applied to train and test the audio files to get the recognized spoken word. The proposed system is divided into four main steps: speech database, pre-processing, feature extraction and recognition. Starting with first step, five audio files are recorded and each audio file contains ten different pronunciation audio files. These speech samples after pre-processing are extracted to features using MFCC. Finally, HMM is used to classify the desired spoken word using MFCC coefficient as input. The experiments showed that the average % accuracy/recognition rate is most and better if the number of states(audio files) are five.

Aleš Pražák, J.V. Psutka, Jan Hoidekr, Jakub Kanis, Ludˇek Müller, and Josef Psutka in 2006 [2], implemented an Automatic Online Subtitling of Czech Parliament Meeting using a LVCSR system. The recognition system is based on HMM, lexical trees and bigram language model. The acoustic model is trained on 40 hours of parliament speech and the language model on more than 10M tokens of parliament speech transcriptions. The application framework consists of DirectShow filter. There are 2 detached DirectShow filters used – one for audio and one for video stream. The audio DirectShow filter acquires the speech signal from any media type, which is then passed directly to the LVCSR system engine. The recognized word sequence is forwarded via system pipes to the subtitle displayer implemented as a video DirectShow filter. Video DirectShow filter incorporates the recognized subtitles to the source video stream. Each video frame, transparent subtitle bitmap is combined with the source bitmap. The test data consisted of five different parliament speech, half an hour each. The recognition accuracy of the speech depends on a discussed topic, with a varying accuracy of 80 to 95%.

Priyanka P. Patil and Sanjay A. Pardesi in 2014 [3], developed a Marathi connected speech recognition system, using MFCC feature extraction technique and CDHMM. The objective of this system is to develop speech

recognition system for Marathi language for educationally underprivileged people or illiterate rural communities. Initially, the recorded speech is segmented into different words using the speech segmentation algorithm. The two basic features of speech signal on which speech segmentation algorithm depends are : Short Time Energy(STE) and Spectral Centroide(SC). MFCC technique is used to extract features from the speech signals of isolated words. The CDHMM for each word of speech signal is developed using these extracted features. Observation properties generated by each state is defined by Gaussian mixture density function with model parameters being re-estimated by Baum-Welch algorithm. Bi-gram pairs are found using Bi-gram language model. Finally the maximum likelihood state sequence path is found using the log Viterbi beam search.

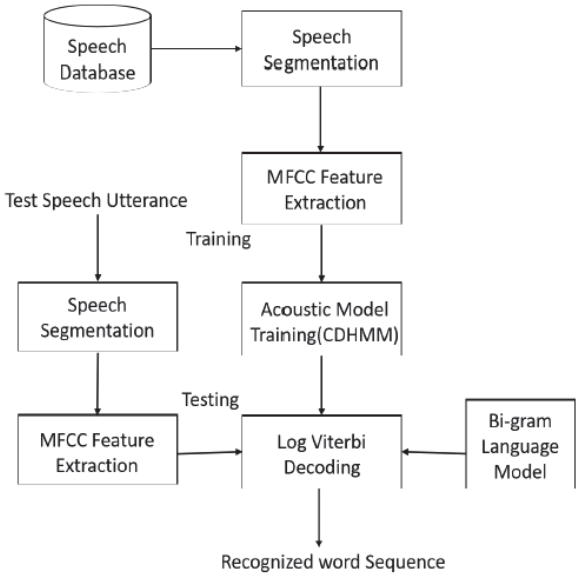

Fig. 1. System Block Diagram

The experiments showed, if energy value and centroid value of frame is greater than threshold T_E and T_C then respective frame is termed as voiced, otherwise unvoiced. The recognition rate of the system is calculated using he given formula:

Recognition rate big(%big) = No. of correct utterances / Total No. of test utterances

The system used combination of CDHMM and MFCC giving better accuracy than other techniques.

Yogita H. Ghadage and Sushama D. Shelke in 2016 [4], designed a multilingual Speech-to-Text conversion system with focus on being Marathi-English mix speech. The objective of this system was to present a Speech-to-Text conversion system for Marathi, English, Marathi-English mix languages for illiterate rural communities or educationally under-privileged people. This work is based on MFCC, SVM and Minimum Distance Classifier. The input speech signal is given to the MFCC which converts it into feature vectors. Further the word recognition and classification purpose are done using Minimum Distance Classifier and SVM techniques. The % accuracy achieved for the proposed system is higher as compared to the one using MFCC feature extraction technique and CDHMM classifier.

