LANGUAGE IDENTIFICATION USING VISUAL FEATURES

Mr. Adars U[1]
Research Scholar, Department of ECE, Vinayaka Mission's Research Foundation (Deemed to be University), Salem, TamilNadu & Assistant Professor, Dept. of Computer Science, PMSA PTM Arts & Science College, Kadakkal, Kollam, Kerala

Dr. T. Muthumanickam[2]
Professor & Head of Department of ECE, VMKV Engineering College, Vinayaka Mission's Research Foundation (Deemed to be University), Salem, Tamil Nadu

Abstract:- Lip-analyzing is the system of the use of the visible look of the mouth at some stage in speech to deduce its acoustic content. Visual speech cues are recognised to have an effect on sound belief and are utilized by people and machines to enhance speech intelligibility. This isn't always a trivial assignment however. Since it's miles widely recognized that many speech sounds which can be extraordinary acoustically are indistinguishable at the lips. Language may be portrayed with the aid of using one-of-a-kind modes of communication. Such as writing, talking and gesturing (as in signal language).Language Identification(LID)is the system of ascertaining which language the records is being provided in. Audio language identity is a mature technology, capable of discriminate among tens of spoken language with only some seconds of consultant speech.

1 INTRODUCTION

Lip-analyzing is the system of the use of the visible look of the mouth at some stage in speech to deduce its acoustic content. Visual speech cues are recognised to have an effect on sound belief and are utilized by people and machines to enhance speech intelligibility. This isn't always a trivial assignment however. Since it's miles widely recognized that many speech sounds which can be extraordinary acoustically are indistinguishable at the lips. Language may be portrayed with the aid of using one-of-a-kind modes of communication. Such as writing, talking and gesturing (as in signal language).Language Identification(LID)is the system of ascertaining which language the records is being provided in. Audio language identity is a mature technology, capable of discriminate among tens of spoken language with only some seconds of consultant speech.

Although Visual Language Identification (VLID) is a brand new and unexplored place of studies, it has numerous applications, each for clinical studies and for sensible deployment. The use of visible-simplest lip-analyzing may be taken into consideration as an severe case of audio-visible speech

reputation, wherein the audio channel is absolutely uninformative, because of a low sign to noise ratio, or wherein there may be a disparity among the situations of the education and checking out environments. VLID can help audio-visible speech reputation through enhancing visible reputation capabilities, both in phrases of the extraction process, the capabilities themselves, or how they're implemented to reputation. If the usefulness of the visible sign may be bettered, it can then be included into present audio-visible technology.The purpose of this venture is to expand a device which could discover a language the use of simplest visible speech information.

2 LITERATURE SURVEY

Jacob L Newman and Stephen J Cox, [1] (2013): Automatic visual language identification (VLID) is the technology of using information derived from the visual appearance and movement of the speech articulators to identify the language being spoken, without the use of any audio information. This technique for language identification (LID) is useful in situations in which conventional audio processing is ineffective (very noisy environments), or impossible (no audio signal is available). Research in this field is also beneficial in the related field of automatic lip-reading. This paper introduces several methods for visual language identification (VLID). They are based upon audio LID techniques, which exploit language phonology and phonotactics to discriminate languages

Massaro, D. W. and Cohen, M. M, [2] (1983): Communication between humans is comprised of several channels of information. The primary mode of communication is speech, which is foremost an audio channel. Speech is produced by the movement of speech articulators, most notably the tongue, jaw, lips and vocal chords. The external appearance of the articulators, as seen by a video camera for instance, constitutes a visual mode of speech. The recognition of speech from its visual components is known as lip or speech-reading. Not all of the articulatory parameters are consistently visible from an external view; for example, the teeth and tongue are periodically occluded whenever the lips close. Other articulators are permanently occluded, such as the velum and the vocal chords. The parameters with some degree of visibility are the measurable characteristics of visual speech.

