Automatic Bilingual Caption Generation

[1]Hemant Kumar Kushwaha, [2]Mahesh Dhumal, [3]Vaibhav Gupta, [4]Sona R Pawara
[1,2,3,4]Dept. of Computer Engineering, STES's Sinhgad Academy of Engineering, Pune, India

Abstract
In the past few years, the use of video sharing platform either for knowledge sharing or entertainment has in-creased exponentially. But the language barrier has made this growth useless for a major portion of the population. Also, the hearing impaired people can't understand the content of the video. In such cases, Use of subtitle can be an effective way to resolve these issues. Currently, various video sharing platform are using either manual subtitle generation or automatic subtitle generation in the English language. This traditional approach increases the cost and also the time utilization for subtitle production. Therefore we have developed the system which instantly generates bilingual subtitles in various Indian regional languages like Marathi, Gujarati, etc. The proposed system consists of Audio extraction, Speech recognition, subtitle generation and translation subsystems which are based on Mel Frequency Cepstral Coefficient and Hidden Markov Model. In the end, implementation of the proposed system can significantly reduce the cost and time utilization for subtitle generation without any reduction in quality of subtitles.

Keywords
Mel Frequency Cepstral Coefficients (MFCC), Hidden Markov Model (HMM), Speech-To-Text (STT), Large Vocabulary Continuous Speech Recognition (LVCSR), Support Vector Machine (SVM), Continuous and Discrete Hidden Markov Model (CDHMM)

I. Introduction

With the exponential growth of the Internet in past years, right from 16 million users in 1995 which was 0.4% of the total population to an extent of 4,156 million users in 2017 which is 54.4% of the population, video sharing platform has also observed tremendous growth. Some videos sharing platforms are restricted to few users either due to the language barrier or users hearing impairment. Subtitles are the solution for these users to understand the content of the video.

Speech recognition is the ability of a machine or program to identify the corresponding text from the input speech signal and convert it into a machine-readable format. Speech recognition works using acoustic and language modeling algorithms. Also, it can be classified further into two categories: speaker independent and speaker dependent. Speech recognition mechanism acts as an interface between the user and a computer system for human-computer interaction.

Nowadays, many video sharing industries are using manual subtitle generation systems, but these systems result in a huge waste of time as well as resources. However,
some industries are also adopting the automatic subtitle generation systems, but most of these speech recognition technologies are designed for the English language, making it restricted to use only by urban communities and educationally privileged people. The native rural communities and underprivileged people are mostly able to understand their native language, like Marathi is a native language of Maharashtra, and are kept away from these technologies. In our day to day life, most of the time we mix the English language with native language. The system is designed for bilingual subtitle generation for various native languages.

The objective of the proposed system is to design and implement an STT conversion system for English to Marathi, Hindi, Gujarati and Bengali languages. The system gives English subtitle along with user preferred native language for the given input video. The system works around a dataset of 1000 English sentences with their respective native language equivalents. This work is based on MFCC and HMM. The outline of the paper is as follows. Section II gives a brief idea of implementation of current systems and their corresponding algorithms. Section III gives the overview of the system along with the various steps and component involved. Section IV describes the audio extraction and feature extraction mechanism used in system for speech recognition. It also explains the classification step done by HMM. Section V is all about the experimental setup based on MFCC and HMM along with result analysis. Section VI concludes the paper. Section VII says about future work.

II. Related Work

In [5], various speech to text models along with the different techniques are compared. It also discusses various issues related to various techniques like Linear Predictive Coding (LPC), Template Based, Artificial Neural Network, Statistical Machine Translation (SMT), etc and also what is the result generated after the execution of each technique. In [1], speech to text system is implemented in four steps using MFCC and HMM. Here the dataset consist of a recording of five audio files and each audio files contains sentence pronunciation in ten different ways. Here the percent accuracy is better and improved when the system is implemented for live audio files. In [2], LVCSR along with HMM, lexical trees and bigram language model is used for generating a system which generates online subtitles for a Czech Parliament meeting. Here 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 percent accuracy depends upon the varying topic and it is varying from around 80% to 95%.

