ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AUTOMATIC SPEECH RECOGNITION SYSTEM FOR LECTURE VIDEOS

Aditi V. Lawate*, Prof. M. M. Wankhade

* M.E.(Signal Processing), Department of Electronics and Telecommunications, Sinhgad Collage of Engineering, Vadgaon(Bk) Pune, India.

Assistant Professor, Department of Electronics and Telecommunications, Sinhgad Collage of Engineering, Vadgaon(Bk) Pune, India.

ABSTRACT

The increase of video lecture data on World Wide Web is rapid therefore an efficient method of data retrieval is needed. So the system providing a method for data retrieval from the lecture video is implemented which will extract the speech data. Automatic video segmentation is applied first. Automatic Speech Recognition (ASR) applied on the lecture video tracks that are the speech contents given by lecturer during the video. The extracted text data is saved in the form of templates for future reference. The 80.97% of accuracy in recognising the letters from lecture videos is achieved and the extracted information is saved so that the quality of learning is improved.

KEYWORDS: ASR, MFCC, DTW, feature extraction.

INTRODUCTION

In the age of e-learning, the amount of lecture data or any video data providing an approach for e-leaning is increasing daily. So, it is required to design a system which will retrieve the speech data from the lecture video. In order to fulfill the need of data retrieval from the lecture videos, a system which provides an approach for data retrieval from the lecture video can be implemented.

The main aim of the system is to provide an efficient way of data retrieval from a lecture video. First of all apply the automatic video segmentation. Subsequently extract the speech data by applying video Automatic Speech Recognition (ASR) technology on the video tracks. Automatic Speech Recognition (ASR) transcripts as are taken for keyword extraction and to prepare a relevant document giving the brief idea about the corresponding lecture video. So one can also store it in the text lines in the form of templates and can be used whenever it is needed without going through the lecture video again.

SYSTEM OVERVIEW

This system presents an approach for getting the textual information from a given video lecture. The video lecture here must be taken in .avi format. If the video is in any other unspecified format then it must be converted into .avi format. Audio from this video is also extracted as we are taking audio into consideration. Then the word segmentation, feature extraction algorithms are applied which are explained in details.

Audio Extraction

The speaker-independent isolated word recognition system is used as speech recognition method in this work presented. Its fundamental success subject is to determine the boundary of the speech signal. The speech signal is acquired via lecture video. The real-time speech signal of a certain speaker is peeked using the MATLAB data acquisition toolbox.

Audio Pre-Processing

DC offset is the drift of the speech signal from zero that occurs due the variations in recording conditions and problems in recording device. So to adjust the DC value of the speech signal to zero, input signal is subtracted from mean of the stored speech samples. A direct current offset in a speech wave is typically an artifact of the recording process. One of the first processing steps involved in speech recognition is to remove the direct current offset of the speech wave.

Word Segmentation

Word boundary detection of the speech signal acquired from the input is performed using its energy and zero-crossing rate. Pre-processing of Speech Signal serves various purposes in any speech processing application. It includes Noise Removal, Endpoint Detection, Pre-emphasis, Framing, Windowing, Echo Cancelling etc. Out of these, silence/unvoiced portion removal along with endpoint detection is the fundamental step for applications like Speech and Speaker Recognition. The proposed method uses Probability Density Function (PDF) of the background noise and a Linear Pattern Classifier for classification of Voiced part of a speech from silence/unvoiced part. The work shows better end point detection as well as silence removal than conventional Zero Crossing Rate (ZCR) and Short Time Energy (STE) function methods.

There are several ways of classifying (labeling) events in speech. It is accepted convention to use a three-state representation in which states are Silence (S), where no speech is produced; Unvoiced (U), in which the vocal cords are not vibrating, so the resulting speech waveform is aperiodic or random in nature and Voiced (V), in which the vocal chords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic. It should be clear that the segmentation of the waveform into well-defined regions of silence, unvoiced, signals is not exact; it is often difficult to distinguish a weak, unvoiced sound from silence, or weak voiced sound from unvoiced sounds or even silence.

However, it is usually not critical to segment the signal to a precision much less than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications. Since for most of the practical cases the unvoiced part has low energy content and thus silence (background noise) and unvoiced part is classified together as silence/unvoiced and is distinguished from voiced part.

