ADVANCED IMAGE TO SPEECH CONVERSION

Project report submitted in fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

HARSHIT SAINI (141327) &
ANIMESH PRATAP SINGH (141362)

Under the supervision of

MS. RUHI MAHAJAN

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Certificate

Candidate's Declaration

We hereby declare that the work presented in this report entitled **" IMAGE TO SPEECH CONVERSION"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of our own work carried out over a period from August 2017 to May 2018 under the supervision of Ms.Ruhi Mahajan(**Assistant Professor (Grade-II) , Cse Department**).The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
HARSHIT SAINI (141327)

(Student Signature)
ANIMESH PRATAP SINGH (141362)

This is to certify that the above statement made by the candidate is true to the best of our knowledge.

(Supervisor Signature)
MS. RUHI MAHAJAN
ASSISTANT PROFESSOR
COMPUTER SCIENCE DEPARTMENT
Dated:

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and We are extremely privileged to have got this all along the completion of our project. All that we have done is only due to such supervision and assistance and we would not forget to thank them.

We owe my deep gratitude to our project guide **MS. RUHI MAHAJAN**, who took keen interest on our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

- **TTS – text to speech**
- **NLP- natural language processing**
- **DSP- digital signal processing**
- **ARM- Advanced RISC Machines**
- **RISC- reduced instruction set computer**
- **HMM- Hidden Markov Model**
- **OCR- Optical Character Recognizer**
- **MT- Machine Translation**
- **MAHT- Machine Aided Human Translation**
- **SVM- Support Vector Machine**

LIST OF FIGURES / FLOWCHARTS

ABSTRACT

This project is a motivation to develop an advanced software engine which could scrap textual data from clean and distorted documented images and transfer the corresponding electronic data into speech signals. At the heart of this software engine lies an OCR Engine (Optical Character Recognizer) which inherits crucial morphological operations required for image conditioning & transformation, accompanied with non-parametric machine learning algorithms (typically KNN & CNN) used for character classification. Further the processed textual data is transformed into speech signals using various Text-to-Speech synthesis techniques such as formant and concatenative synthesis. The process of text to speech conversion is based on pipelining of sectoral analytical procedures including textual normalization, phonetic and prosodic analysis.

CHAPTER 1
INTRODUCTION

1.1 Introduction

This project is aimed to adapt a machine learning based architecture which consist of primarily three components:

1.1.1 OPTICAL CHARACTER RECOGNIZER (OCR)

- Optical Character Engine (Optical character recognizer motor, OCR) is the mechanical and electronic change of pictures of composed, manually written or printed content into machine-encoded (content written in machine-lucid shape), regardless of whether from an examined report, a photograph of an archive, a scene-photograph (for instance the content installed on signs and boards in a scene photograph) or from subtitle content superimposed or twixed on a picture (for instance from a transmission or a video podcast).

- It is generally utilized as a type of data passage from printed paper information records, regardless of whether identification archives, solicitations, bank explanations, modernized receipts, business cards, mail, printouts of static-information, or any reasonable documentation. It is a typical strategy for digitizing printed messages with the goal that they can be electronically altered, looked, put away more minimalistically, showed on-line, and utilized as a part of machine procedures, for example, intellectual processing, machine interpretation, (separated) content to-discourse, key information and content mining. OCR is a field of research in design acknowledgment, manmade brainpower, and PC vision.

1.1.2 LANGUAGE TRANSLATOR

- Machine translation, abbreviated as MT (which can be differentiated computer-aided translation(CAT), Interactive Translation(IT), machine-aided human translation (MAHT) or) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

- On a basic level, MT performs simple substitution of words in one language for words in another, but that alone usually cannot produce a good translation of a text because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus statistical, and neural techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.

- Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used. It follows that machine translation of government and legal documents more readily produces usable output than conversation or less standardised text.

1.1.3 SPEECH PROCESSING AND SYNTHESIS

- The role of Digital speech processing is very vital in current speech communication study scenario and all its applications. As we know that the most basic purpose of speech is to communicate with others; in other words we can also say it means conduction of message within the human and machine systems. Text to speech system (TTS) helps in converting the input text into voice using a speech synthesizer . As we all can also call it the artificial production of human speech.

- A speech synthesizer as we know it is a computer system that is majorly used for this specific purpose, and it is also implemented in both the forms such as the hardware as well as the software form like ARM7 microcontroller that helps in the conversion of Text to Speech and Speech to Text. A text-to-speech (TTS) system tends to convert the normal input language text into different English and other accents as per requirement.

- This output is a artificial speech that cannot be understood by a person whose having standard communication skills in English language. Text to speech system processes that produce speech are very much different from the live human speech that is produced. The production of the live human speech mostly depends on complex fluid mechanics which in turn are dependent upon changes in the lung pressure and also the vocal tract constrictions .The primary aim of a text to speech system is to convert any random given text into a similarly corresponding spoken waveform.