In [3], the implemented system focuses on speech recognition for Marathi to English mix speech. Here the main objective was to generated a system for English- Marathi mix speech using MFCC, SVM, and Minimum Distance classifier. The percent accuracy achieved for the proposed system is higher as compared to the one using MFCC feature extraction technique and CDHMM classifier. In [4], a system was proposed to improve the representation of speech features in HMM-based system by recognizing the speech signal using data of frequency spectral from Mel Frequency. This system fails when the condition is varying accordingly. In order to overcome this failure condition. A sub-band decomposition technique known as frequency isolation was used. The system generated was highly robust to noise. In [6], a framework was proposed to generate automatic bilingual subtitles using Sentence Boundary Detection (SBD), Automatic Speech recognition(ASR) and Machine Translation (MT). Here IBM Watson speech to text

framework was used for speech recognition. This framework generates the transcript file in 1.5 times the duration of audio files and uses Microsoft Translator API as a machine translation tool. The proposed framework can reduce the production time by 1/3, with no decline in quality.

III. System Overview

The system operation is divided into 3 phases:

1. Recording and Audio Extraction
2. Feature Extraction and Recognition
3. Speech Translation and Subtitle Integration

In first phase video is recorded and then audio is extraction from the video for further processing.

In next phase the acoustic features are extracted using MFCC method. These features and MFCC coefficients are compared with the features in training set. The sentence with the minimum difference between values from the data in dataset is given as the recognized sentence.

In the last phase, the recognized English sentence is translated into the respective preferred language. And the recognized and translated sentences are incorporated with the original video.

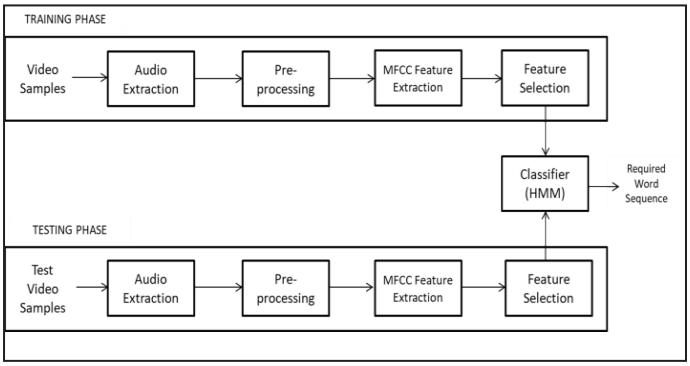

Fig. 1: System Block Diagram

A. Dataset

A video dataset is a dataset consisting of video files which contain utterances of ten different sentences by 10 different users. A video dataset is a very important and crucial point in any subtitle generation system. The dataset is categorized into 2 types:

1. Training dataset
2. Testing dataset

Table 1 : Dataset Table

S.No	Original Language	Target Language
1	Black is my favouritecolour	काला मेरा पसंदीदा रंग है
2	I am fine	मैं ठीक हूँ
3	How are you?	तू कैसा है?
4	India is my country	भारत मेरा देश है
5	We are sorry	हमें खेद है
6	He is not well	वह ठीक नहीं है
7	Rakesh Sharma was the first Indian astronaut	राकेश शर्मा पहले भारतीय अंतरिक्षयात्री थे
8	Ram is playing football	राम फुटबॉल खेल रहा है
9	Where are the students?	विद्यार्थी कहाँ हैं?
10	Who is that person?	वह व्यक्ति कौन है?

A. Training dataset

Training dataset contains recorded speech utterances of 10 different users for 10 English sentences. Each sentence being uttered 10 times by each user i.e. 80 utterances of each sentence are used to train the model and a total of 800 samples are used to train the model.

B. Testing dataset

Testing dataset contains approximately 20% of the recorded speech utterances from the original dataset. The speech utterances in this dataset are separated from the training dataset.

