A Hybrid Approach for Speech Recognition Using Visual Features

Syed Sabeel Abbas Naqvi[1], Gaurav Shrivastav [2]

[1]Computer Science & Engineering, Institute of Technology & Management, Aligarh, sabeelnaqvi@gmail.com

[2] Associate Professor, Computer Science & Engineering, Institute of Technology & Management, Aligarh

***Abstract*— Lip perusing, otherwise called visual discourse preparing, and implies acknowledgment of talked word in light of the example of lip developments while talking. In lip perusing choice of components assume import part. In lip perusing applications database is video, so 3 Dimensional changes is proper to concentrate lip movement data. Proposed lip perusing depends on casing standardization and edge insightful element extraction. In this paper 3D change based system is proposed for highlight extraction. These components are the contribution to Multi-Layer Perceptron (MLP) neural system show for discriminative investigation. The proposed visual words approach utilizes a mark 2-dimensional element framework that speaks to a whole talked word. The mark of a talked word is an accumulation of 2features. These incorporate lip geometric and lip appearance highlights.**

Keywords— HMM, ASR (Automatic Speech Recognition), GUI (Graphical User Interface), Speech Recognition, lip detection.

I. INTRODUCTION

Talked dialect learning for hearing weakened individuals, besides leftover tuning in, lip perusing is an imperative channel to comprehend the data. On the off chance that the hearing impeded individuals need to have typical social lives, they should have clear talked and lip perusing capacities to speak with other individuals. At that point they can without much of a stretch adjust to learning, occupation and family life. Lip-perusing acknowledgment is not completely precise and solid, ordinary hearing individuals and hearing debilitated individuals really need to utilize it consistently. In the event that hearing weakened individuals need to upgrade their correspondence with other individuals besides wearing listening device instrument, lip-perusing acknowledgment is one fundamental learning strategy. [1]

Examines on programmed acknowledgment of cluttered discourse have been cantered on acknowledgment of acoustic discourse just whose execution corrupts within the sight of surrounding commotion. Be that as it may, visual elements produced from the speaker's lip locale have as of late been proposed as a dynamic methodology to upgrade ordinary discourse acknowledgment. [2] Clients Friendly System GUI based So that each client needn't bother with any significant instructional meetings before utilizing the framework. [3] A framework which enables the client to communicate the framework for attempt on lipstick. Client can choose shading on a shading scale and see as though she tries on this shade of lipstick. For this framework, the calculation is sorted out as takes after. Initially, confront range is distinguished, at that point confront picture is portioned and lip zone is isolated from the face picture. The picture of lips zone is changed over the HSV (Hue Saturation Validation) shading space. In the wake of building up the shading division handle, lips are recognized by the framework. At long last; lip's tone esteem is changed with the tint estimation of the shading which is chosen by the client. At last it is changed over to RGB shading space and changed picture is appeared on the screen. [4]

Lip discovery utilizes the upgraded form of Lip-Map that proposed in for better division amongst lip and skin pixels, duplicate Lip-Map by immersion segment of HSI shading space. Keeping in mind the end goal to lessen required figuring, the principal evacuate upper portion of the face picture. After this progression, evaluate the lip zone. For this reason, isolate remaining lower half piece of into a few sections and ascertain standard deviation for each part. In light of standard deviation of each part, decide lip range. At long last, for concentrate lip pixels from skin pixels, get ideal limit an incentive to change over the dim scale picture into parallel picture. White pixels in double picture are lip area. [5] Various methodologies, for example, splines, dynamic forms, and parametric models in the writing so as to speak to and remove the lip shape. Established dynamic shapes and splines experience the ill effects of complex parameter tuning and they are for the most part not able to flawlessly fit to the trademark lip parts, for example, Cupidon's bow due to the incorrect slope data because of enlightenment contrasts. Proposed to fit cubic polynomials on the external lip shape utilizing the shading data of the lip picture. [6] The structure of LCACM, through which the shading objects consolidating complex appearances or force in homogeneities, can be viably sectioned. As needs be, introduced a two-stage nearby locale based way to deal with lip following. Being versatile to the lip developments, the proposed approach highlights: (1) less pre-preparing steps, for example, teeth evacuation, preparing information catch and preparing procedure, and (2) execution heartiness to the presence of teeth, tongue and dark hole.[7] lip movement portrayals and proposed a two-organize discriminative lip include choice strategy for speaker distinguishing proof and discourse perusing response.

