

AUTOMATIC SYNCHRONIZATION OF AUDIO-VISUAL CONTENTS WITH SUBTITLES

Santosh Hiremath* and Jenila Livingston LM

Received: September 30, 2020; Revised: June 30, 2021; Accepted: July 02, 2021

Abstract

This paper presents an automated Kannada subtitle generator from Kannada video which is implemented to assist people with auditory problems for watching videos. Subtitle generation has become an important task for supporting such special people and it integrates an audio extraction and a speech recognition module and synchronization. The proposed work involves three phase such as extraction of audio form video, recognition of speech, generating subtitle and synchronization of subtitle with audio and video. An adaptive speech recognition module is implemented AMFCC for feature extraction. For decoder acoustic module, AHMM is used to reduce the computational time and memory usage. The text file from the speech recognition module is rendered to synchronize the missing offset with the video using parallel processing. One amongst the foremost essential factors with audience perception is the quality of the video subtitle for those people who are deaf and difficult for hearing. In this paper we are discussing the technique for Sub-Synchronization tool for automatic synchronization between media substance and subtitles, by taking an advantage of current familiar techniques which are utilized for word sequences alignment. The main objective of this paper is to deal the lack of synchronizing that happens within the subtitles once made during the video that have some improvised parts. In order to synchronize the audio-Video content with subtitle an efficient technique is used for sequence alignment is PLCS algorithm which gives best performance in terms of time and space, accuracy compare to other existing technique for sequence alignment algorithm.

Keywords: DTW, LCS, PLCS, CGF, Needleman-Wunch, MFCC, HMM

Introduction

According to the study, the major auditory impairment affects 360 million people (5%) in India. Deafness means hearing deficiency, which can be partial or total. It is difficult to see hearing impairment and therefore the symptoms are not able to seen anyone, therefore hearing-impaired suffer in silence. The extreme hearing impairment for those people whose audible range loss is more than 70dB. Although a people with audible suffering from rigorous hearing harm as per the World Health Organization (WHO) is automatically excluded

from the hearing handicap category Stan *et al.* (2016). The WHO definition of “deafness” refers to the loss of hearing ability in one or two ears. The WHO definition of “hearing impairment” has been refers to type of people who having complete and partial ability to hear Vanderplank and Robert (2016). Speaking is a foremost activity of all human beings to communicate with one another. An important issue in today’s world is speech estimation and speech separation that is often no stationary and harmonic. In actual, separating the speech form

¹ Researcher Scholar, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India, 600127.
E-mail: santoshswami410@gmail.com; jenila.lm@vit.ac.in

* Corresponding author

music became one of the major challenge Loni and Shaila (2015).

For speech separation, the objective is to isolate the target speech from various types of background noise which can be broadband or narrowband, periodic or aperiodic. Also, the background noise is independent of speech in most cases so that their spectral contents are uncorrelated Biswas *et al.* (2015). It has been applied to many aspects of daily life. But it has some limitations. For example, some important signals, such as language, music, digital version of film and television which have a huge amount of data, have a higher transmission and storage costs Szeto and Elson (2015) In today's world, the use of subtitles has become important for an understanding of the video. Subtitling is essential for those people who are deaf, who are having reading and literacy problems and can help those who are learning to read. Subtitles provide information for people who have difficulty understanding the speech and auditory components of the visual. This leads to a valid subject of research in the field of automatic subtitle generation Aparício *et al.* (2016). It is very useful for those people who are deaf, who have reading and literacy problems also to those who are learning to read.

A subtitle causes the individuals to understand the speech and hear-able parts of the visual, which prompts a legitimate subject of examination in the Automatic Subtitle Generation field. However, manually creation of subtitles is a very long and boring activity which requires an extensive presence of the user is essential Schmidhuber and Jürgen (2015). Video is the most popular multimedia used on PCs. So it is essential to make information encompassed in the form of sound in videos comprehensible to people with auditory problems. However, manual subtitle creation is a tedious activity and requires the active participation of the user Wright *et al.* (2013). The problem taken up falls under two major areas of Computer Science-Speech Recognition and Natural Language Processing. Speech Recognition is the used to translation of spoken words into text. The Automatic Speech Recognition (ASR), Computer Speech Recognition, or Speech to Text (STT) and Natural Language Processing (NLP) deals with interactions between the computer and human languages. It makes the computer to understand and interpret the human language Anagnostopoulos *et al.* (2015).