Ibrahim Patel1 Dr. Y. Srinivas Rao2 in 2010 [5], proposed system to improve the representation of speech feature in HMM based system by recognising the speech signal using data of frequency spectral with Mel frequency. In the various techniques for extracting speech parameter for efficient speech recognition, MFCC technique along with advance recognition method like HMM is most widely used, but this technique may fail in highly varying conditions. Thus to avoid such scenarios this system integrates sub band decomposition technique which is a method for frequency isolation, to the MFCC method to extract speech feature. Initially this system computes a measure matrix for dissimilarity measurement of HMM, and then the K clusters are obtained by applying clustering technique to the measurement matrix computed earlier. In each step new clusters are generated by merging the previous stage clusters.

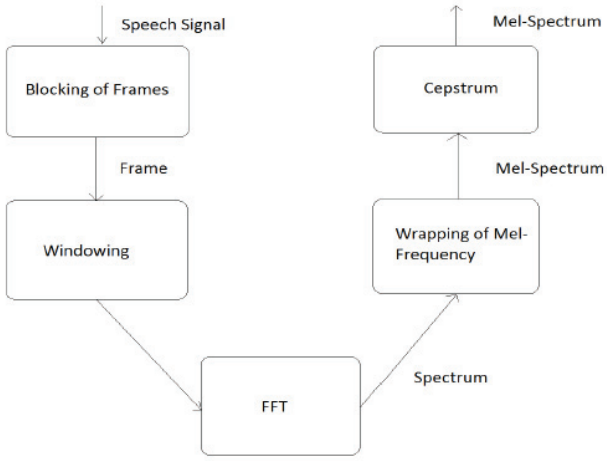

Fig. 2. Model for speech processing

By using this method set of Mel-frequency cepstral coefficient is calculated for 30 milli-sec speech frame which are also known as acoustic vectors.

These vectors provide the set of training vectors. The collection of word are maintained for training of the HMM network. This system provides an approach for speech recognition which is robust to noise.

Chanwoo Kim and Richard M. Stern in 2016 [6], proposed power normalized cepstral coefficient algorithm which is new feature extraction approach. PNCC makes the use of power-law nonlinearity instead of traditional log nonlinearity which is there in MFCC.PNCC also involves algorithm for suppression of noise which is based on asymmetric filtering. In noisy conditions the accuracy of PNCC is way better than vector Taylor series ETSI advanced front end. Initial stages in PNCC processing are similar to that of MFCC and PLP except it uses gammatone filters for frequency analysis. After the completion of frequency analysis, non-linear time varying operations are performed using longer-duration temporal analysis in order to reduce noise. Processing of PNCC which differs from MFCC and PLP can be shown in Fig. 3

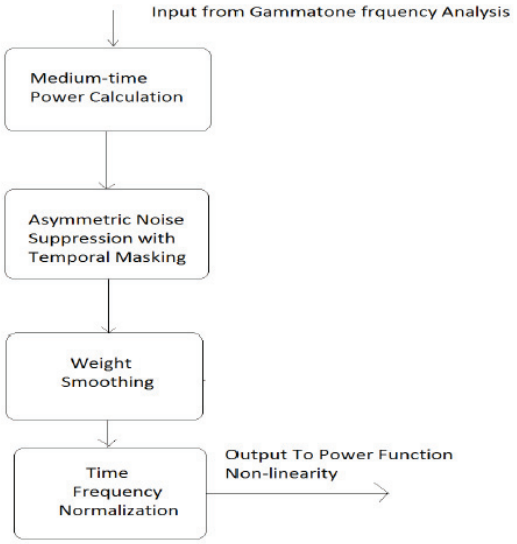

Fig. 3. Addition processing in PNCC Compared to MFCC

If the above mentioned section is omitted from the PNCC processing then the remaining processing section in PNCC is called as simple power-normalized cepstral coefficient. Final stages of PNCC and MFCC are also similar except PNCC uses Power-law nonlinearity with exponent 1/15 rather than using traditional log non-linearity. This algorithm provides more accuracy in noisy environment as compared to MFCC with 33% higher computational cost than MFCC.

Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik and Supriya agarwal in

2018 [7], reviewed different existing algorithms and techniques that are applied to achieve various functionlities such as articulatory and acoustic-based speech recognition, conversion from speech signals to text and vice-versa.

1. Basic Speech Recognition Model

The speech recognition systems follow some standard steps as shown in Fig. 4.

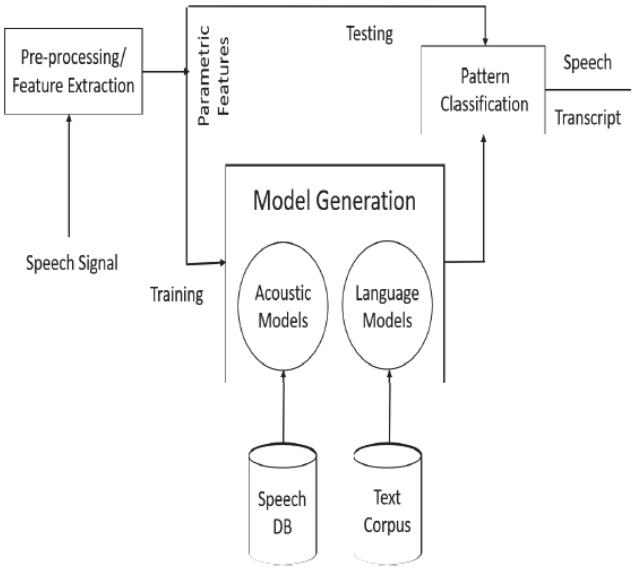

Fig. 4. Architecture for Speech Recognition System

2. Speech To Text Conversion Methods

It is the process of converting spoken words from speech into written texts. STT is done using some conversion methods mentioned below:

HMM - Simple, automatically trained and computationally feasible to use.

ANN - Simple, fast convergence rate, increase the recognition accuracy of the speech recognition system.

3. Text To Speech Conversion

It is a conversion process in which input text is analyzed, pre-processed and understood, and then is converted into speech signals. Various steps involved in TTS Conversion are shown in Fig. 5.

After looking up closely at different algorithms and techniques, a conclusion was made stating, HMM is a better STT technique because of its computational feasibility and use of parallel and cascade synthesis works the best under TTS systems.

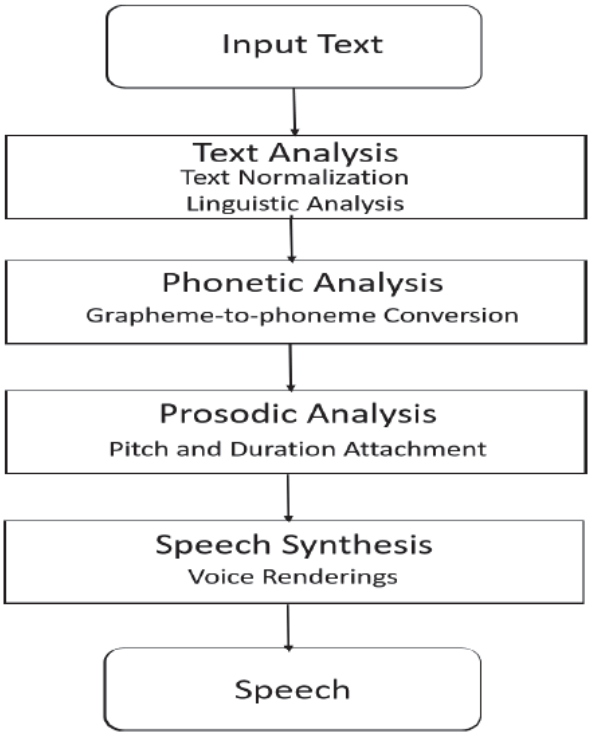

Fig. 5. Text To Speech System Flow

Xiaoyin Che and Sheng Luo in 2017 [8], implemented the integrated framework for automatic bi-lingual subtitle generation.

This framework involves Sentence boundary detection(SBD), machine translation(MT) , Automatic speech recognition(ASR).