IV. Bi-Lingual Captions Generation

A. Audio Extraction

ffmpeg library is used for audio extraction from recorded video sample.It is a very fast command line converter that can convert different multimedia files between various formats. ffmpeg does read from arbitrary number of input files, these files can be regular files, network streams, etc. The file type is specified by the "-i" option. Also, it writes to a number of output files, which are specified by a plain output url.

"-map" option is used to select which stream from which input will go into which output. It can also be done automatically.

"-y" is a global attribute which overwrites output files without asking.

"-ac" and "-ar" are the two stream specifiers which are used for input/output per stream. -ac is used to set the number of audio channels and -ar is used to set the audio sampling frequency. Both -ar and -ac are set by default to the corresponding value from the input stream and input audio channel.

"-vn" attribute can be used as both input and output option. If used as an input option, it blocks all video streams of a file from being filtered or being automatically selected. And, if selected as an output option, it disables video recording.

"-vf" attribute takes a parameter as an input and create the filtergraph specified by the parameter and uses it to filter the stream. It is an alias "-filter:v".

The extracted audio sample is then passed to the feature extraction subsystem.

B. Feature Extraction

The very first step in any speech recognition system is of feature extraction. Its function is to identify the components of the audio signal that can be used for the identification of the linguistic content.

MFCC is a widely used feature extraction technique in speech recognition systems. The main purpose of feature extraction is to compress the input signal into features and use these features. The features are insensitive to speech variations, changes in environmental conditions and are independent of the speaker.

Steps used in MFCC feature extraction are:

A. Pre-Emphasis
Pre-emphasis is basically applied to flatten the input speech signal. It also helps in improving the Signal-to-Noise Ratio.

B. Framing
The frequency of audio signal changes over time. Applying Fourier Transform across the entire signal can result in loss of frequency contours over a short period of time. To avoid this, we assume that frequencies are stationary for a short period of time. So, we frame the entire signal into 20-40ms frames. Hamming window gets rid of some data from the start and end of the frame, so to re-incorporate this data overlapping is applied on frames.

C. Windowing
Windowing is done in order to avoid or reduce the spectral leakage and distortion in the spectrum, introduced by the framing process. Hamming window is used for this purpose.

D. DFT
FFT is an efficient algorithm for DFT implementation. DFT can be used for frequency spectrum estimation, which is also called as Short-Time Fourier-Transform (STFT).

E. Filter Bank / Mel Frequency Filtering
Triangular filters are applied on a Mel-scale to extract frequency bands. The Mel-scale aims to mimic the perception of the non-linear human auditory system by being less discriminative at lower frequencies and more discriminative at higher frequencies. Given frequency can be converted to mel frequency as follows:

$$mel(f) = 2595 \times \log10(1 + f \div 700) \quad (1)$$

F. DCT
DCT is applied to perform the transformation of the Mel coefficients back to the time domain. This results in the generation of MFCC. Typically, the resulting 13 cepstral coefficients are used.

C. Classification
HMM is a statistical model and can be represented as the simplest dynamic Bayesian network. In HMM the system which is being modeled is assumed to be a Markov process with hidden/ unobserved states.

In simple Markov models, the state is visible to the observer whereas in HMM, the state is not directly visible, but the output is visible. Here the output is dependent on the state. In HMM, an inference model is being constructed based on the assumptions of a Markov process, hence the name Hidden Markov Model.

In Markov process "future is independent of the past given the present".

Let us consider a scenario where the Body type is the hidden variable and can be Skeletal, Fit or Overweight and the observed variables denote the calories intake of a person. The arrows represent the transitions from a hidden state to another hidden state or to an observed state.

According to the Markov model assumption, each state depends only on the previous state and not on any other prior state.