Subtitling for Videos is generally done manually which is time-consuming and tedious. Manual subtitling of a video on an average takes eight to ten times the length of the video and requires skilled people to do the job. Automation of this process saves time, money and human labour.

For a machine to do subtitling, it must be provided with the capabilities to understand speech and transcribe it into text format. It must also be able to obtain timing information so that the subtitle could be aligned to the speech. Variations that exist in speech between different speakers are a major concern in Speech Recognition. Many approaches exist in STT like speaker-dependent, speaker-independent, command and control for single speaker etc. Araújo *et al.* (2013).

People with hearing impairments are very difficult to communication in society over the world around Orero (2007). These hearing difficulties disrupt the overall communication process in order to access the information from a barrier problem they must defeat on a day by day basis. Also, these situations create a social exclusion, one of them reasons not to getsensory-disabled Equal rights. These rights are very important this situation which has been recognized by the United Nations at the international conference. December 13, 2006. Eur. Health Law (2007) Great States must comply with Article 30 Take the suitable process for the people with suffering the disabilities problem by including access to TV program, movies, theatres and other activates, in easily reached formats. Considering the report from Global Disability Disabil (2011), there are over a 1,000 million people who are suffering any one type disability. Therefore, demand increase disabled communities' educational programs which are available especially for TV program topics are subtitled, and yet, for this the subtitle is high-quality Gao *et al.* (2009). This condition lead to governments & dictatorial bodies are mandatory requirements for broadcaster to ensure that TV programs should contain a subtitle or an language with the sign. Services and Audio for the hearing-impaired interpretation services for the blind or visually impaired Ley (2010).

Literature Survey

Xuemei *et al.* (2017) demonstrates a multi-hierarchical semantic information descriptive model based on video-segmentation and extraction technology. The goal of their research mapping of video semantic information from a lower level to High-level feature when translating subtitles. Moreover, batch processes technique can be used for web services for synchronization. For example, Montagud *et al.* (2017) Launches a web-based application for synchronization of subtitles on both Single and multiscreen scenes. This web application can also allow the dynamic customize subtitle design and its position as per the customer preferences. Synchronization between text, audio is also indentified and developed for the web applications excluding subtitle. In Lert wong

khanakool et al. (2013), an algorithm for synchronizing broadcast speech with its transliteration are done by using Syllabic unit for the real time speech. The objective of this work is to create the audio dictionary in structured language for Thai. Garcia et al. (2009) demonstrated a system for Spain language in public television. The system was developed without human intervention for syncs of subtitle for news Television program. By combine the STA and TR, which can interact with the news reduction computer (NRC) in order to obtain the programme which is submitted to the screen. ASA engine based STA; The texts are received and makes a provisional arrangement for the incoming of audio signal. The work has been carried and tested on programs with online script by using an HMM and GMM for the feature vectors.

In the Hybrid Broadcast Broadband TV (HBBTV), various proposals have been attempted to work carried out with the synchronization between Audio-visual substance and subtitle. A novel application is proposed to check the synchronization of subtitle and video content in the broadcast Video via an InternetD. Kedačić et al. (2018). Moreover, media synchronization (e.g. Broadcast video with subtitles or alternate audio Via the Internet) The novel is gaining new attention with connected device ecosystems initiating media use Models Deventer et al. (2017) The novel be gaining new interest with the connected device by an eco-systems which initiating the video use the Models. Another goal is to make content available to everyone Over come existing barriers to accessing content for users Must meet specific requirements, or different platforms Makes content fully usable and accessible. Cuzco-Calle et al. (2018) Provides an appropriate system for the automatic production of Sync video content and subtitles for the people who is suffering from hearing impaired using an Natural Language Processing. Rodriguez-Alsina et al. (2013), In the video subtitle system enabled uses of customize, alteration and synchronization of subtitles across the different single and multiple screens; Allows the people to access the content anywhere. In Huang et al. (2003), ASR is used for live scripted programs For synchronizing pre-defined subtitles. ASR produces and contains a transcription of the program's based on audio Time stamp of each word by applying the Dynamic programming. Kidneys and others. Demonstrate robotic solution for practical application of ASR using context-tuned model and Machine learning in a large organization of data-namely most hours are news broadcasts with exactly the title programs. The main Aim of this work is to create the identification and clear patterns

to turn it into raw automated transcription interested in broadcast titles to introduce technology into the live production environment. Finally, with regard to ASR methods, it is usually the top speech Identification systems are based on advanced hybrid algorithms and manual process steps. Conventional speech systems use a number of processing steps, with unique input features and sound. By using Hidden Markov Models.