The problem of automatic word boundary detection in quite environment in presence of noise is addressed in this part. A fast and robust algorithm for accurately locating the endpoints of isolated words is described below. The algorithm utilizes the energy to acquire reference points. The required characteristics of an ideal word boundary detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no a priori knowledge of the noise. All these issues are solved with this algorithm. The algorithm calculates average ambient noise for one frame and takes frames up to say one seconds and averages it overall the frames in this one second to get the reference value of the noise. We calculate the energy per frame first i.e.

$$P[1..m] = \sum_{k=1}^{j} s[k]^2 \qquad (1)$$

Where, s[k] are the speech data in frame. Similarly P is calculated for all the frames and an average is taken for the final noise value [E].

$$E = \frac{\sum_{k=1}^{m} p[k]^2}{m} \qquad (2)$$

The threshold is set at (constant* E), as the detecting criterion.

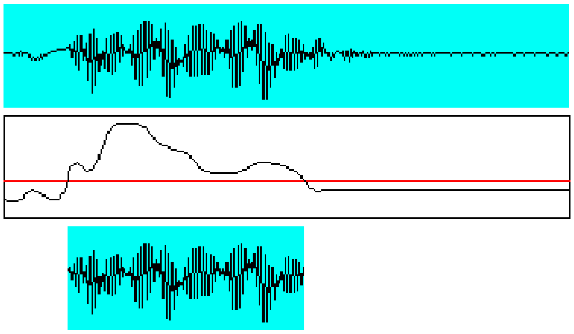

Figure 1: Showing silence removal with the clipped signal [7]

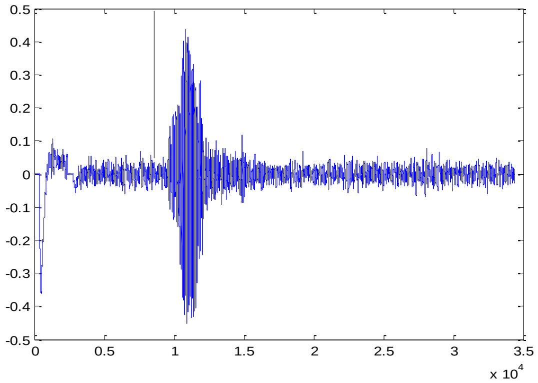

Figure 2: Speech input from the video. The black line shows the place where the endpoint detecting algorithm suggested the truncation. [7]

Feature Extraction

The Mel-Frequency Cepstral Coefficient (MFCC) method is utilized for extracting speech features. As was shown in perception experiments, the human ear does not show a linear frequency resolution but builds several groups of frequencies and integrates the spectral energies within a given group. Furthermore, the mid-frequency and bandwidth of these groups are non–linearly distributed. The non-linear warping of the frequency axis can be modeled by the so called Mel-scale. The frequency groups are assumed to be linearly distributed along the Mel-scale. The Mel frequency *fMel* can be computed from the frequency *f* as follows:

$$fMel(f) = 2595.\log\left(1 + \frac{f}{700Hz}\right) \qquad (3)$$

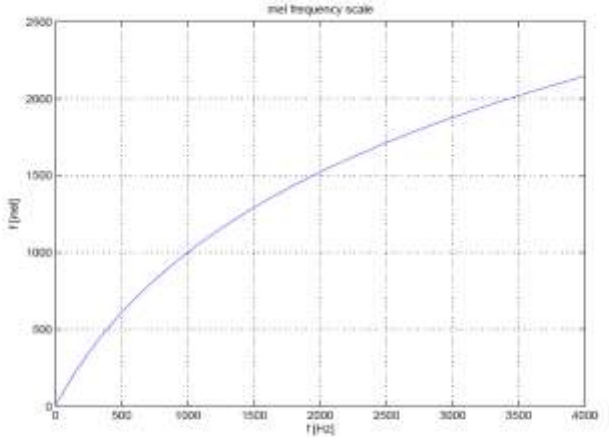

Figure 3: A plot of the Mel scale [7]

The human ear has high frequency resolution in low–frequency parts of the spectrum and low frequency resolution in the high–frequency parts of the spectrum. The coefficients of the power spectrum $|V(n)^2|$ are now transformed to reflect the frequency resolution of the human ear. A common way to do this is to use K triangle–shaped windows in the spectral domain to build a weighted sum over those power spectrum coefficients $|V(n)^2|$ which lie within the window. We denote the windowing coefficients as,

ηkn*:* $k = 0,1,\ldots k\text{-}1$; $n = 0,1,\ldots \mathrm{N}/2$ (4)

This gives us a new set of coefficients,

$G(k)$; $\mathrm{k} = 0,1,\ldots k\text{-}1$ (5)

The so called Mel spectral coefficients,

$$G(k) = \sum_{n=0}^{N/2} \eta kn \cdot \left|V(n)\right|^2 ; k = 0,1,\ldots k-1 \qquad (6)$$

Caused by the symmetry of the original spectrum, the Mel power spectrum is also symmetric in k:

$G(k) = G(-k)$ (7)

Therefore, it is sufficient to consider only the positive range of k, k = 0, 1, . .K− 1.