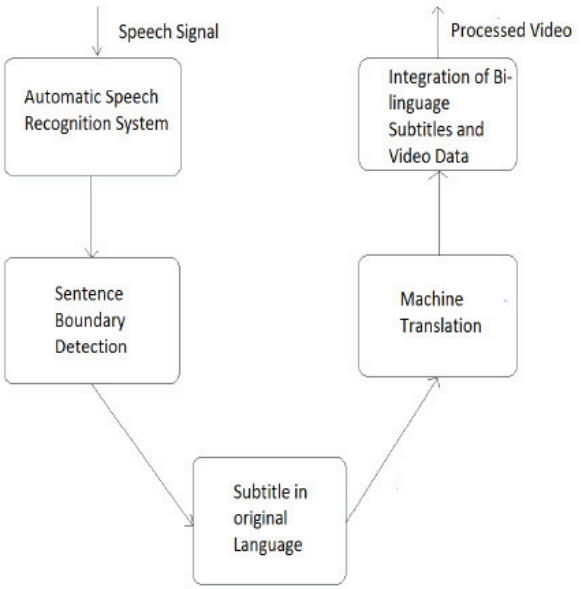

Fig. 6. Working model For speech recognition

This framework uses the IBM Watson Speech-to-Text API for speech recognition. This service provides the transcript file within 1.5 time of audio duration. This framework ignores the default segmentation done by ASR system and arranges all the words in list as per their timestamp. Sentence boundary detection mechanism uses the state-of-the-art lexical model and DNN for segmentation of transcript generated from ASR system. This framework uses Microsoft Translator API as machine translation tool. Text content of the subtitle in original language generated in previous stage will be submitted to server and the resulting text is directly added as second line in subtitles. If the proposed framework is applied, the total working time in preparing bilingual subtitles can be shortened by approximately 1/3, with no decline in quality.

Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk in 2012 [9], described an approach of speech recognition by using MFCC. In this paper, PCA technique is applied to MFCC features for extraction of the most significant components from those features. The adequate data training and testing set required for next classification state is built using analyzed MFCC features which are projected onto a two dimensional space. Maximum Likelihood(ML) and SVM, two classifiers, are selected to train and test on two dimensional MFCC dataset and then are compared to each other for correct classification. The tests conducted on all different percentages of dataset for recognitions showed that distributions of SVM classifier seems to give more tense, consistent and reliable performance than those of ML.

In [10] an approach has been specified to resolve the issues regarding the HMM framework like discarding the information related to time dependencies, being prone to over-generalization using straightforward template matching. Recognizer in this approach is based on Dynamic time wrapping algorithm, as the continuous speech recognition will result in explosion of search space , the conventional top-down search is complemented with data-driven selection for DTW alignment. This template-based framework provides flexibility of unit selection which leads to the new approaches to speaker and environment adaptation. The combination of HMM and DTW will lead to decrease in word error rate by 17% as compare to HMM results.

Mercedes de Castro, Diego Carrero, Luis Puente, Belen Ruiz. Universidad Carlos III de Madrid in 2011 [11], described about the application of ASR in live television system and in synchronizing system. It also tells about the challenging issues faced while accessing the live multimedia and the disturbing effect which causes the lack in synchronization for the live multimedia. It also tells about how subtitle helps a hearing-impaired people in daily life to watch television and access the audio content. It tells about the need for the subtitle in daily life. It talks about the subtitle delay in live video the setup consists of two phases, presentation phase where the tv is connected to the broadcast and the presentation phase which consists of the setup box. Here the live video is synchronized to a particular channel along with the video and the subtitle is delivered to a dedicated IPTV channel. It tells briefly about how ASR is applied to live TV programs. Favorable condition has led to 90-90% of success rate in speech recognition technology. Practically ASR still produces some delay. It discusses about the various factors that determine the complexity of subtitle generation. In real time subtitle generation (Fig 7), encoding and packetization of input audio is done in parallel along with the video signal then the audio signal is processed to obtain the transcript which is then converted into subtitles.