The link between the subtitle and Position is the context of a scientific problem: the arrangement of the word Scenes. In Wunsch and Needleman (1987) Needleman-Wunch proposed the technique to identify the best fit between the 2 scenes. Mainly it was developed for the localization of sequence alignment by using the position of similarity between arrays words. This algorithm is used as a tool for different domains such as pattern matching for conversational speech Wilkerson and Casas (2017), In some application Such as bioinformatics, in which the length of the scenes is too long, memory usage is prohibited. The efficiency can be measured in terms of time and space resources is given as $O(nm)$. Therefore, An optimizations technique which was introduced by Hirschberg (1975) which can reduce space upto $O(n+m)$, but, at the overhead of computational time. Karpukhin et al. (2017) Other proposals include Levenshtein distance technique to sync the video, audio and text transcriptions, which was implemented for the Parliament Sessions for Alignment of text with Media Contents and also for videos lecture class or TV series with video script and subtitle arrangement Zhang et al. (2015).

Existing Methodology

In most of the cases, the sound plays a very important role now a days, hence it became an essential part to make the understanding of a video available for people who are with hearing problems. By making the use subtitles for an video which can play an important role to understand the videos content. For solving this problem an automated subtitle generator is implemented. Additionally, some other problems which lead to the implementation of an automatic subtitle generator include the impossibility of the manual writings of the subtitles. Moreover, the application of video watching in some noisy environments such as airports and transliterations where it is not possible to hear TV broadcasts is another challenge. Time synchronization is also an important consideration in the process of subtitle generation. The robustness, automatic generation of the word lexicons, word accuracy rate, error rate, silence and synchronization between the signals is also be important concern which is to be considered for the

video contents. The subtitling provides information for the individuals with the difficulty of understanding the speech and also for the deaf. The manual process of creating subtitle includes huge time and hence automatic subtitle generation is required. Hence the different techniques are used for the automatic subtitle generation with synchronization of subtitle with audio video and subtitle. Figure 1. shows the different stages and phases Process developed for synchronization of subtitles in parallel with Subtitle transcription process.

The first step, called transcription, uses the Automatic Speech Recognition (ASR) engine to gets a steady stream of words in line with Audio-Video transcription. The Audio-Video transcription tool was developed and demonstrates for the both scientific literature and technology market. An commercial ASR system can be used for speech recognition, speech to text or text to speech as long as it is provided transcript word, time stamp indicating any sound corresponding to a specific text is generated. This system gives a better accuracy in word recognition and time stamp. The detector generate a current sequence words and make a comparison with the transcription which was stored in the dictionary, once the subtitle is matched; it will stop process.

In step 2, the goal is to figure out what the perfect moment to have each subtitle a process that authors call chronization. Subtitle and Transcription are the two available information resources which are used for the production of the program. The objective is to find out which terms of transcription storage correspond to the subtitle and from them and their time impressions determine their present moment. The eraser is set up according to the length of the subtitle. Depending on the situations defined above, whenever subtitle is received, the words associated with it is already in available in the transcription storage. Thus, a subtitle Memory are not being running, which can further delay the transmission. This step is a key component of the subtitle synchronization application.

Finally, in the 3rd Step, which is called as synchronization, the main objective to synchronize the audio-video content with the subtitle based on

the transcription received for the given video file. For this step, storage is required audio-visual material for a long timesubtitles should be available in the transcription memory.

The main element of the chronization process goal is to find the right (past) time each subtitle is displayed. The goals is used to develop an algorithm for synchronization between video content and subtitle are:

Need to find the best word fit alignment between the word forms the appearance of subtitles and word order the transcription is in storage.