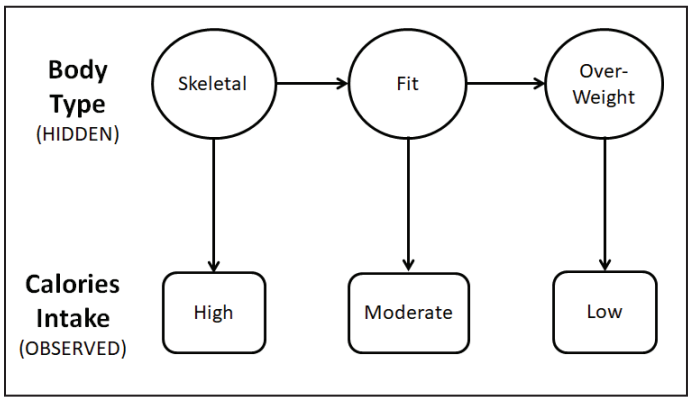

Fig. 2: HMM State Transitions

The joint probability for a sequence of states using the conditional probability chain rule and Markov assumptions can be given as:

The probability of observing a sequence of length L is given by where the sum runs over all possible hidden-node sequences

Table 2: Priors

Skeletal	0.3
Fit	0.5
Overweight	0.2

Table 3: Transition

	Skeletal	Fit	Overweight
Skeletal	0.3	0.4	0.3
Fit	0.2	0.5	0.5
Overweight	0.5	0.1	0.2

Table 4: Observations

	Skeletal	Fit	Overweight
High	0.8	0.19	0.01
Moderate	0.2	0.7	0.1
Low	0.05	0.25	0.7

HMMs are probabilistic models. Once given a set of observed states, joint probability of a set of hidden states can be computed using HMM. Joint probability of a sequence of hidden states helps in determining the best possible sequence.

Applications of HMM:
- Speech Recognition
- Machine Translation
- Handwriting Recognition
- And many more.

V. Experimental Setup
The proposed system can be considered as an incorporation of Audio extraction, Feature extraction, Classification and Subtitle Integration. Input to the system is video data recorded by user,

which is passed to Audio extraction sub-system. Output of Audio extraction sub-system is given as input to Feature extraction sub-system. From the features and MFCC coefficients given by feature extraction mechanism, first 13 MFCC coefficients along with features such as pitch, amplitude, sample rate, etc. are used in speech recognition. During training phase of system, these coefficients are used to train HMM model. For individual sentence in dataset, one model of HMM is trained. Proposed system consist of 10 trained models where each model is trained using 80 samples of audio data. These trained models are saved using pickle module of python which are then used for testing phase.

During testing phase, the output of feature extraction subsystem is given to each of 10 trained models. Each model gives likelihood for each sentence given to model.

The model having maximum likelihood is considered as required output. The sentence corresponding to maximum likelihood is given further to translation and integration Subsystem.

Translation mechanism of system takes the sentence generated from the classification step, which is then broken down to separate words based on spaces between words. Each word is given certain numeric value based on the functioning of word in system. For example, verb having maximum priority or numeric value. After labelling the words, these words are sorted in ascending order according to label. Individual word is then translated into target language and combined together to generate required sentence. This sentence is then integrated with video data usingffmpeg to generate required output.

A. Result

The percent accuracy achieved for the proposed system using MFCC and HMM techniques for native languages is 86%. The video file consist of audio data which is a clean speech signal comprising a voice uttering of single sentence without any background noise. For simplicity, we have considered a video file of 5 seconds which has audio signal corresponding roughly to a sentence in the .wav file.

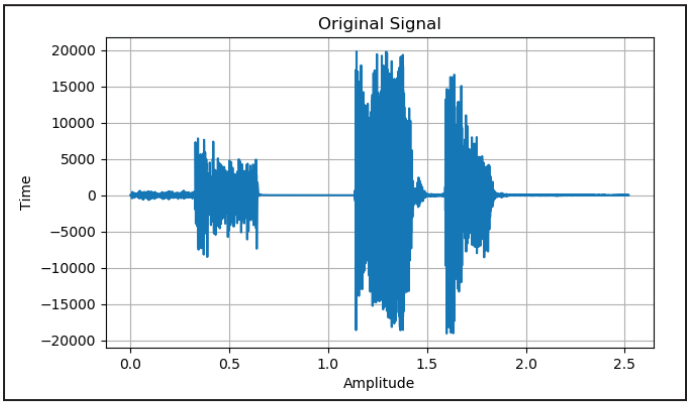

Fig. 3: Original Audio Signal

The raw audio data extracted by the audio extraction mechanism from video is in unnormalized format in time domain which can be graphically represented as in fig. 3.

