International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

Automatic Subtitle Generation In Videos Using Mel Frequency Cepstral Coefficients

[1]Aseem Mahajan [2] Lakshmi Sureshbabu [3] Jessal Manidhar R [4] Kumar Bhagwat [5] Poonam P.Bari
[1][2][3][4]B.E., Student [5] Assistant Professor,
[1][2][3][4][5] Department of Information Technology
Fr. C. Rodrigues Institute of Technology, Vashi
[1] aseem.mahaj@gmail.com [2] lamysureshbabu@gmail.com [3] jessalmanidhar@gmail.com
[4]kumarb@gmail.com [5] poonambari26@gmail.com

***Abstract*:** **Video is currently one of the most popular multimedia over pcs and the internet. It is very difficult to comprehend the meaning of videos without proper subtitles for deaf and hearing impaired people. The proposed system aims to create a media player that generates subtitles automatically using speech recognition. The input is in the form of a video file. The video file is subjected to audio extraction to create an audio file. Speech recognition using Mel Frequency Cepstral Coefficients (MFCC) is then performed on the extracted audio file. Java speech Application Programming Interface (API) is used for speech processing with the help of the grammar design .The subtitle file generated is then synchronized with the input video file. The subtitle generation is performed offline and for English videos only.**

Index Terms: **Audio extraction, Java Speech API, MFCC, Subtitles.**

I. INTRODUCTION

Video plays an important role in various sectors such as business, education, entertainment and other private sectors. Due to globalization and the advent of social media, people with varied cultural and linguistic backgrounds interact, exchange and view ideas through videos across the world. The language barrier that is faced by most in videos affects the understandability of the content in videos adversely. Subtitles are usually displayed at the bottom of the screen or at the top if some other text is present at the bottom already. They can either be a written translation of the audio in videos in a foreign language or written form of the audio in the same language. Subtitles help viewers who are deaf or people who cannot understand the spoken language in the video or who find it difficult to recognize the accent. The understandability of videos is greatly improved by the presence of subtitles. However, the availability of subtitles file for videos is low and embedding subtitle files to every video is not feasible. To address this problem, the proposed system aims to create a media player that performs automatic subtitle generation of audio and video files using Mel Frequency Cepstral Coefficients (MFCC).

1.1 Mel Frequency Cepstral Coefficients (MFCC)

In any automatic speech recognition system, it is necessary to identify the components of the audio signal that contains the linguistic content and discard all the other parts of the signal which carries information like background noise, sounds other than speech, etc. Mel Frequency Cepstral Coefficient (MFCC) technique is often used to create the speech-skeleton of the sound files. The MFCC are based on the known variation of the human ear’s critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech [1][2].

A brief overview of steps in MFCC used in speech recognition is:

1. Frame the signal into short frames. The audio signal is divided into 20-40 ms frames. Due to constant change in audio signal, it is divided into short frames, assuming audio signal does not change over short frames.
2. Calculate the period gram estimate of each frame of the power spectrum. This is motivated by an organ in the ear, cochlea. Depending on the incoming sound frequency, the cochlea vibrates at different spots. The frequencies that are present in the frame are identified by the period gram estimate.
3. Apply the Mel filter bank to the power spectra, sum the energy in each filter. The period gram spectral estimate contains a lot of information not required for Automatic Speech Recognition (ASR) [3][4]. The cochlea cannot discern the difference between two closely spaced frequencies. This effect becomes more pronounced as the frequencies increase [5]. Due to this, we take clumps of period gram bins and sum them up to calculate the energy that exists in various

frequency regions. This is performed by the Mel filter bank.

4. Take the logarithm of all filter bank energies. This is also motivated by human hearing principle that we do not hear loudness on a linear scale. This compression operation makes features calculated match more closely what humans actually hear.
5. Take the discrete cosine transform (DCT) of the log filter bank energies [5].The filter bank energies are quite correlated with each other due to which the filter banks overlap each other. The DCT de-correlates the energies.
6. Keep DCT coefficients 2-13, discard the rest. The higher DCT coefficients represent fast changes in the filter bank energies and it turns out that these fast changes actually degrade ASR performance, so we get a small improvement by dropping them [5].