To accept that there will be omitted extra and modified words.

To establish comparison estimation method for different setting alternatives, which can gives importance for the rearrangement the word which are longer than short words.

To studies the arrangement of similar words.

To give additional weightage to initial calibration efforts than the latter.

Considering these requirements, the various alignment algorithms have been introduced. They are as follows:

- A. Dynamic Time Warping
- B. Longest Common Subsequence Algorithm.
- C. Needleman-Wunch
- D. Context Free Grammar.

Dynamic-Time Warping

This technique is used to find a best alignment between the two given sequences of words based on the time dependent. Here sequences which are warped into a nonlinear format is used to match with the other sequence of words. This technique mainly used to compare the speech patterns with different speech patterns for the speech recognition.

The main objective of dynamic-time warping (DTW) is to align the two sequences in some optimal sense based on the certain constraints. The sequences can be discrete signals, character of sequences, sequences of futures, or any other futures based on time series.

To determine the optimal warping path for two sequences A and B, one should compute the total cost of all the possible (n, m)-warping paths and

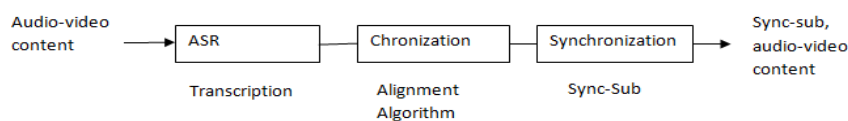


Figure 1. Phases for synchronization of subtitles.

then obtain the minimal costs. The no's of different (n, m)-warping paths is exponential between n and m. To compute optimal solution between two sequences (n,m) words can be done by making use of dynamic programming technique .For example: In the 1st rows, we will take no values from transcript and the distance between this and subtitle test will be considered as Zero. So we will assign Zero to the first row. Similarly, value for the first column. If we take no values from Subtitle text, the distance between this one and transcript will also be consider as Zero. And the distance between 1st rows and 1st columns initially set as zero. For the entire step, will be considering the distance between from each point in concern and add it's with the minimum distance can be found so future. The resulting will give the optimal distance between two sequences up to that position. In order to get the distance between the two sequence, DTW can be computed by applying the dynamic programming using the below formula,

$$\text{Str}[j][k] = d(A[j], B[k]) + \text{Minimum}(\text{Str}[j-1][k-1], \text{Str}[j][k-1], \text{Str}[j-1][k]) \quad (1)$$

Algorithm DTW (A, B):

Step 1: Find the length of two sequence i.e. n, m
 $n = \text{length}(A)$
 $m = \text{length}(B)$
 Step 2: Create Table with size of (n+1) (m+1)
 Step 3: Initialize the first row and first column as Zero.
 Step 4: Find the optimal distance between two sequences
 for j from 1 to n
 for k from 1 to m
 $\text{Str}[j][k] = d(A[j], B[k]) + \text{Minimum}(\text{Str}[j-1][k-1], \text{Str}[j][k-1], \text{Str}[j-1][k])$
 Step 5: Return the optimal result
 $\text{str}[n+1][m+1]$

Longest Common Subsequence algorithm

The Longest Common Subsequence (LCS) algorithm is used to finding the longest subsequence common to all sequences in the set of given sequences. The LCS algorithm differs from the longest common substring problem such as substrings, sub-sequences which are not required to occupy following positions within the original sequences.

The LCS algorithm has an optimal organisation, Here the problem can be divided into smaller and simpler sub problems, which can sequentially be divided into simpler sub problems, and so on, until the final solution becomes trivial. LCS has an overlapping sub-problems: the solutions to high-level sub-problems repeatedly reuse solutions to lower level sub-problems. Problems with these two properties are agreeable to dynamic programming approaches. To obtain the score matrix S of size (n+1)*(m+1) in the records such as $s[i,j]$ ($1 < i < n, 1 < j < m$) of length of the LCS for substring X1, X2, Xi and Y1, Y2, Yj by the following recursive function.