In order to normalize the raw audio data, pre-emphasis is applied which balances the frequency spectrum by amplifying higher frequencies. Fig. 4 represents the normalized audio data in time domain

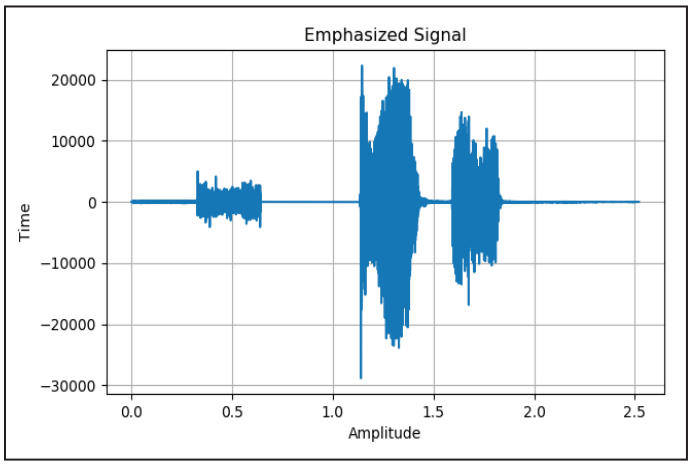

Fig. 4 : Emphasized Audio Signal

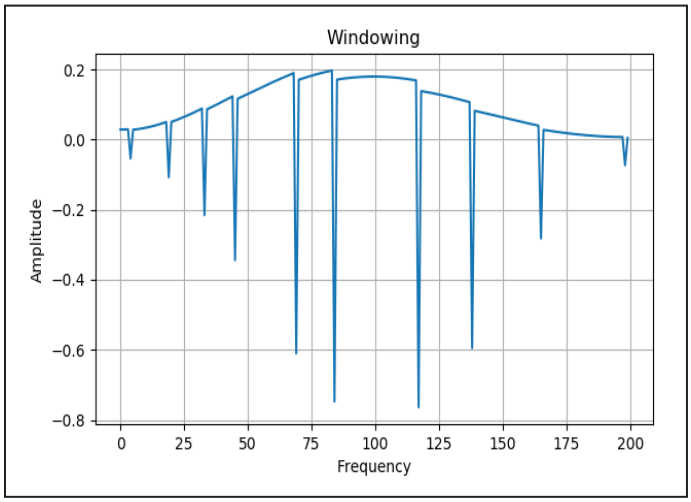

Fig. 5: Windowing

To avoid or reduce the discontinuities in audio segment Hamming window is applied on the audio data which transforms the data as shown in fig. 5.

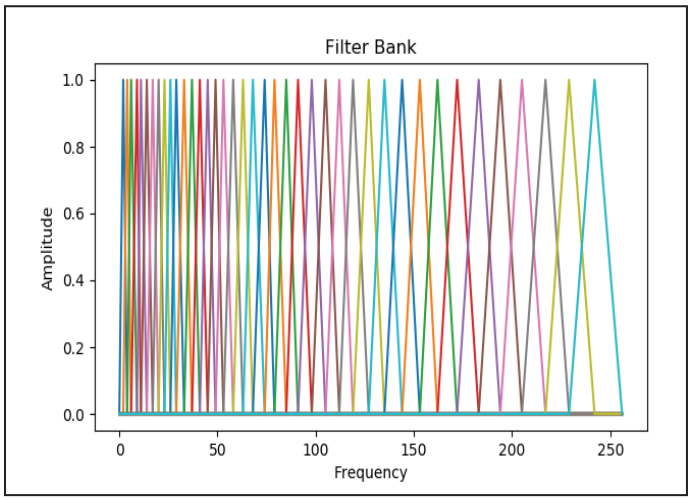

Fig. 6: Filter Bank

Here the frequency is converted from Hertz (f) scale to Mel (m) scale. Each filter in the filter bank is triangular having a response of 1 at the center frequency and decrease linearly towards 0 till it reaches the center frequencies of the two adjacent filters where the response is 0 and this is shown in the fig. 6.

