Topic Name: Visual Speech Recognition Based on Lip Movement for Bangla Languages.

Literature Review (Five Papers)

Paper No: 01

Paper Topic: Visual Speech Recognition Based on Lip Movement for Indian Languages.

Literature Review

Everyone can speak without the help of any device and mainly the technical skill set is not needed. The problem with the primitive interfacing devices is, some percentage of basic level of skill set is much necessary to use those interfaces. So it will be difficult to interact with such devices for people who are all not aware of technical skill set. As in this work, main concentration is on speech recognition, any technical skill set is not required so this will be helpful for the people to speak to the computers in known language rather than giving inputs from the other devices of the systems.

In this work whole experiment is worked on 120 samples which are for three Indian languages. We are using the number system from 0 to 9 for recognition of English, Kannada and Telugu languages. The accuracy expected to this dataset is 90%.

In image processing, pre-processing step plays the very important role, as it helps in obtaining the better results by applying several operations on the input image. It is core or basic step performed on all image processing tasks. In our approach frames which are extracted from input video is considered as the input image. It is much necessary for removing noise from the considered frames so that the accuracy and the clarity will be in the better way.

The task of recognizing sharp discontinuity and locating points in a considered image is referred as the edge detection process. For pattern recognition, image edge is the core feature of the image. So for lip pattern recognition edge detection is used.

In this paper, overall experimental description of proposed VSR (visual speech recognition) using the lip parameters is demonstrated. With the visual speech recognition for Indian languages with an accuracy of (yet to add) is achieved. The effective usage of pre-processing step such as de-noising and resizing, followed by canny edge detection algorithm in order to find out true edges of considered image for ROI extraction. Four features like entropy, energy, contrast and correlation are extracted by using the GLCM algorithm along with ANN classifier for accurate classification of visual properties from the considered video. The performance of proposed system is more accurate than that of other conventional methods; experimental results witness the efficiency and accuracy of proposed system. For future works, it can be possible to add both audio and video input parameters for the better performance in the visual speech recognition.

Paper No: 02

Paper Topic: Voiceless Bangla vowel recognition using sEMG signal.

Literature Review

Some people cannot produce sound although their facial muscles work properly due to having problem in their vocal cords. Therefore, recognition of alphabets as well as sentences uttered by these voiceless people is a complex task. This paper proposes a novel method to solve this problem using non-invasive surface Electromyogram.

The voiceless Bangla vowels classification process started with data collection, then it is pre-processed for de-noising and removing DC components. Then the feature extraction and feature selection algorithm have been applied, and finally classification was done using the selected features.

Human body is treated electrically neutral due to the same number of positive and negative charges. The nerve cell membrane is polarized in the resting state. When a neuron is stimulated the muscle fiber depolarizes as the signal spreads along the surface and muscle fiber contraction happens. This depolarization, along with the movement of ions, makes an electric field near the muscle fiber. An EMG signal is the train of Motor Unit Action Potential (MUAP) showing the muscle response to neural stimulation. In case of speech delivery, EMG signals are generated in the facial muscles by opening or closing lips, mouth and jaw as well. Consequently, EMG signals also appear in the extrinsic muscles of the tongue.

A limitation of the present study is the number of subjects. Only 8 subjects were used in this study that may hampers on the accuracy rate. However, the aim of this work does not propose a final system, but to explore the possibility of developing such system. This novel work can be extended in a number of ways. The experiment needs to be done on voiceless people to validate the proposed method. The performance of the other neural networks like HMM, SVM with different kernels are needed to investigate for better accuracy. Our near future research will solve the present constraints and explore and extend the present methodology. Finally, it can be said that this novel method for Bangla vowels classification will help the voiceless people who can't produce sound but can move their facial muscles just as the normal people.
The accuracy can be increased by adding more features. However, it is our intention to keep the number of features small that provide better classification accuracy. The methodology developed in this paper is not only useful in Bangla vowels classification but also useful in many biomedical research areas such as EEG seizure detection, brain-computer interface (BCI) etc.