Christiansen, M. H. and Kirby, S, [3] (2003): The term language refers to the broadest system by which humans convey information to each other. Human languages have evolved to be uniquely complex amongst those used by other species in terms of the limitless potential for informational exchange and the rules governing the languages themselves. Languages can be communicated in a variety of ways, including by written text, visually (sign language and body language), or more commonly acoustically, by speech. At its lowest level, spoken language is comprised of sound units, known as phones, which are distinctive. Phones that are considered contrastive within a certain language are phonemes of that language are phonemes in English but not in Japanese). Each language has its own

inventory of phones, although many are often shared. These are the basic building blocks of any spoken language. Phones are combined together to form larger units called words, that are subject to morphological rules determining which phones the morpheme (or root word) should contain for a given context. Phonotactics are a more general set of rules specifying which phone sequences are permissible in a language.

Santhi.S and Raja Sekar, [4] (2013): An automatic Language Identification (LID) is the task of automatically recognizing a language from the given spoken utterance. Language identification is used to identify the language of the particular audio and reduce the complexity of the audio sample. LID systems that rely on multiple language phone recognition language modeling (PRLM) and n-gram language modeling produces the best performance in formal LID evaluations. By contrast, Gaussian mixture model (GMM) systems, which measure acoustic characteristics, are far more computationally efficient but tended to provide inferior levels of performance. We have described here the efficiency of an LID system for two different languages namely English and Hindi. The evaluation of languages is done on the standard recorded databases, from which features are extracted using Mel-frequency cepstral coefficients (MFCC). The language models are done using PRLM and classification is done using Gaussian mixture model (GMM). The obtained results ensure that accuracy of LID is efficient for the chosen languages and the system performance is evaluated on both PRLM and GMM.

Sugiyama, M, [5] (1991): Audio language identification is a mature field of research, with many successful techniques developed to achieve high levels of language discrimination with only a few seconds of test data. These techniques make use of discriminatory features of language as identified by linguists and phonologists. Many of the features are not expressed visually and are therefore not identifiable in the visual-domain, by lip reading. Such methods include measuring the spectral similarity between languages which encompasses the stress and pitch of speech.

Zissman M, [6] (1996): Several approaches exist which exploit the phonetic content differences between languages to achieve language discrimination. Such techniques require the training of a phone recogniser, usually comprising a set of Hidden Markov Models (HMMs), which are used to segment input speech into composite phones. Phonetic transcriptions of training data are required to train a phone recogniser and this may be a prohibitive factor as transcriptions may not be readily available. Furthermore, some LID systems use language-dependent phone recognisers, which introduces a further limitation in that the recognition of additional languages is not a trivial task and is once again reliant upon transcription availability. In an approach called Phone Recognition followed by Language Modelling (PRLM) phonotactics is the feature of language used for discrimination. The contention

here is that diff erent languages have di erent rules regarding the order in which phones may occur in speech. In this technique, a single phone recognition system can be used to tokenise an utterance using a shared phone set, trained using one language. The phone sequences produced by this system can then be analysed in terms of the co-occurrence probabilities of phones in an utterance. Statistical models are built using language specific training data, and these models generate a likelihood score of input utterances being produced by that model. For classification, simple maximum likelihood approaches can be used, or more complex back-end classifiers such as Gaussian Probability Density Functions (GPDF), neural networks .This system can be extended by building PRLM systems using language-specific phone recognisers, and running the recognition systems in parallel.

Mendoza.S., Gillick, L., Ito, Y., Lowe, S., and Newman, M, [7] (1996): The syntax of a language governs its allowable sentence structures and is a feature which di ers between languages. Continuous Automatic Speech Recognition (ASR) systems can make use of higher level language information such as this by using word level language models. Large Vocabulary Continuous Speech Recognition (LVCSR) has been applied to audio identification in the hope that recognising these high level features of language will improve discrimination. Using a set of language dependent ASR systems, one for each language to be recognised, the class of a test utterance can be classified as the ASR system producing the highest likelihood score. A disadvantage of this approach is that training an ASR system requires a large amount of transcribed, language-specific training data, in order to train reliable models which generalise well. Also, there is a significant processing overhead associated with training and testing multiple ASR systems.