- Explicit lip movement is helpful notwithstanding lip force as well as geometry;
- Grid-based thick lip movement elements are better and heartier thought about than shape based lip movement.

The Bayesian discriminative component choice serves likewise as a middle of the road measurement lessening venture before the transient LDA, by effectively choosing the lip highlights that are custom-made for the particular acknowledgment issue. [8]

In this paper, we use the multilayer Perceptron and fuzzy logic, the combination of the both to achieve the best accuracy of the video. C++ compiler is also used. Fuzzy logic is work on the true or false. Multilayer Perceptron deal with the nonlinear classification problems because it can form more complicated decision regions, and the haar cascade classifier is used for the video. Videos are taken from the Smartphone and apply in the neural networks taken out the process time of the each video. The process time defines the accuracy of the video. The process times of the videos are show in the table. Fig 4.1 & 4.2 show the process time and the dimensions of the lip movements.

II. RELATED WORK

Yun-Long Lay et al. Inside the correspondence procedure of people, the speaker's outward appearance and lip-shape development contains greatly rich dialect data. The hearing debilitated, besides utilizing leftover tuning in to speak with other individuals, can likewise utilize lip perusing as a specialized instrument. As the hearing disabled take in the lip perusing utilizing a PC helped lip-perusing framework, they can unreservedly learn lip perusing without the imperatives of time, place or circumstance. Consequently, they propose a PC helped lip-perusing framework (CALRS) for phonetic elocution acknowledgment of the right lip-shape with a picture handling technique, protest situated dialect and neuro-organize. This framework can precisely think about the lip picture of Mandarin phonetic articulation utilizing self-arranging map neuro-organize (SOMNN) and augmentation hypothesis to enable hearing weakened to rectify their elocution.

Elham S. Salama et al. Discourse acknowledgment of turmoil individuals is a troublesome undertaking because of the absence of engine control of the discourse articulators. Multimodal discourse acknowledgment can be utilized to upgrade the strength of scattered discourse. A programmed discourse acknowledgment framework for individuals with dysarthria discourse issue in light of both discourse and visual parts. The Mel-Frequency Cepestral Coefficients (MFCC) is utilized as elements speaking to the acoustic discourse flag. For the visual partner, the Discrete Cosine Transform (DCT) Coefficients are extricated from the speaker's mouth area. Face and mouth areas are recognized utilizing the Viola-Jones calculation. The acoustic and visual information elements are then linked on one component vector. At that point, the Hidden Markov Model (HMM) classifier is connected on the consolidated element vector of acoustic and visual parts.

Arvinder Singh, Gagandeep Singh Speech is one of the regular types of correspondence. Late advancement has made it conceivable to utilize this in the security framework and controlling the gadgets. In discourse acknowledgment, the errand is to utilize a discourse test to choose the personality of the individual that created the discourse from among a populace of speakers. An imperative pre-handling venture in Automatic Speech Recognition frameworks is to identify the nearness of clamor. It has been demonstrated that precise discourse endpoint location enhances the disconnected word acknowledgment exactness. Additionally, appropriate area of locales of discourse decreases the measure of preparing. This perspective is likewise essential for portable communication. Discourse acknowledgment frameworks work sensibly well with a tranquil foundation however ineffectively under boisterous conditions or in twisted channels. Such a crisscross in the preparation and testing has extremely constrained. The goal of this calculation is the advancement of flag preparing and examination methods that would give pointedly enhanced discourse acknowledgment precision in an uproarious situations. Discourse is a characteristic medium of correspondence for people, and in the most recent decade different discourse innovations like programmed discourse acknowledgment (ASR), Voice reaction frameworks and another comparable framework have significantly developed.

Gozde Yolcu Oztel, Serap Kazan Online shopping has turned out to be extremely famous as of late. In spite of the upsides of internet shopping as far as time and assorted qualities, it is a burden for clients can't attempt items. This review means to essentially attempt on items. Application produced for lipstick trial. Client can see as though she tries on a lipstick on a screen. Utilizes confront division for distinguishing lips region, shading division for cutting lips and shading space change for better outcome. Regardless of some framework delays, application functions as practical.