1.2 Java Speech Application Programming Interface (JSAPI)

It allows developers to integrate speech technology into user interfaces for their Java applications and applets. This API stipulates a cross-platform interface to support command and control recognizers, speech synthesizers and dictation systems. Speech recognition provides computers with the ability to listen to spoken language and determine what has been said [6]. It processes audio input containing speech by converting it to text.

A brief overview of a typical speech recognizer is:

1. Grammar design: Defines the words that the user may speak along with the patterns in which the user may speak.
2. Signal processing: The spectrum (i.e., the frequency) characteristics of the input audio is analyzed.
3. Phoneme recognition: The phoneme patterns of the language that is subjected to recognition is compared to the spectrum patterns is done.
4. Word recognition: Comparison of likely phoneme sequences against the words and patterns of words specified by the active grammars is done.
5. Result generation: The application is provided with the information of the words that are recognized and detected in the input audio by the recognizer.

A grammar is an object in the Java Speech API that indicates what words a user is expected to say and in what patterns those words may occur [6]. Grammars constrain the recognition process due to which it is important for speech recognizers and it enables faster and more accurate recognition.

II. THE PROPOSED SYSTEM

As a result of globalization and widespread reach of social media, multimedia in the form of videos reach everywhere and are viewed by everybody. But in videos, the understandability is a major issue. To address this issue, subtitles are introduced into videos.Various attempts regarding subtitle generation have been done but with major issues regarding format compatibility and accuracy of subtitles. Thus, a new system is proposed in which a media player accepts video inputs and performs speech recognition and speech processing on the audio file extracted and generates subtitles for the video and plays the file.

In the proposed system, the input is a video file and provides subtitles in the output. The system accepts multiple video formats and performs subtitle generation only for English videos. The proposed approach performs audio extraction on the input video file and produces an audio (mp3 format) file. This audio file is input to the speech recognition phase in the form of audio signals. Speech recognition is performed using Mel Frequency Cepstral Coefficients (MFCC). In this process, the audio signal is filtered to give the audio signals consisting content only of linguistic nature. The audio signal is then subjected to speech processing using java speech application programming interface. The speech is converted into the text for subtitles as per the grammar design predefined. The subtitles generated is finally synchronized with the input video file. The player completely automates the process of subtitle generation and the entire process is performed offline.

A. Scope of our System

The media player is developed on the basis of requirements of two major types of users, that is, hearing impaired users and users who are not familiar with the language or its accent used in the video.

1) The subtitle generation is performed for English videos only.
2) The accuracy of subtitles is of utmost importance while generation- Our system aims to generate subtitles with maximized accuracy by the use of Mel Frequency Cepstral Coefficients.
3) The player supports multiple video formats- Every video format is converted to a common audio format in the first step.
4) Speech recognition is performed using Mel Frequency Cepstral Coefficients- in which the background noise is minimized.
5) Speech recognition is performed using java speech Application Programming Interface- in which grammar design is specified.

B. *Future scope:*

1) Making the subtitle generation for videos real-time
2) Multilingual subtitle generation support

III. METHODOLOGY

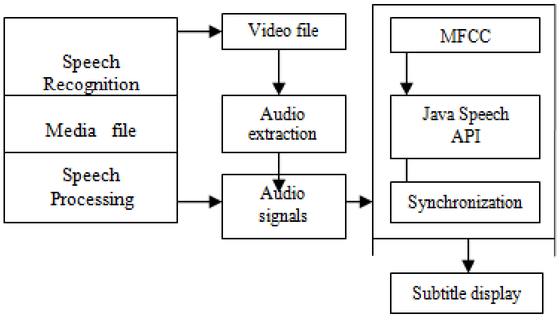

Fig 3.1: General Architecture of the proposed system

Fig 3.1 shows the overall block architecture and stages involved in automatic subtitle generation. A media file is input into the system. The media file can be an audio file or a video file. If the input file is a video file, audio extraction is performed. The audio signals that are extracted are subjected to Mel Frequency Cepstral Coefficients for speech recognition and removal of noise elements from the signal. Speech processing is performed using java speech application programming interface to identify the speech present in the signals. The generated subtitles are then synchronized with the video file and displayed.