Almajai, I. and Milner, B, [8] (2009): The field of computer vision has been applied to speech recognition in order to use the information present in visual speech to improve standard ASR performance. This research area is known as Audio-Visual (AV) speech recognition, and its main application is for speech recognition in noisy environments, where conventional audio features become ine ective, but where visual features are una ected .Approaches for AV ASR use features derived from the mouth region, since it is the area on the face which conveys the most information regarding the speech articulators. Face, mouth and lip tracking approaches have all been employed in the past as methods to locate automatically the areas of the face that contain the most speech information, from which feature extraction can take place. Face detection software usually operates by splitting an image into a pyramid of rectangular regions, the shape, scale and rotation of which are limited to pre-defined parameters. The greyscale intensities within those regions are analysed and classifiers trained on the regions. Once the face has been recognised, smaller facial features such as the mouth can be more easily located by the same technique, using the constraints introduced by the knowledge of the face location.

Broadly speaking, there are two types of features that we can extract from the mouth region for use in visual speech recognition; geometric and appearance-based.

Matthews, I., Cootes, T., Bangham, J., Cox, S., and Harvey, R, [9] (2002): Describes and evaluates two method of visual feature extraction for integration into an audio-visual speech recogniser. Also presented in that paper are the video-only recognition results, which pertain to multi-speaker (test speakers are included in the training set), word-level, isolated letters recognition, using HMMs for speech modelling, and using low resolution grayscale video. They used 520 utterances for training, which equates to two recitals of each letter by each speaker, and they use a further recital for testing (260 utterances).

Sunil S. Morade and B. Suprava Patnaik, [10] (2013): Lip tracking is very crucial for visual lip reading recognition system. This paper presents a novel active contour guided geometrical feature extraction approach for lip reading. Three active contour methods are studied, namely snake, region scalable fitting energy method and loacalised active contour model. These methods are adopted for salient geometrical feature calculation. A joint feature model, obtained by combinatining inner area, height and width has been proposed. Results of experimentations on digit utterances are given to show the improvement achieved by visual speech recognition Systems.

3 EXISTING METHOD

The various existing methods for language identification is described as follows,

3.1AUDIO LANGUAGE IDENTIFICATION

Audio language identification is a mature field of research, with many successful techniques developed to achieve high levels of language discrimination with only a few seconds of test data. The main approaches make use of the phonetic and phonotactic characteristics of languages which are proven to be an identifiable discriminatory feature between languages. In the next sections, we briefly review the techniques used in these approaches.

3.1.1Phone-Based Tokenization:

There are several approaches to LID which exploit the difference in phonetic content between languages to achieve language discrimination. Such techniques require the training of a phone recogniser, usually comprising a set of Hidden Markov models (HMMs), which are used to segment input speech into a sequence of phones. In an approach called Phone Recognition followed by Language Modelling (PRLM), phonotactics is the feature of language used for discrimination. The contention here is that different languages have different rules regarding the syntax of phones, and this can be captured in a language model. In this technique, a single phone recognition system is used to tokenise an utterance

using a shared phone set, trained using one language. The phone sequences produced by this system can then be analysed in terms of the co- occurrence (or ngram) probabilities of phones in an utterance. Statistical models are built using language specific training data, and these models generate a likelihood score of input utterances being produced by that model. For classification, simple maximum likelihood approaches can be used, or more complex backend classifiers such as Gaussian Mixture Models (GMMs), neural networks or Support Vector Machines (SVMs) can be applied. This system can be extended by building PRLM systems using language-specific phone recognisers, and running the recognition systems in parallel (Parallel PRLM =PPRLM).

3.1.2Gaussian Mixture Model Tokenization:

The tokenization sub-system within the LID architecture is usually applied at a phone level. It presents a variant to the standard PPRLM LID approaches which uses sub-phone, frame- level tokenization. In this method, a Gaussian mixture model (GMM) is trained for each language from language-specific acoustic data. Each GMM can be considered to be an acoustic dictionary of sounds, with each mixture component modelling a distinct sound from the training data. Given an MFCC frame, the mixture component is found which produces the highest likelihood score, and the index of that component becomes the token for that frame. For a stream of input frames, a stream of component indices will be produced, on which language modelling followed by back-end classification can be performed, as is common in audio LID .For the NIST 1996 12 language evaluation task report a minimum error rate of 17%, which is higher than standard PRLM techniques. Despite this increase in error rate, several advantages are offered by this approach.