The cepstral coefficients, which are the coefficients of the Fourier transform representation of the log magnitude spectrum, have been shown to be a more robust, reliable feature set for speech recognition than the LPC coefficients. Because of the sensitivity of the low order cepstral coefficients to overall spectral slope and the sensitivity of the high-order cepstral coefficients to noise, it had become a standard technique to weight the cepstral coefficients by a tapered window so as to minimize these sensitivities.

Speech Recognition

Dynamic Time Warping (DTW) measures the similarity between test signal which is the speech acquired from the video input and the reference signal pre-stored in the database. It then equalizes the length of the test signal with length of the reference signal in time. Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another they were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics indeed, any data which can be turned into a linear representation can be analyzed with DTW.

A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequence with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

In this type of speech recognition technique the test data is converted to templates. The recognition process then consists of matching the incoming speech with stored templates. The template with the lowest distance measure from the input pattern is the recognized word. The best match (lowest distance measure) is based upon dynamic programming. This is called a Dynamic Time Warping (DTW) word recognizer. In order to understand DTW, two concepts need to be dealt with,

i. Features: the information in each signal has to be represented in some manner.

ii. Distances: some form of metric has be used in order to obtain a match path. There are two types:

(a)Local: a computational difference between a feature of one signal and a feature of the other.

(b)Global: the overall computational difference between an entire signal and another signal of possibly different length.

Since the feature vectors could possibly have multiple elements, a means of calculating the local distance is required. The distance measure between two feature vectors is calculated using the *Euclidean* distance metric. Therefore the local distance between feature vector x of signal 1 and feature vector y of signal 2 is given by,

$$d(x,y)=\sqrt{\sum_{j}\left(x_{j}-y_{j}\right)^{2}} \qquad (8)$$

DTW Algorithm:

Speech is a time-dependent process. Hence the utterances of the same word will have different durations, and utterances of the same word with the same duration will differ in the middle, due to different parts of the words being spoken at different rates.

This problem is illustrated in figure 4, in which a “time-time" matrix is used to visualize the alignment. As with all the time alignment examples the reference pattern (template) goes up the side and the input pattern goes along the bottom. In this illustration the input “SsPEEhH” is a ‘noisy’ version of the template “SPEECH”. The idea is that ‘h’ is a closer match to ‘H’ compared with anything else in the template. The input “SsPEEhH” will be matched against all templates in the system’s repository. The best matching template is the one for which there is the lowest distance path aligning the input pattern to the template

This algorithm is known as the Dynamic Programming (DP). When applied to template based speech recognition, it is referred to as Dynamic Time Warping (DTW). DP is guaranteed to find the lowest distance path through the matrix, while minimizing the amount of computation. The DP algorithm operates in a time-synchronous manner: each column of the time-time matrix is considered in succession so that, for a template of length N, the maximum number of paths being considered at any time is N.

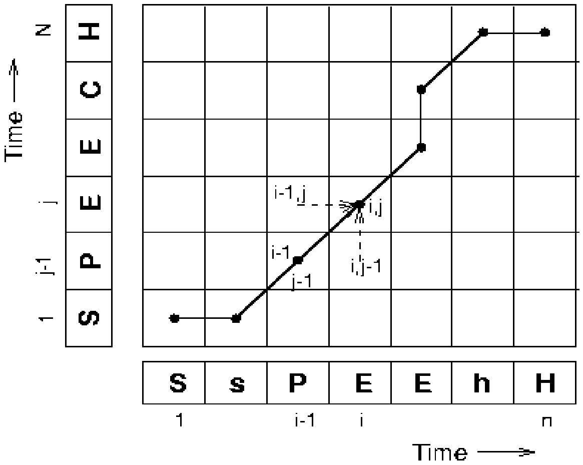

Figure 4: Time alignment between the test and the training pattern [7]

If D (i, j) is the global distance up to (i, j) and the local distance at (i, j) is given by d (i,j) then,

$$D(i,j) = \min[D(i\text{-}1,j\text{-}1), D(i\text{-}1,j), D(i,j\text{-}1)] + d(i,j) \qquad (9)$$

Given that $D(1,1) = d(1,1)$ (is the initial condition), we have the basis for an efficient recursive algorithm for computing *D*(i, j). The final global distance *D*(n, N) gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score.

System flowchart

The system flowchart shows the algorithm for the speech extraction from the lecture video. Using this algorithm the speech data is extracted and is saved in the form of text templates which can be used for future reference (Figure 5).