A. *Audio Extraction*

The audio content of the video file is extracted.The extracted audio content is then converted to a common format. The input media file can be of two types:

A. ***Video file:***

Audio extraction process is required only if the input media file is in a video format. Multiple video formats are accepted as input files to the system. The audio file generated post extraction process is in mp3 file format. The audio file is then converted to its respective audio signal format.

B. ***Audio file:***

The input audio file format is converted to mp3 format followed by its audio signal generation.

B. *Speech Recognition Using Mel Frequency Cepstral Coefficients*

The speech recognition is performed using Mel Frequency Cepstral Coefficients (MFCC). The audio signal is subjected to MFCC to remove all the background noise and hence, obtain purely the speech content of the audio signalsdiscarding all the other information like background noise,emotion,etc. The input audio signals are segmented into short frames followed by the calculation of the periodogram estimate of each frame of the power spectrum. This calculation enables identification of various frequencies present in the signal. Mel filterbank is applied to the power spectra and hence, the sum of energies in each filer is obtained. The logarithm of all filterbank energies is taken since human hearing and loudness is on a logarithmic scale. Finally, discrete cosine transform (DCT) of the log filterbank energies is taken [5]. DCT coefficients of range 2-13 is used and the rest is discardedas the higher DCT coefficients represent fast changes in the filterbank energies which degrade the automatic speech recognition performance. The output signal contains the speech content of the original signal only [5][7].

C. *Speech Processing*

After the process of speech recognition, the output signals are subjected to processing using Java Speech Application programming interface (JSAPI). In this process, grammar designs are specified and defined as per the various accents of English language. This will include the words and its patterns used in the language. Grammars are important to speech recognizers because they constrain the recognition process [6]. These constraints make recognition faster and more accurate. The spectrum (frequency) characteristics of the speech signal is analyzed for the incoming audio. The spectrum patterns are compared to the patterns of the phonemes of the language beingrecognized. Word recognition is performed by comparing the sequence of likely phonemes against the words and its patterns specified by the active grammars. The subtitles are created for each corresponding speech signal by providing information about the words that have been detected in the incoming audio.

D. *Synchronization*

Once the corresponding subtitles are generated for each speech signal, synchronization of subtitles with the video is crucial for attaining accuracy. In order to achieve this, the text generated must be synchronized with the audio in time domain. Timestamps are provided to the words as per the video timing. Each word is assigned initial and final timestamps corresponding to the time of the video. The timestamps of the words generated are matched with the video time. The text is embedded into video ready for display with subtitles appearing at bottom of the screen.

IV. CONCLUSION

Various existing systems are studied along with new technologies to get deeper insight of the proposed system to be developed. Many technical papers were analyzed to get clear view of documenting and developing the system.

The proposed system of the project attempts to bridge the understandability gap created amongst users by videos due to unfamiliar languages or its accents by automating subtitle generation of videos with minimum user involvement. The system also enables the hearing impaired users to view and understand videos to the fullest extent. However, the major focus of the proposed system is to create a media player dedicated for the generation of subtitles automatically for English videos.

REFERENCES

[1]Lindasalwa Muda, Mumtaj Begam and Elamvazuthi., "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and DTW Techniques", Journal of Computing, Volume 2, Issue 3, March 2010.

[2]Mahdi Shaneh and Azizollah Taheri,"Voice Command Recognition System based on MFCC and VQ Algorithms”, World Academy of Science, Engineering and Technology Journal, 2009.

[3]R. P. Lippmann, “Speech recognition by machines and humans,” SpeechCommun., vol. 22, no. 1, pp. 1–15, 1997.

[4]Gerasimos Potamianos, Member, IEEE et. al. "Recent Advances in the Automatic Recognition of Audio-Visual Speech"

[5]http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[6]Wikipedia:http://en.m.wikipedia.org/wiki/Java_Speech_API

[7]Mudit Ratana Bhalla,"Performance Improvement of Speaker Recognition System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012