Paper No: 03

Paper Topic: Audio-Visual Speech Recognition Using Lip Movement for Amharic Language.

Literature Review

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to a written text. In recent years, there have been many advances in automatic speech reading system with the inclusion of visual speech features to improve recognition accuracy under noisy conditions. By identifying lip movements and characterizing their associations with speech sounds, the performance of speech recognition systems can be improved, particularly when operating in noisy environments.

The aim of this study is to design and develop automatic audio-visual Amharic speech recognition using lip reading. In this study, for face and mouth detection we use Viola-Jones object recognizer called haarcascade face detection and haarcascade mouth detection respectively, after the mouth detection ROI is extracted. Extracted ROI is used as an input for visual feature extraction. DWT is used for visual feature extraction and LDA is used to reduce visual feature vector. For audio feature extraction, we use MFCC. Integration of audio and visual features are done by decision fusion. As a result of this, we used three classifiers. The first one is the HMM classifier for audio only speech recognition, the second one is HMM classifier for visual only speech recognition and the third one is CHHM for audio- visual integration.

Audio features are still the main contribution and play a more important role, than visual features. However, in some cases, it is difficult to extract useful information from the audio. There are many applications in which it is necessary to recognize speech under extremely adverse acoustic environments. Detecting a person's speech from a distance or through a glass window, understanding a person speaking among a very noisy crowd of people. In these applications, the performance of traditional speech recognition is very limited.

The purpose of this study is to develop an automatic audio-visual speech recognition for Amharic language using the lip movement which include face and lip detection, region of interest (ROI), visual features extraction, visual speech recognition and integration of visual with audio. The architecture of the system that we adopted in our study is the decision fusion architecture. As a result of this architecture, we used three classifiers. The first one is the HMM classifier for audio only speech recognition, the second one is HMM classifier for visual only speech recognition and the third one was CHHM for audio-visual integration.

For implementation we use python programing language and OpenCV. For face and mouth detection we use Viola-Jones object recognizer called haarcascade face detection and haarcascade mouth detection respectively, after the mouth detection ROI extracted.

Paper No: 04

Paper Topic: Visual Speech Recognition Using Artificial Neural Networking

Literature Review

Automatic Speech Recognition plays an important role in human-computer interaction, which can be applied in various application like crime-fighting and helping the hearing- impaired consists of two domain Audio Speech Recognition and Visual Speech recognition. This thesis is based on Recognition of Speech in the visual domain only.

This paper provides a new approach to lip reading Bengali words using a combination of the curvature of the inner and outer lips and Neural Networks. The method uses a more robust a faster algorithm to detect the lip contour than conventional methods used so far.
Processing multiple frames and by collecting the contours, we can predict the Bengali words that are stored inside the database. Our thesis will mainly focus on detecting some specific Bengali words.

Research on Lip-reading suggests that the field of Lip-reading is still in its infancy. A completely accurate and efficient lip-reading algorithm is yet to be developed. The objective of this research is to provide its contribution to this developing field with an algorithm that saves time and provides at least an above average accuracy. Contributions of this thesis are in three areas of lip reading. These are Lip Segmentation, Feature Extraction and Viseme Recognition.

Many algorithms and models have been proposed by researchers for detecting objects and segmenting them. We have done a brief literature review illustrating some of these methods and provided a unique mixture of Active Contour Model and butterfly method which is supposed to outperform all the mentioned algorithms of literature review in many performance measuring parameters.

A webcam or a camera is utilized to get the video of a man talking such that from his articulates each syllable must be recognized, however video should maintain continuity, with no sound. This articulation of words ought to be taken a couple of times, to choose the perfect radiance, to such an extent that it is simpler to play out the resulting steps. This video is then spared in avi or mpgeav arrange. The procured video is then separated into outlines or a picture arrangement, with the end goal that every video outline is presently a different picture record. This method is finished by utilizing MATLAB's image processing toolbox.