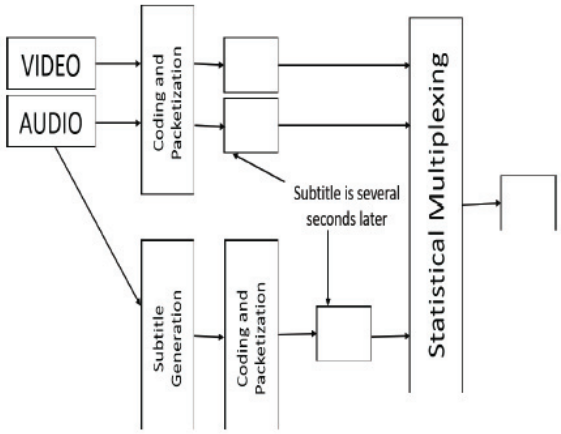

Fig. 7. Transcriptions process

In this project ASR has been applied directly to the audio of a TV channel to characterize and compensate the individual delays introduced by the live subtitling process. This paper is based on the experiments on two main types of TV programs: live interviews with several speakers at TV studio and live interviews with a TV correspondent during TV news program the average results for both the programs was roughly around 68% and 81%. Here the Subtitle delays calculated by considering the difference between audio utterances and their corresponding text transcriptions, without considering further processing time required by typical live subtitling processes like editing, partitioning, encoding and packetization. Average delay is 3.84 seconds. Standard deviation is 3.42 seconds. The system is able to generate a DVB/MPEG signal with subtitles in sync with

the original audio/video to be broadcast with a slight delay compared to the original signal, via an additional, user-selectable IPTV channel. The objectives of the project also include research in the fields of applied ASR in live TV selected scenarios. The functional steps in this project are:

(a) Detection of broadcast terrestrial TV input channels

(b) Extract the AUDIO signal from the input channel

Andrew Lambourne*, Jill Hewitt, Caroline Lyon, Sandra Warren in 2004 [12], describes about the speak title project, how it met the challenges which occurred while performing speech recognition and live-subtitling. Main purpose was to investigate and develop a system to provide subtitling service on live television. Here test users were considered and project was in a evaluation phase. The project goal was to use speech recognition technology to focus on real time speech recognition and deliver a transcript with a minimal delay and high accuracy.

Limitation: topic specific vocabulary files need to be provided

Criteria: accuracy 97-98%, throughput delay 5-6 sec

LUUK VAN WAES , MARIËLLE LEIJTEN, ALINE REMAEL in 2013 provided an approach to understand the causes and consequences of reduction in quantitative text reduction in live subtitling.

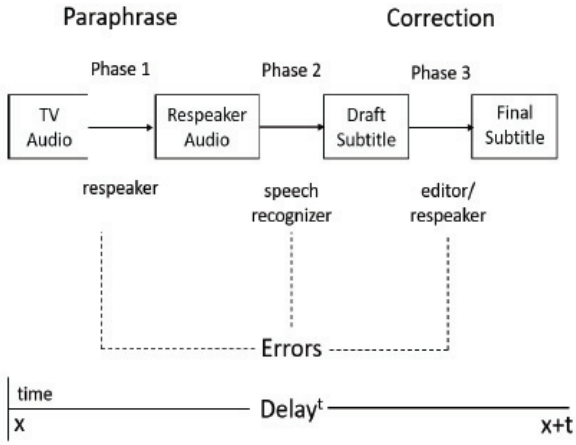

Fig. 8. Temporal representation of the live subtitling process

Experiment has been performed where three experts from infotainment talk show were subtitled by twelve repeaters. Different features such as reduction percentage, delay and also the measures of speakers working memory were collected. From the experiment it is concluded that quantitative reduction is not random process, but largely determined by external factors such as delay, amount of source text.

In [14] Major technological perspective and evolution in technology for speech to text conversion has been given. This also discusses technique developed in each stage of classification of speech to text conversion. This system provides on-line speech-to-text engine. Speech signal will be acquired by the system using microphone and processed with sample speech data to identify the uttered word. The input speech signal contains different representations which are generated during speech production. The language code generator step converts text symbol to phonetic symbol. Vocal tract system is the final step of speech production process that physically creates necessary sound source.

Speech acquisition step is followed by speech recognition step which mainly involves speech analysis, feature extraction, modelling. This study provides step-by-step process for speech-to-text conversion with significant accuracy.