Firstly, the training of the tokeniser does not require transcribed data, which simplifies the incorporation of additional languages into the system and is especially advantageous for VLID where there is no agreed protocol for transcriptions. Secondly, there is a reduction in computational cost using this technique compared with phone recognition. Except the language models themselves, rather than the scores they produce, become the vectors used by the back- end classifier. Instead of a maximum likelihood or Linear Discriminant Analysis (LDA) back- end, SVMs are built from the bigram language models of the tokenised training data, an SVM for each language, each comparing that target language against all other languages.

3.1.3Disadvantages of Existing Systems:

A person‟s voice can be easily recorded and used for unauthorized PC or network.

Low accuracy.

An illness such as a cold can change a person‟s voice, making absolute identification difficult or impossible.

The amount of memory required to store the voice signal is too large.

The correlation between each voice segment should be less.

4 PROPOSED METHOD

Language identity (LID) is the system of classifying a representative sample of speech, text, or extraordinary media, thru manner of approach of the language in which it is spoken or written. The cause of this mission is to expand a tool that can end up privy to amongst languages using fine seen speech information. Given representative schooling records for each of the languages to be discriminated, the linguistic identity of an unseen sample of speech need to be determined. Although seen language identity (VLID) is a modern-day and unexplored vicinity of research, it has several applications, every for scientific research and for realistic deployment.

A preliminary have a have a take a observe on speaker-independent, seen language identity (VLID), in which fine lip shape, appearance and motion are used to determine the language of a spoken utterance. Instead of recognizing phones from acoustic information, we're capable of use the seen appearance of the mouth location to represent speech.

There are modes to be performed to determine VLID system. It consists of training mode and recognition mode.

4.1TRAINING MODE

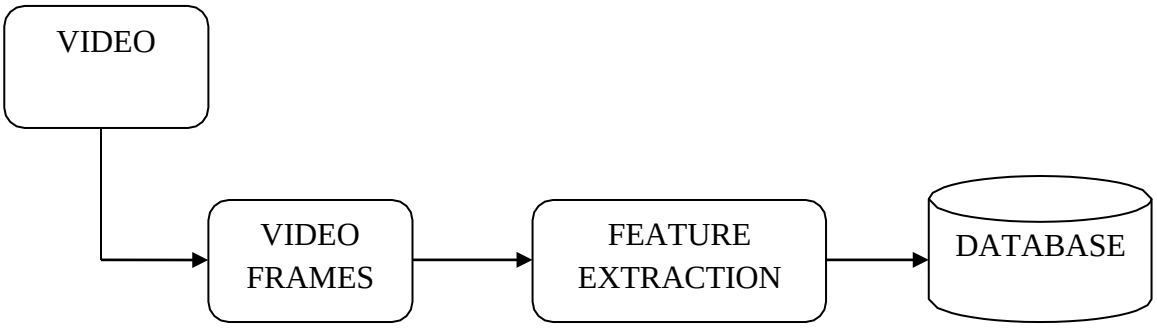


Fig:4.1 Block Diagram of Training Mode

A video record includes frames. These frames while seem earlier than us in a price extra than our notion of vision, offers a sensation of an item transferring earlier than us, through searching simply on the display screen on which frames are performing at excessive price. Thus the usage of MATLAB video is transformed to frames.

Video output turned into break up into some of extraordinary sequences, in which every collection turned into made from some of frames and every repetition of every expression fashioned one collection. Then vital functions are extracted from the frames.

The predominant benefit, in speech recognition, of the usage of visible cues is that they're complementary to the acoustic signal: a few phonemes which might be hard to apprehend acoustically in noisy environments may be simpler to differentiate visually, and vice versa. We distinguish demanding situations in lip functions processing.

• Detection of lips

• Extraction of functions. The first hassle quantities to locating and monitoring a particular facial part (mouth, lips, lip contours etc.)

Successful mouth monitoring remains difficult in instances in which the background, head pose 5and lighting fixtures range greatly. After a success face detection, the location is processed similarly to attain lip functions. Though now no longer very particular in phrases of lip movement description, even the bounding packing containers of lip areas can monitor beneficial lip functions if they're envisioned for each body independently due to the fact such rectangles monitor the dynamic evolvement of the peak and width for the duration of speech production. However, the lip statistics inside the mouth location is maximum normally extracted. The extracted functions are saved in database for the detection purpose, that is the second one mode in VLID process.