Algorithm: LCS(X, Y)

Step 1: Find the length of two sequence i.e. n, m
 $m = \text{length}(X)$
 $n = \text{length}(Y)$
 Step 3: Initialize the first row and first column as Zero.
 Step 4: Find the optimal distance between two sequences
 for j from 1 to m
 for k from 1 to n
 if $x_j = y_k$
 $\text{Str}[j, k] = \text{Str}[j-1, k-1] + 1$
 else
 if $\text{Str}[j-1, k] \geq \text{Str}[j, k-1]$
 $\text{Str}[j, k] = \text{Str}[j-1, k] + 1$
 else
 $\text{Str}[j, k] = \text{Str}[j, k-1]$

Needleman-Wunch

Needleman-Wunch (NW) is one of the most popular algorithm for dynamic programming which can be used for global sequence alignment. The objective of algorithm is the find the best fit sequence words alignment between the given two sequences. The score given for the most possible setting and the most correct setting is the Max score.

This algorithm make uses 2D arrays, where each cell contains score and pointer with different values. The pointer which can point to either left, up, or diagonally part of the matrix. Which can indicate the maximum score of the current cell. The pointers can also use to retrieve the two sequence of words that are aligned by applying the Trace back process. Whereas, the values which are assigned to each cell are called scores. To calculate the score the different score scheme are used which contains the Miss, Gap and Match of the words. With the help of scoring scheme 3 different scores are calculated out of which the max score is selected.

This algorithm contains the 3 scores such as horizontal, vertical and diagonal gap scores. The

horizontal gap score are left and right of the cell of matrix value. Whereas vertical gap scores is the sum of all upward matrix cell and diagonal gap score is top-left cell values and matched or mismatched score. Matched score is added if there are paired side scenes in the present indexed matched; otherwise, mismatched score Added. This calculation process is repeated until all the cells are filled. Once the process is completed, the value present at the right most bottom of the cell is consider as the maximum score.

This algorithm contains the three different stages such as initializing the matrix, filling the values in the cell, and trace back approach. There are no counts in the trace back phase.

Algorithm: Needleman-Wunch(X, Y)

Step 1: Find the length of two sequence i.e. n, m
 $m = \text{length}(X)$
 $n = \text{length}(Y)$
 $d = \text{MismatchScore}$

Step 2: Initialize the first row and first column as mismatch score by incrementing value starting from -1 to So on and at the index (0,0) with Zero.

Step 3: Find the optimal distance between two sequences

```

for j form 0 to m
  Str [i, 0]=d*j
for k from 0 to n
  Str [k, 0]=d*k
for j form 0 to m
  for k from 0 to n
    Str[i, j]=max(Str[j-1, k-1]+Str[Xj, Yk], // Match
                  Str[j-1, k] + d, // Delete
                  Str[j, k-1] + d ) //Insert

```

Context Free Grammar

Context free grammar is the most popular algorithm which is used to generate the grammar and also used to check the substring matching, pattern matching and subsequence matching form the given String. Let us consider a String $M = S_1, S_2, S_3 \dots S_n$ can be an string A is substring $S_i, S_j, S_k \dots S_n$ of M and substring A is not essentially connecting in the string of M if $1 < i < j < k < n$. Here the Subsequence matching problem was specified by using CFG G and a substring M ensures there occur in a String M' which belongs to $L(G)$ such that M is a substring of M'. In order to solve the problem of substring matching problem an CYK algorithm and edit distance problem was implemented. Here Edit distance algorithm was implemented in order to find the distance between the String and substring for the porpoise of sequence alignment of string.

Alorithm CYK():

Step 1: Fill the all digonl matrix as closure($S_{i,i}$)
 for I form 1 to n
 $S_{i,i} = \{A \mid A \rightarrow a_i, \in P\}$
 $S_{i,i} = \text{Closure}(S_{i,i})$

Step 2: generate all the substring matching
 for j from 2 to n
 for i from 1 to n-k+1
 $k = i + j - 1$
 $S_i, i = -\phi$
 for l from I to k-1
 $S_i, k = S_i, k \cup \{A \mid A \rightarrow BC \in P, B \in S_i, k, C \in S_{k+1, j}\}$
 $S_{ij} = \text{Cosure}(S_i, k)$

Step 3: Find the distance between to string for sequence alignment
 for j form 1 to m
 for k from 1 to n
 $S[j][k] = 1 + \min(\text{Distance}(S[j][k-1], // \text{Insert}$
 $S[j-1][k], // \text{Remove}$
 $S[j-1][k-1])); // \text{Replace}$