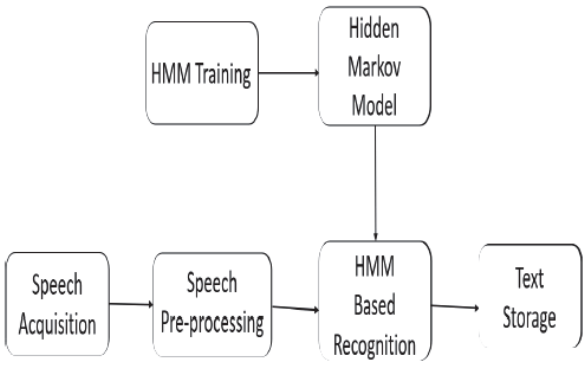

Fig. 9. System Architecture

Subtitles provide a gateway for data communication without any kind of barrier. It requires moderate quantity of hardware for data acquisition for example microphone, video recorder etc. In this system we are going to develop instant bi-lingual subtitle generation mechanism for video system. The proposed system design is as shown in fig.10

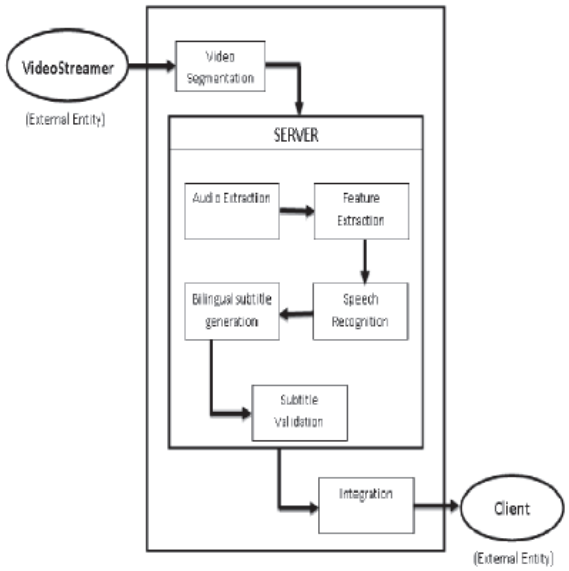

Fig. 10. System Design

1. Mel-Frequency Cepstral Coefficient :

Feature extraction plays important role in speech recognition system. Mel-frequency cepstral coefficient algorithm provides robust and efficient way for feature extraction. MFCC involves pre-emphasis, Framing, Windowing, Discrete Fourier Transform, Mel frequency filtering, logarithmic function and Discrete Cosine Transform. Fig. 11 shows working of MFCC.

Pre-emphasis : In this step , high frequency component of speech signal are flatten by using first order high pass FIR filter.

Framing : In this step constantly changing audio signal is converted into frames of 20-40ms frame time length. Loss of information can be avoided by overlapping.

Windowing : In order to prevent any kind of discontinuities in speech signal produced by framing, windowing is used. In case of speech recognition hamming window is mostly preferred.

DFT : In case of speech recognition system Discrete Fourier transform is used as fast Fourier transform for conversion of frames from time domain to frequency domain. Frequency domain provides more accurate calculation as compare to time domain.

Mel frequency filtering : Normally human ear cannot understand the frequency content linearly, therefore subjective pitch is measured corresponding to each tone on Mel scale. Mel scale follows linear frequency spacing for the frequency below 1000Hz and for the frequency above 1000Hz it follows Logarithmic spacing. From the give frequency the Mel frequency can be calculated as.

$$mel(f) = 2595 \times \log_{10}(1 + f \div 700) \quad (1)$$

DCT- In this step Mel-Filtered spectrum is converted back to the time domain as in the recognition stage Mel-frequency cepstral coefficients are used as time index.