Hashem Kalbkhani, Mehdi Chehel Amirani Lip location is utilized as a part of numerous applications, for example, confront recognition and lips perusing. Scientists have considered entire of face picture for lip identification. Propose another calculation. In the calculation for diminishing required computation and increment exactness of right location, don't consider entire of the face picture. Right off the bat expel the upper half piece of the face picture. At that point, for gauge lip zone, isolate remained bring down half face picture to a balance of. For each part compute factual data, for example, standard deviation, and in light of them we identify lip zone in face picture. For particular lip pixels from skin pixels, utilizes YCbCr and HSI shading spaces at this work.

M Haider Mehraj, Ajaz Hussain Mir There is an expanding prerequisite for powerful and dependable individual validation frameworks in territories of high security or secure get to. The greater part of current strategies for individual acknowledgment concentrate on either static facial data or speaker acknowledgment through the discourse flag. While in clean conditions, the discourse flag has turned out to be an important wellspring of speaker ward data, issues happen in boisterous or channel befuddle conditions. While lip data exhibits dominatingly discourse subordinate data, important speaker subordinate data is likewise contained inside the static and dynamic elements of the lips. Presents a similar survey of different lip based biometric systems and makes correlations between them by utilizing Relative Operating Characteristics (ROC) bend and Recognition Time (RT) as execution metric. The ROC bend and acknowledgment time has been gotten at different casing rates to decide the best lip acknowledgment method.

Yiu-ming Cheung et al. Lip following has assumed a noteworthy part in a lip perusing framework. In this paper, introduce a neighborhood locale based way to deal with lip following, which comprises of two stages: (i) lip form extraction for the principal lip outline (ii) lip following in the ensuing lip outlines. At first, develop a limited shading dynamic shading model gave that the closer view and foundation areas around the protest are locally extraordinary in shading space. In the primary stage, finds a joined semi-circle around the lip as the underlying developing bend and register the confined energies for bend advancement with the end goal that the lip picture is isolated into lip and non-lip districts. The proposed approach adjusts to the lip development, as well as vigorous against the presence of teeth, tongue and dark opening. Broad analyses demonstrate the proficiency of the proposed lip following calculation in examination with the current techniques.

H. Ertan Çetingül et al. There have been a few reviews that mutually utilize sound, lip force, and lip geometry data for speaker distinguishing proof and discourse perusing applications. Utilizing unequivocal lip movement data, rather than or notwithstanding lip power and additionally geometry data, for speaker ID and discourse perusing inside a brought together element choice and segregation examination structure, and addresses two imperative issues: 1) Is utilizing express lip movement data helpful, and, 2) assuming this is the case, what are the best lip movement highlights for these two applications. The best lip movement highlights for speaker distinguishing proof are thought to be those that outcome in the most elevated separation of individual speakers in a populace, while for discourse perusing; the best elements are those giving the most astounding phoneme/word/state acknowledgment rate. A few lip movement highlight hopefuls have been considered including thick movement includes inside a bouncing box about the lip, lip form movement components, and mix of these with lip shape highlights. Moreover, a novel two-organize, spatial, and worldly segregation investigation is acquainted with select the best lip movement highlights for speaker recognizable proof and discourse perusing applications.

III. PROPOSED WORK

A. Problem Statement

A highly efficient method of speech recognition using visual features is presented. The method is used for identification of the Hindi words.

B. Objectives

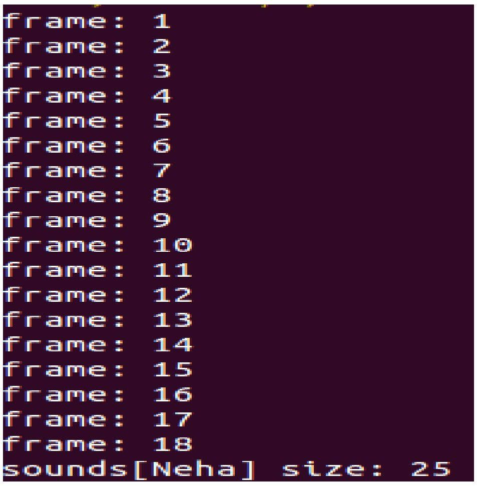
To develop an automatic lip reading system for practical applications