Step 4: Return distance
 $S[m][n]$

Proposed Methodology

This work emphases on the Automatic synchronization of audio-video contents with subtitles. The key element of this process is to guarantee the quality accessible service for people who having the hearing disabilities to make understanding of video contents through subtitle. In order to achieve above key element, an automatic Sub-synchronization of audio-video and sub-titles is implemented, by applying speech recognition technique and making use of parallel word alignment algorithm. There are many algorithm exists which are discussed in the existing methodology for the alignment of word sequence. In this paper, we discussing about proposed methodology for the alignment of word sequence based on time and delay. An efficient technique called Parallel Longest Common Subsequence (PLCS) algorithm has been introduced for the sequence alignment using parallelization of LCS algorithm. While calculating the score matrix for data dependency is done by each Ith row data can be obtained by (I-1)th rows data and each Jth column data can be obtained by (J-1)th columns data. The score matrix table can be obtained for the finite word sequence let $C[1, \text{lenth}]$ and maximum number before J can be obtained $S[i, j] = C[i] * S[i, j]$ which can be computed as:

$$P[I, j] = \begin{cases} 0 & \text{if } j=0 \\ j-1 & \text{if } p_{j-1}=C[i] \\ P[I, j-1] & \text{if } p_{j-1} \neq C[i] \end{cases} \quad (3)$$

The efficient algorithm for Parallel LCS is given below to calculate the distance and score between two sequence alignments:

Algorithm: PLCS(X, Y)

Step 1: Find the length of two sequence i.e. n, m
 $m = \text{length}(X)$
 $n = \text{length}(Y)$
 Step 2: Initialize the first row and first column as Zero.
 Step 3: calculate the row data
 For j form 1 to m //parallelization
 For k form 1 to n
 if $P[j, k-1] \neq c[j]$
 $P[j][k] = P[j, k-1]$
 else
 $P[j][k] = k$
 Step 4: Find the optimal distance between two sequences
 For j from 1 to m
 For k from 1 to n par-do //parallelization
 If $P[c, k] = 0$
 $\text{Str}[j, k] = \text{Str}[j-1, k] + 1$
 else
 $\text{Str}[j, k] = \text{Str}[j-1, k] + \text{Str}[j-1, P[c, k]-1] + 1$
 else
 $\text{Str}[j, k] = S[j, k-1]$

Here we are extracting audio from the video file and the audio file is given for the feature extraction that can be done by using Adaptive MFCC technique and these features are to be trained using HMMBW algorithm. As a result, each sample have the probability of a voice. The probability of a voice sample are stored in transcription Memory in the form of array.

Besides the audio, we have not synchronized subtitles in a SRT file, whereas all subtitle contains a timestamp form its beginning till its ending of subtitle. After getting each subtitle length, easily we can build an array for the data extracted through

MFCC and whereas each cell represents the same time length. In this approach, the transcription and subtitle are aligned by using the PLCS alignment algorithm between the transcription memory storage and subtitles. Whereas transcription storage contains the sequence of words. Once the sequence words alignment is recognized. Then in order to estimate the appearance time of subtitles (t_{pi}) can obtained as soon as the word generated in the transcription storage (t) and to compute the best score of the aligned word from its position (k) in the subtitle given as.

$$T_{pi} = t - k * 0.385 \quad (4)$$

Once the generated word time has been established and to define the deletion time can be computed as:

$$T_b = t_{pi} + d \quad (5)$$

$d = \text{length}(\text{subtitle}) / \text{No of character (per second)}$.

The below Figure 2 show the proposed Architecture for Audio-video content and its subtitle synchronization.