2. Hidden markov model :

Hidden markov model can be specified by the parameter (A,p,B) where A refers to the probability of state transition, p refers to initial state probability and B is emission probability density function which is shown in fig.12

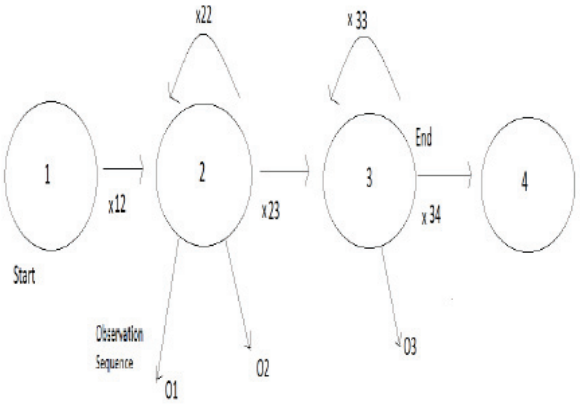

Fig. 12. Typical left-right HMM

Every model is used to calculate the probability of input sequence (O=O1,…,OT) so that probability of input sequence can be maximized by finding the corresponding state sequence. Fig. 12 shows the left-right HMM Machine Translation:

3. Machine Translation:

For translation of subtitles the microsoft's translator API has been used. Translator server is feeded with the subtitle generated in previous step in original language. The resultant text generated by server is directly added to subtitle to generate bilingual subtitles.

IV. CONCLUSION

In this paper, we have discussed about various techniques proposed by different authors to recognize speech signal and generate subtitles. We saw how each methods works, how much accuracy these methods give and concluded that Subtitle generation framework involving PNCC and HMM is more robust to noise and can generate subtitles with high accuracy. We have also observed that the conventional speech recognition mechanism involving MFCC and HMM performs extremely well in low varying environment with low computational cost as compared to PNCC. By studying advantages and disadvantages of algorithms, we have proposed a system using MFCC and HMM to generate instant bi-lingual subtitles for video.

REFERENCE

[1] Su Myat Mon, Hla Myo Tun ,"Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 06, JUNE 2015.

[2] Aleš Pražák, J.V. Psutka, Jan Hoidekr, Jakub Kanis, Lud˘ek Müller, Josef Psutka, "Automatic Online Subtitling of Czech Parliament Meetings", 2006.

[3] Priyanka P. Patil, Sanjay A. Pardeshi, "Marathi Connected Word Speech Recognition System", IEEE First International Conference on Networks & Soft Computing, pp 314-318, Aug. 2014.

[4] Yogita H. Ghadage, Sushama D. Shelke," Speech to Text Conversion for Multilingual Languages", International Conference on Communication and Signal Processing, April 6-8, 2016, IEEE, India.

[5] Ibrahim Patel1, Dr. Y. Srinivas Rao2, "SPEECH RECOGNITION USING HMM WITH MFCC- AN ANALYSIS USING FREQUENCY SPECRAL DECOMPOSION TECHNIQUE", Signal & Image Processing : An International Journal(SIPIJ) Vol.1, No.2, December 2010.

[6] Chanwoo Kim, Richard M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 24, NO. 7, JULY 2016 .

[7] Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik, Supriya Agrawal, "Speech to text and text to speech recognition systems-Areview", IOSR Journal of Computer Engineering, Volume 20, Issue 2, Ver. I (Mar.- Apr. 2018).

[8] Xiaoyin Che, Haojin Yang, Christoph Meinel, Sheng Luo, "Automatic Lecture Subtitle Generation and How It Help", IEEE 17th International Conference on Advanced Learning Technologies, 2017.

[9] Chadawan Ittichaichareon, Siwat Suksri, Thaweesak Yingthawornsuk, "Speech Recognition using MFCC", International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand).

[10] Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq "Template-Based Continuous Speech Recognition" IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 4, MAY 2007.

[11] Mercedes de Castro, Diego Carrero, Luis Puente, Belen Ruiz. Universidad Carlos III de Madrid, "Real-time Subtitle Synchronization in Live Television Programs", May 2011.

[12] Andrew Lambourne*, Jill Hewitt, Caroline Lyon, Sandra Warren, "SPEECH-BASED REAL-TIME SUBTITLING SERVICES", 2004.

[13] LUUK VAN WAES, MARIËLLE LEIJTEN, ALINE REMAEL , "LIVE SUBTITLING WITH SPEECH RECOGNITION. CAUSES AND CONSEQUENCES OF TEXT REDUCTION", Across Languages and Cultures 14 (1), pp. 15–46 (2013).

[14] Prachi Khilari, Bhope V. P. "A REVIEW ON SPEECH TO TEXT CONVERSION METHODS", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 7, July 2015.