To develop a hybrid approach using lip geometric and lip appearance features and to use the same for recognition purpose by using Multi-Layer Perceptron (MLP) neural network and fuzzy logic in the presented approach

IV. RESULTS AND DISCUSSION

Table 4.1 Sample Video Database

File Number	File Name	Process Time
1	Neha	25
2	Namaste	28
3	Nayeem	25
4	Ladwa	25



```
frame: 1
frame: 2
frame: 3
frame: 4
frame: 5
frame: 6
frame: 7
frame: 8
frame: 9
frame: 10
frame: 11
frame: 12
frame: 13
frame: 14
frame: 15
frame: 16
frame: 17
frame: 18
sounds[Neha] size: 25
```

FIG. 4.1

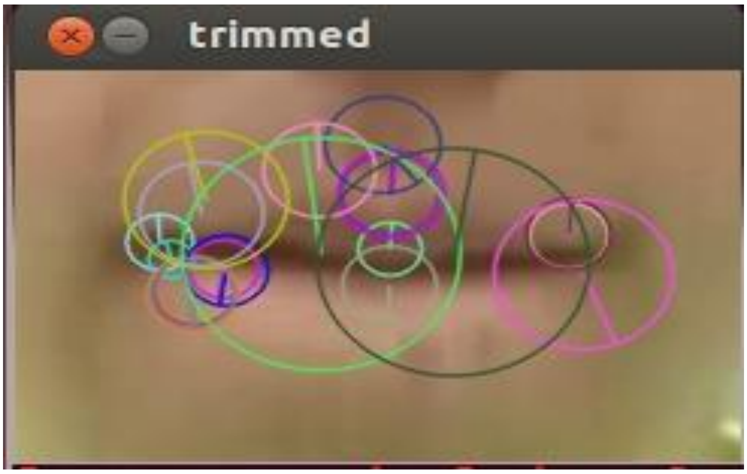


Fig.4.2

V. CONCLUSION

Despite the fact that the visual side of Hindi dialect discourse does not give much data to perceive discourse, either by human or machine, the proposed plot utilizes a way to deal with handle the VSR issue, where the framework perceives the entire word as opposed to simply parts of it (visemes). In this approach, a word is spoken to by a mark that comprises of a few flags or highlights vectors (or highlight network). Each flag is built by fleeting estimations of its related component. The proposed confront confinement strategy is a half and half of the information based approach, format coordinating methodology and highlight invariant approach (skin shading). This technique utilizes a wavelet change to diminish the time required for different checking steps contrasted with filtering the spatial area. The shading data was not influenced much by the lighting conditions, since it was utilized just to pick the best face area utilizing fluffy rationale.

To dispense with the video-particular components, the creator will attempt to discover better elements to substitute for the picture based elements, and to utilize a pre-handling method to settle the light issues.

REFERENCES

[1] Yun-Long Lay, Chung-Ho Tsai, Hui-Jen Yang, Chern-Sheng Lin, Chuan-Zhao Lai, “The application of extension neuro-network on computer-assisted lip-reading recognition for hearing impaired”, IEEE, 2008, PP. 1465–1473.
[2] Elham S. Salama, Reda A. El-Khoribi, Mahmoud E. Shoman, “Audio-Visual Speech Recognition for People with Speech Disorders”, IEEE, 2014, PP. 51-56.
[3] Arvinder Singh, Gagandeep Singh, “Speech Recognition Based System to Control Electrical Appliances”, IEEE, 2012, PP. 81-83.
[4] Gozde Yolcu Oztel, Serap Kazan, “Virtual Makeup Application Using Image Processing Methods”, IEEE, 2015, PP. 401-404.
[5] Hashem Kalbkhani, Mehdi Chehel Amirani, “An Efficient Algorithm for Lip Segmentation in Color Face Images Based on Local Information”, IEEE, 2012, PP. 12-16.
[6] M Haider Mehraj, Ajaz Hussain Mir, “Lip Based Recognition: A Comparative Analysis”, IEEE, 2014, PP. 153-157.
[7] Yiu-ming Cheung, Xin Liu, Xinge You, “A local region based approach to lip tracking”, IEEE, 2012, PP. 3336–3347.
[8] H. Ertan Çetingül, Yücel Yemez, Engin Erzin, A. Murat Tekalp, “Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading”, IEEE, 2006, PP. 2879-2891.xetdcq