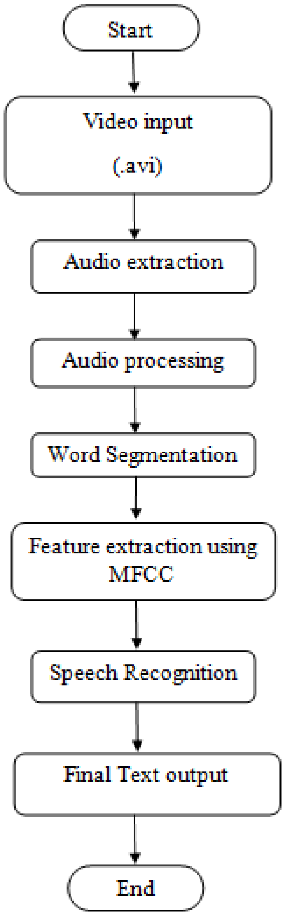

Figure 5: System Flowchart

RESULTS AND DISCUSSION

For automatic speech recognition, the system uses the Mel frequency cepstral coefficient matching for recognition of the word. The table 1 shows the number of actual words to be recognized and the number of words which are recognized using ASR algorithm. Here the five videos are taken from the YouTube video channel. This will show the accuracy of speech recognition for five individual videos and the average accuracy which gives the overall accuracy of the system. The overall accuracy of the system for automatic speech recognition is 80.97%.

Formula:

Accuracy of the system can be calculated as,

$$\text{Accuracy} = \frac{\text{Total number of recognized words}}{\text{Total number of words to be recognized}} \text{ x } 100$$

Table 1: *System accuracy results for ASR*

Video	Actual Words	Recognized Words	Accuracy	Average Accuracy
1	182	148	81.31%	
2	170	139	81.76%	
3	194	153	78.86%	80.97%
4	178	144	80.89%	
5	167	137	82.03%	

CONCLUSION

The system provides an approach for data retrieval from the lecture video which will extract the speech information from the lecture video. ASR technique have to be implemented on the lecture audio tracks respectively. Extracted information is saved which will give the brief idea of the video and which can be used for future reference to understand the lecture contents more efficiently. The system will improve the quality of e-learning and hence is more useful in the age of e-learning. Accuracy of the system differs from video to video but the overall accuracy do not have much difference. Therefore the proposed system gives more efficient way for the metadata extraction using ASR methodology and therefore it can be implemented where the data retrieval is needed.

ACKNOWLEDGEMENTS

I am indeed thankful to my guide Prof. M. M. Wankhade for her able guidance to complete this paper. I extend my special thanks to Head of Department of Electronics and Telecommunications Dr. M. B. Mali who extended the preparatory steps of this paper-work. I am also thankful to the Principal Dr. S. D. Lokhande, Sinhgad College of Engineering for his valued support and faith on me.

REFERENCES

1. Haojin Yang and ChristophMeinel, “Content Based Lecture Video Retrieval Using Speech and Video Text Information” in IEEE Transactions on Learning Technologies, Vol. 7, No. 2,April-June 2014.
2. E. Leeuwis, M. Federico, and M.0020Cettolo, “Language modeling and transcription of the ted corpus lectures,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2003, pp. 232–235.
3. C. Munteanu, G. Penn, R. Baecker, and Y. C. Zhang, “Automatic speech recognition for webcasts: How good is good enough and what to do when it isn’t,” in Proc. 8th Int. Conf. Multimodal Interfaces, 2006.

4. D. Lee and G. G. Lee, “A korean spoken document retrieval system for lecture search,” in Proc. ACM Special Interest Group Inf. Retrieval Searching Spontaneous Conversational Speech Workshop, 2008.
5. M. Grcar, D. Mladenic, and P. Kese, “Semi-automatic categorization of videos on videolectures.net,” in Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases, 2009, pp. 730–733.
6. www.computer.org/cdsl/trans/lt/2014/02/06750040-abs.html.
7. http://staffwww.dcs.shef.ac.uk/people/S.Wrigley/com326/sym.html.
8. J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, “Analysis and processing of lecture audio data: Preliminary investigations,” in Proc. HLT-NAACL Workshop Interdisciplinary Approaches Speech Indexing Retrieval, 2004, pp. 9–12.
9. https://www.youtube.com/watch?v=8NiMtbsa4Ks&feature=youtu.be.
10. https://www.youtube.com/watch?v=0IwjxVxuaNk&feature=youtu.be.
11. https://www.youtube.com/watch?v=GSwnFoTVPgY&feature=youtu.be.
12. https://www.youtube.com/watch?v=piEWwnCV3LY&feature=youtu.be.
13. https://www.youtube.com/watch?v=GR27Bh-bzL0&feature=youtu.be.