An experimental result shows automated synchronization of audio-visual and subtitle in the Table 1. Here we have tested the different Kannada video file with different duration. In this paper various algorithm has been tested for the sequences alignment such as DTW, LCS and Needleman-wanch algorithm was selected to measure the character sequence from the transcription memory (storage) and with subtitle. The efficiency of the entire algorithm is computed based on the delay time, length of subtitle and time for processing. If subsequence is too large then time efficiency may goes high to find the distance and delay between the transcription and subtitle. By considering the this problem we have implemented the proposed

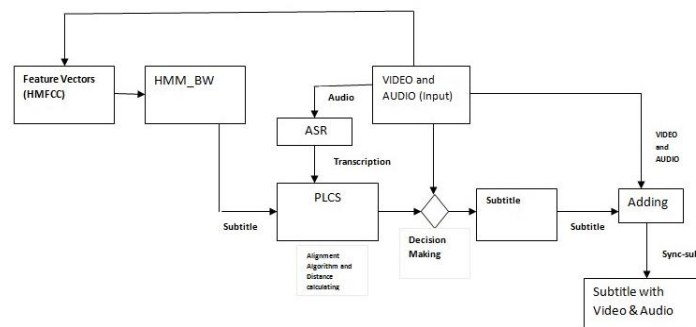


Figure 2. Architecture for Audio-video content and its subtitle synchronization.

algorithm for sequence alignment such as PLCS. The proposed algorithm gives the best result in terms of time efficiency and computing the delay between the transcription and subtitle.

Table 1. The average of all 50 video file in-terms of delay, number of subtitle, duration.

Technique	Number of subtitle	Duration	Delay
LCS	36300	3h25mn	4.65sec
DTW	36300	3h25min	4.10sec
NEEDLEMA	36300	3h25min	3.02sec
N-WUNCH	36300	3h25min	3.02sec
CFG	36300	3h25min	3.02sec
PLCS	36300	3h25min	2.83sec

Conclusion

In this work, automated subtitle synchronization with audio and video is implemented for Kannada video file which contains News, Songs, short movies etc. The system is based on automatic speech recognition engine which receives the audio file, then convert it into text which are stored in transcription memory (Array). In order to make synchronization with audio and subtitle a proposed algorithm is implemented such as Parallel Longest Common Subsequence algorithm (PLCS) which is used to make the alignment between Transcription received from ASR engine along with subtitle already generated from ASR during training. Using PLCS algorithm we will be able to calculate the score, distance between the sequences of word alignment. Finally it will calculate the delay between the audio and subtitle. Based on the delay subtitle will be rearranged and synchronized with audio-video. In this paper, we also implemented other sequence alignment techniques in order to make comparison with the proposed approach. By observing the result the proposed method gives the best result in terms of time efficiency compared to other techniques. Here result is recorded by testing 50 different video file with different duration of video. The result shown in the Table 1 is the average of all 50 video file in-terms of delay, time, Number of subtitle and duration.