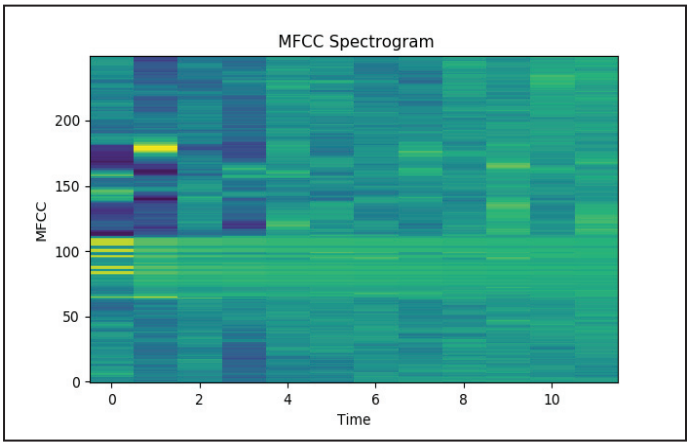

Fig. 7: MFCC Spectrogram

In order to de-correlate the filter bank coefficients generated in previous step, Discrete Cosine Transform is applied. From the coefficients generated in this step, only two to thirteen coefficients are retained which are represented in the spectrogram as shown in fig. 7.

VI. Conclusion

In this paper, we have introduced the bilingual subtitle generation system for video sharing platform which consists of feature extraction, Speech recognition, translation, subtitle integration. We have evaluated auto-generated subtitles in terms of accuracy and efficiency. Results show that the manually generated subtitles are equivalent in quality to auto-generated bilingual subtitles. Implementation of the proposed system may reduce the cost and time required for subtitle production.

VII. Future Scope

The proposed speech recognition system is developed for speaker independent English to Marathi, Hindi, Gujarati, and Bengali language subtitle generation. This work may extend for other regional languages like Urdu, Punjabi, etc. The currently proposed system works for recorded videos however the system can be implemented for live videos as well such as Parliament speech or cricket match commentary.

References

[1] SuMyat Mon, HlaMyoTun,"Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)", International Journal of Scientific amp; Technology Research, Vol. 4, Issue 06, June 2015.

[2] Ale Prazc ̧k, J.V. Psutka, Jan Hoidekr, Jakub Kanis, Ludek M ̃uller, ̈Josef Psutka,"Automatic Online Subtitling of Czech Parliament Meetings".

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

[4] Ibrahim Patel, Dr. Y. Srinivas Rao,"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.

[5] AyushiTrivedi, Navya Pant, Pinal Shah, Simran Sonik, Supriya Agrawal,"Speech to text and text to speech recognition systems-Areview", IOSR Journal of Computer Engineering, Vol. 20, Issue 2, Ver. 1, 2018.

[6] 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.

Hemant Kumar Kushwaha has completed his Bachelor of Engineering degree in Computer from Sinhgad Academy of Engineering, Pune, India in 2019. His fields of interest are Android and Web Development.

Mahesh Dhumal has completed his Bachelor of Engineering degree in Computer from Sinhgad Academy of Engineering, Pune, India in 2019. His fields of interest are Machine learning and Web Development.

Vaibhav Gupta has completed his Bachelor of Engineering degree in Computer from Sinhgad Academy of Engineering, Pune, India in 2019. His fields of interest are Data Science and Web Development.

Sona R Pawara is a professor at Sinhgad Academy of Engineering, Pune, India in 2019. Her field of interest is Machine Learning.