4.2 RECOGNISING MODE

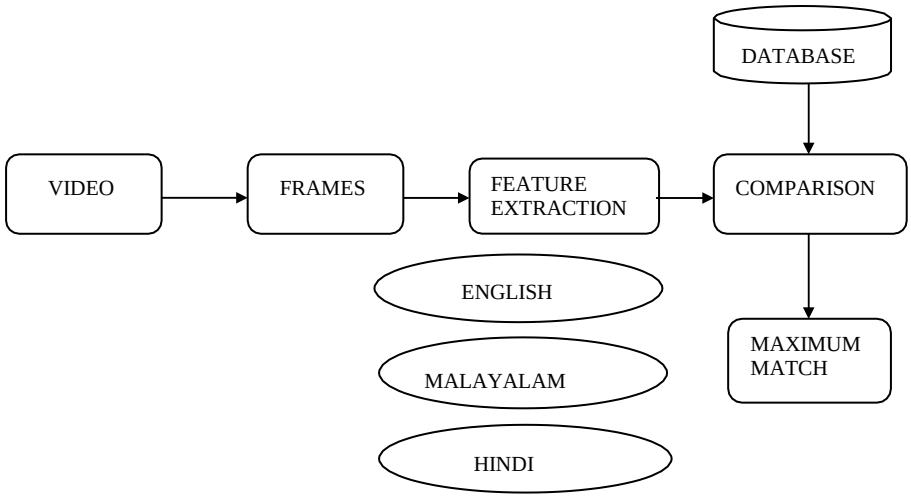


Fig 4.3 Block Diagram of Recognition Mode

The same procedure of training mode takes place, ie, necessary features are extracted from the video, next it is compared with the data base of the training mode from which the most matched result is taken. Thus from the matched result corresponding language (English, Malayalam, Hindi) is obtained. This method avoids the difficulties in identifying the language from audio speech recognition.

5 ALGORITHM

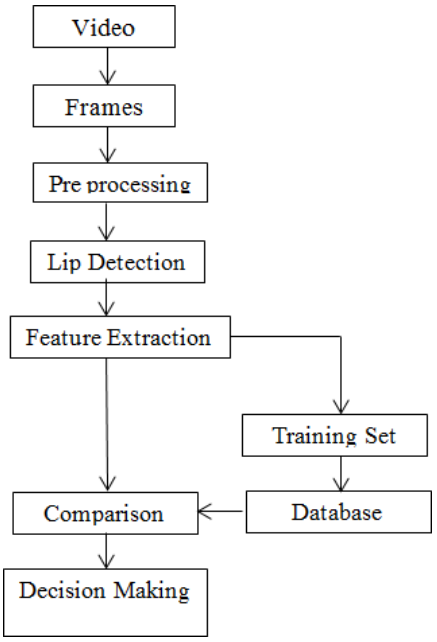
The primary task in both training and recognizing mode is the feature extraction. It aims at extraction of relevant information from the lip region of decent frame set. To extract features, the image has to undergo image preprocessing which include the following steps:

- Image thresholding
- RGB to gray level conversion
- Lip detection and crop the optimal region
- Downsizing cropped image to 64 64 pixel

Four corner points and a center point is detected from the preprocessed image. Using the distance formula, distance of each corner points with respect to center point is computed and stored as vector. This process is done for all images in the datasets. The distance formula is given by,

$$D_x=\sqrt{((x_1-x_2)^2+(y_1-y_2)^2)}$$

These features are then used to compare with the features obtained while recognizing the input unit.



6 DIGITAL IMAGE PROCESSING

Digital Image Processing way processing virtual picture by way of a virtual pc. We also can say that it's miles a use of pc algorithms, with a view to get stronger picture both to extract a few beneficial information.

Image processing specially consist of the subsequent steps:

1.Importing the picture thru picture acquisition equipment;
2.Analysing and manipulating the picture;
3.Output wherein end result may be altered picture or a document that's primarily based totally on analysing that picture.