The mask of the segmented image of the lip is taken which is then reduced using canny edge detection. Then the left and right most pixel points of the mask are found. Then the midpoint of the two points is used to find the boundary region which is 20% of the distance from the midpoint. Afterwards the lowest pixel is selected from within the boundary region, which is called the dipping point.

There are many advantages of this proposed technique. First and foremost, it improves upon some of the drawbacks of the existing methods of contour extraction. Adding robustness and accuracy to image-based algorithms, it it extracts the curvature of the lips and so, the results are independent of the size or quality of the picture, illumination or mouth rotation.

Paper No: 05

Paper Topic: Visual Speech Recognition

Literature Review

Lip reading is used to understand or interpret speech without hearing it, a technique especially mastered by people with hearing difficulties. The ability to lip read enables a person with a hearing impairment to communicate with others and to engage in social activities, which otherwise would be difficult. Recent advances in the fields of computer vision, pattern recognition, and signal processing has led to a growing interest in automating this challenging task of lip reading. Indeed, automating the human ability to lip read, a process referred to as visual speech recognition (VSR) (or sometimes speech reading), could open the door for other novel related applications.

VSR has received a great deal of attention in the last decade for its potential use in applications such as human-computer interaction (HCI), audio-visual speech recognition (AVSR), speaker recognition, talking heads, sign language recognition and video surveillance. Its main aim is to recognize spoken word(s) by using only the visual signal that is produced during speech. Hence, VSR deals with the visual domain of speech and involves image processing, artificial intelligence, object detection, pattern recognition, statistical modelling, etc.

There are two different main approaches to the VSR problem, the visemic* approach and the holistic approach, each with its own strengths and weaknesses. The traditional and most common approaches to automatic lip reading are based on visemes. A Viseme is the mouth shapes (or appearances) or sequences of mouth dynamics that are required to generate a phoneme in the visual domain. However, several problems arise while using visemes in visual speech recognition systems such as the low number of visemes (between 10 and 14) compared to phonemes (between 45 and 53). Visemes cover only a small subspace of the mouth motions represented in the visual domain, and many other problems. These problems contribute to the bad performance of the traditional approaches; hence, the visemic approach is something like digitizing the signal of the spoken word, and digitizing causes a loss of information.

The holistic approach such as the “visual words” (Hassanat, 2009) considers the signature of the whole word rather than only parts of it. This approach can provide a good alternative to the visemic approaches to automatic lip reading. The major problem that faces this approach is that for a complete English language lip reading system, we need to train the whole of the English language words in the dictionary! Or to train (at least) the distinct ones.

The major challenge for VSR is the lack of information in the visual domain, compared to the audio domain, perhaps because humans have yet to evolve to have need of a more sophisticated communication system. For example, it was sufficient for man’s survival to use sound to warn friends if there was an enemy or a predator around without having to see them. Therefore, humans did not worry about producing a sufficient visual signal while talking. This major challenge, along with some others, opens the door for more research in the future, to compensate for the lack of information.

Findings

After researching five papers we found a lot of limitation in our proposed paper. We want to develop an algorithm by which can detect visual speech only using movement of lip for Bangla language. Some major limitations are:

I. Lack of information about appropriate algorithm.
II. Lack of source for Bangla language.
III. We are not familiar to these type of recognition algorithm.
IV. Algorithms are not working on MATLAB's image processing toolbox.
V. Shortage of researching time.
VI. Serious sickness of group member.
VII. A limitation of the present study is the less number of subjects.
VIII. The performance of the other neural networks like HMM, SVM with different kernels are needed to investigate for better accuracy.

Moreover we have to research again and again for implementing the Visual Speech Recognition for Bangla language. We are very hopeful to implementing proposed system though we have still three months.