References

- Anagnostopoulos CN., Iliou T., and Giannoukos. I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review* Springer., 43:155-77.
- Aparício, M., Figueiredo, P., Raposo, F., de Matos, DM., Ribeiro, R., and Marujo L. (2106). Summarization of films and documentaries based on subtitles and scripts. *Pattern Recognition Letters Elsevier.*, (73):7-12.
- Araújo, de., TM., Ferreira FL., dos Santos Silva DA., Lemos FH., Neto GP., Omaia D., de Souza Filho GL., and Tavares TA (2013). Automatic generation of Brazilian sign language windows for digital TV systems. *J. of the Brazilian Computer Society*. Springer., 107-25.
- Biswas., Astik, PrakashKumarSahu., and Mahesh Chandra. (2015). Multiple cameras in-car audiovisual speech recognition using phonetic and viremic information. *Com. and Electrical Eng.*, (47):35-50.
- Cuzzco-Calleet, I., Ingavélez, P., Robles-Bykbaev, V., and Calle-López. D. (2018). An interactive system to automatically generate video summaries and perform subtitles synchronization for persons with hearing loss. *Proc. IEEE 20th Int. Conf. Electron. Electr. Eng. Compute.*, 1-4.
- Deventer van, M.O., Stokking, H., Hammond, M., Le Feuvre, J., and Cesar., P. (2016) Standards for multi-stream and multi-device media synchronization *IEEE Commun. Mag.*, 54(3):16-21.
- Disabil. Soc. (2011), The world report on disability. *Disabil.Soc.*, 26(5):655-658.
- Gao, J., Zhao, Q., Li, T., and Y, Yan. (2009). Simultaneous synchronization of text and speech for broadcast news subtitling. *Advances in Neural Net-works (Lecture Notes in Computer Science).*, 5,553:576-585.
- Garcia, J.E., Ortega, A., Lleida, E., Lozano, T., Bernues. E.D. (2009). Sanchez Audio and text synchronization for TV news subtitling based on auto-matic speech recognition. In: *Proc IEEE Int. Symp. Broadband Multime-dia Syst, Broadcast*, p. 1-6
- Health Law, Eur. J. (2006) Convention on the rights of persons with disabilities. vol. 14, no. 3, p. 98-281
- Hirschberg, D.S. (1975) A linear space algorithm for computing maximal common subsequences. *Commun. ACM.*, 18(6):341-343.
- Huang, C., Hsu, W., and Chang, S. (2003). Automatic closed caption alignment based on speech recognition transcripts. In: *Columbia Univ. NY, USA, Tech. Rep.*, p 007.
- Karpukhin, I.A. and Konushin, A.S., (2017). Constructing a speech audio-videocorpus by aligning long segments of speech and text *MoscowUniv.Comput. Math.Cybern.*, 41(2):97-103.
- Kedačićet, D., Herceg, M., Peković, V., and Mihic, V. (2018). Application for testing of video and subtitle synchronization. *Proc. Int. Conf. Smart Syst.Technol.*, 23-27.
- Lert wong khanakool, N., Punyabukkana, P., and Suchato, A. (2013). Real-time synchronization of live speech with its transcription. *Proc. 10th Int. Conf. Elect. Eng/Electron, Comput, Telecommun. Inf. Technol.*, 1-5.
- Ley. (2010) General De La Comunicación Audiovisual, BOE, Madrid, Spain, p. 2,010
- Loni, Deepali Shaila and Shaila Subbaraman. (2015). Singing voice identification using a harmonic spectral envelope. In: *Information Processing (ICIP), International Conference on IEEE*, p. 119-123.
- Montagud, M., Boronat, F., González, J., and Pastor. J. (2017). Web-based platform for subtitles customization and synchronization in multi-screen scenarios. *Proc. Adjunct Publication ACM Int. Conf. Interact. Experiences TVOnlineVideo.*, 81-82
- Orero., D.A.C.J.P. and Remael. A. (2007). Media for All: 'Subtitling for the Deaf, Audio Description, and Sign Language'. Amsterdam, The Netherlands: Rodopi. vol. 30.
- Rodríguez-Alsina, A., Talavera, G., Orero, P., and J. Carrabina. (2013). Subtitle synchronization across multiple screens and devices. *Sensors.*, 12(7):8,710-8,731.

- Schmidhuber and Jürgen. (2015). Deep learning in neural networks: An overview. *Neural networks*. 61:85-117.
- Stan Adriana, Yoshitaka Mamiya, Junichi Yamagishi, Peter Bell, Oliver Watts, Robert AJ Clark, and Simon King. (2016). ALISA: An automatic lightly supervised speech segmentation and alignment tool. *Computer Speech and Language*., 35:116-133.
- Szeto and Elson. (2015). Community of Inquiry as an instructional approach: What effects of teaching, social and cognitive presences are there in blended synchronous learning and teaching. *Computers and Education*., 81:191-201.
- Vanderplank and Robert. (2016). Some essential themes in building the case for captions in language learning. In: *Captioned Media in Foreign Language Learning and Teaching*. Palgrave Macmillan, UK, p. 7-41.
- Wilkerson J. and Casas. A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. In: *Annu. Rev. Political Sci.*, vol. 20, no. 1, p. 529-544.
- Wright, SJ., Kanevsky, D., Deng, L., He, X., Heigold, G., and Li, H. (2013). Optimization algorithms and applications for speech and language processing. *IEEE Transactions on Audio, speech Language Processing*., 2,231-43.
- Wunsch, C.D., and Needleman. S.B. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol.Biol.*, 48(3):443-453.
- Xuemei, W., Yue, J., and Fang., W. (2017), The study of subtitle translation based on multi-hierarchy semantic segmentation and extraction in digital video. *Humanities Social Sci.*, 5(2):91-96.
- Zhang, Y., Tang, Z., Zhang, C., Liu, J., and Lu, H. (2015). Automatic face annotation in TV series by video/script alignment. *Neurocomputing*., 152:316-321.