6.1 PHASES OF IMAGE PROCESSING:

Acquisition:- It may be as easy as being given an picture that's in virtual form. The primary paintings includes: a)Scaling, b)Color conversion(RGB to Gray or vice-versa)

Image Enhancement:-It is among the handiest and maximum attractive in regions of Image Processing it's also used to extract a few hidden information from an picture and is subjective.

Image Restoration: It additionally offers with attractive of an picture however it's miles objective(Restoration is primarily based totally on mathematical or probabilistic version or picture degradation).

Color Image Processing:-It offers with pseudocolor and complete colour picture processing colour fashions are relevant to virtual picture processing.

Wavelets And Multi-Resolution Processing:-It is basis of representing photographs in diverse egrees.

Image Compression:- It includes in growing a few features to carry out this operation. It specially offers with picture length or resolution.

Morphological Processing:- It offers with equipment for extracting picture additives which are beneficial withinside the illustration & description of shape

.

Segmentation Procedure:- It consists of partitioning an picture into its constituent elements or objects. Autonomous segmentation is the maximum hard project in Image Processing.

Representation & Description:- It follows output of segmentation stage, selecting a illustration is simplest the a part of answer for remodeling uncooked information into processed information.

7 ADVANTAGES

- Errors occurs in LID using audio signal perhaps due to the noise or the background sound while the audio signal is being recorded. And also for perfect working of LID microphones are used, which when may be distinctive leads to increase in error. No such error is added in the case of VLID.
- As the computing time is less, processing speed is high.
- It exploit the advantage of high correlation between the consecutive frames.
- Highly economical

8 APPLICATION

- Incorporated with CCTV for investigating the crime.
- Social Medias.
- Subsystem for speech to speech translation.
- In Global Call Centers.

9 FUTURE SCOPE

The number of languages included in the system could be increased to determine how well this approach generalizes when the chance of language confusion is higher. Groups of phonetically similar languages could be added to see if they are more confusable than those with different phonetic characteristics. The features of language that we have used for discrimination is phonology , specifically phonotactics, which governs the allowable sequence of phones in a language. Phonotactics are not the only aspect of language which can be used to differentiate between them. Further work into VLID could therefore focus on incorporatingboth of these additional language cues and evaluating their contribution to language discrimination. Also VLID could be implemented in real time which increases the application side of the same. By developing this system , word recognition can be achieved.

10 CONCLUSION

Biometric recognition is a popular subject in today‟s research & has been shown to be an important tool for identity establishment. Using lip-motion features for dynamic lip image sequences to be using in different recognition systems. The visual lip features are extracted by assuming successful lip contour tracking. Done preliminary study in identification language purely from visual features. Here we recorded multilingual speakers and attempted to discriminate them reading in two different languages.

11 REFERENCE

[1]. Jacob L Newman and Stephen J Cox, (2013). Language Identification Using Visual Features. IEEE Trans. On Audio, Speech and Language Processing.

[2]. Massaro, D. W. and Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. Journal of
Experimental Psychology: Human Perception and Performance, 9(5):753– 771.
[3]. Christiansen, M. H. and Kirby, S. (2003). Language evolution: consensus and controversies. Trends in Cognitive Sciences, 7(7):300–307.
[4]. Benedetto, D., Caglioti, E., and Loreto, V. (2002). Language trees and zipping. Phys. Rev. Lett., 88(4):048702–048705.
[5]. Sujatha, B. and Santhanam, T. (2010). A novel approach integrating geometric and Gabor wavelet approaches to improvise visual lipreading. International Journal of Soft Computing, 5(1):13–18.
[6]. Zissman, M. (1996). Comparison of four approaches to automatic language identification of telephone speech. Speech and Audio Processing, IEEE Transactions on, 4(1):31–44.
[7]. Mendoza, S., Gillick, L., Ito, Y., Lowe, S., and Newman, M. (1996).

Automatic language identification using large vocabulary continuous speech recognition. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings. 1996 IEEE International Conference on, volume 2, pages 785–788.
[8]. Almajai, I. and Milner, B. (2009). Enhancing audio speech using visual speech features. In INTERSPEECH-2009, pages 1959–1962.
[9]. Matthews, I., Cootes, T., Bangham, J., Cox, S., and Harvey, R. (2002). Extraction of visual features for lipreading. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(2):198–213.