- The two main components of a text to speech system are the text processing and the speech generation. To conduct processing and development of the given input text and produce suitable series of phonemic units is the primary objective of the text processing component. The extraction of these phonemic units by the speech generation component is done either by fusion from parameters or by choice of a unit from a large speech corpus. It is absolutely important for the text processing component to ensure and produce an appropriate series of phonemic units that is in accordance with the random input text that is given by the user for the purpose of natural sounding speech.

1.2 Problem Statement

- To implement an optical character recognizer with a reasonaboly high accuracy ,which could identify each & every optical/lingual/mathematical character imprinted on the image or the imagery module under consideration.

- To implement a machine based langauge tranlator which could translate text written in any language to any type of language as per the requirement of the user with correct grammatical sequencing & anaphorical resolution.

- To implement a speech synthesizer.

- Looking at the overall project , it is expected to develop an overall highly accurate image to speech translator for multiple languages & mathematical expressions.

1.3 OBJECTIVES

- To Develop a reasonably accurate Image to Speech converter for English to Spanish & Spanish to English translations specifically.

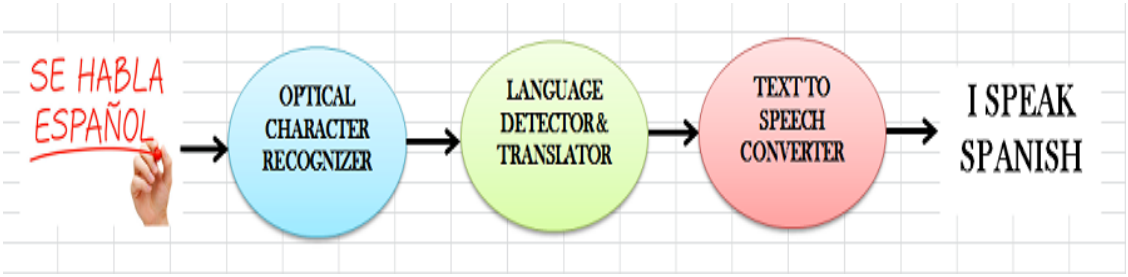


FIG.1

1.4 METHODOLOGY

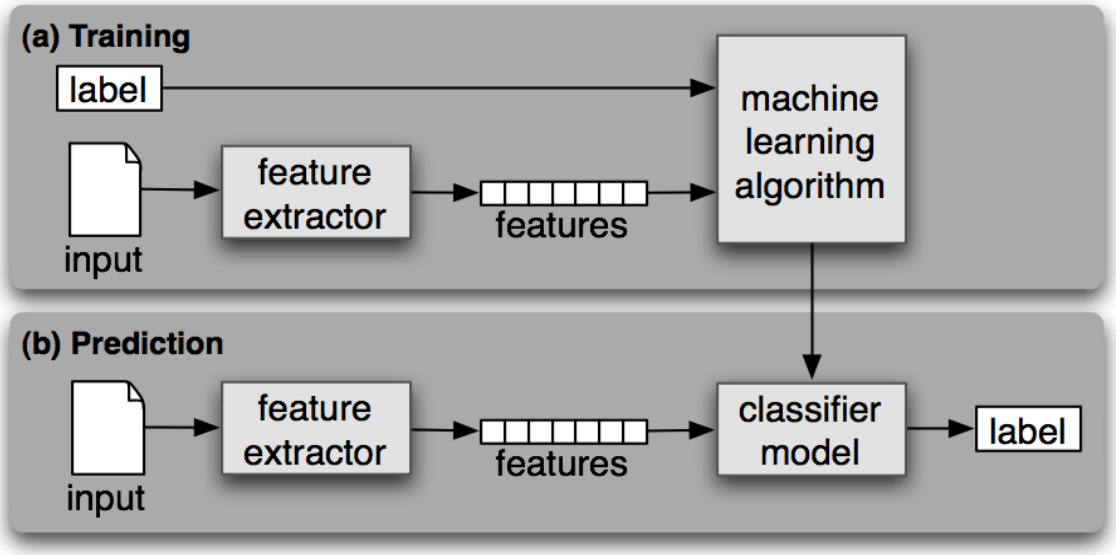


FIG.2

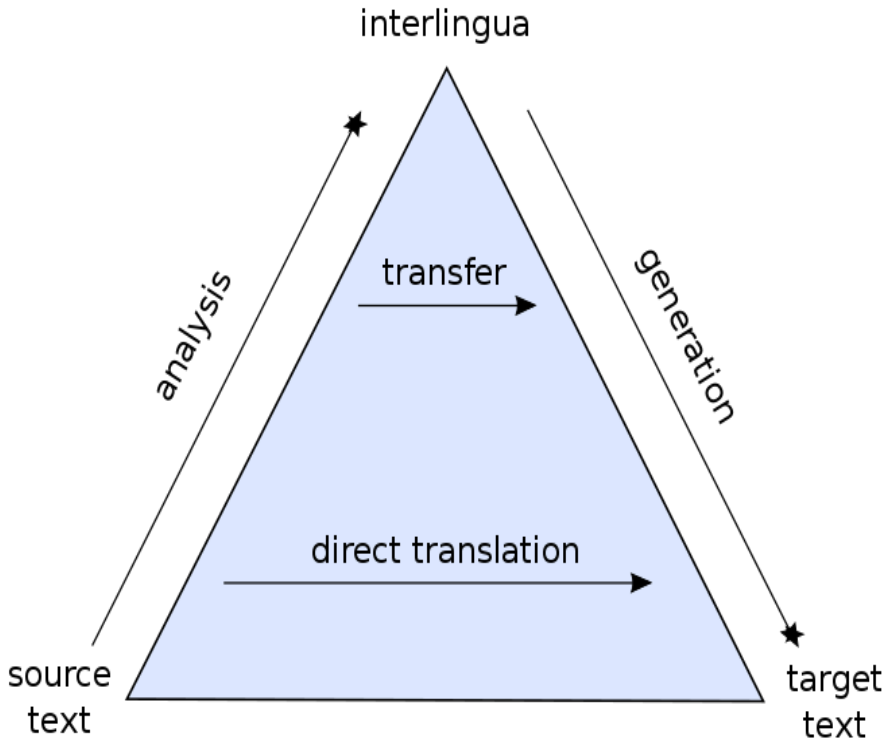


FIG.3

1.4 METHODOLOGY (continued)

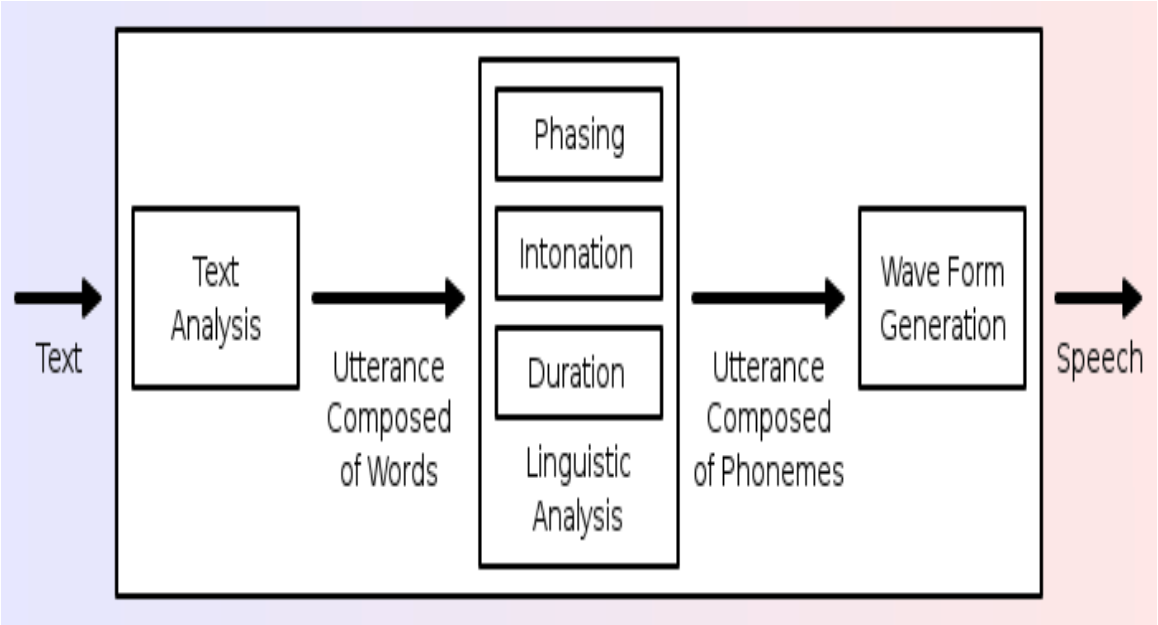


FIG.4

1.5 Organisation

The report is divided into five chapters:

- **Chapter 1- Introduction**
    1. **Intoduction**
    2. **Problem Statement**
    3. **Objectives**
    4. **Methodology**
    5. **Organization**

- **Chapter 2- Literature Survey**

- **Chapter 3- System Developement**
    1. **Stages of OCR algorithm**
    2. **TTS's System main phases**

- **Chapter 4- Performance Analysis**

- **Chapter 5- Conclusions**
    1. **Conclusion**
    2. **Possibilities for improvement**
    3. **Future Scopes**

CHAPTER 2
LITERATURE SURVEY

2.1 Review on Text-To-Speech Synthesizer ,8 AUGUST 2015 , Suhas R. Mache, Manasi R. Baheti ,C. Namrata Mahender

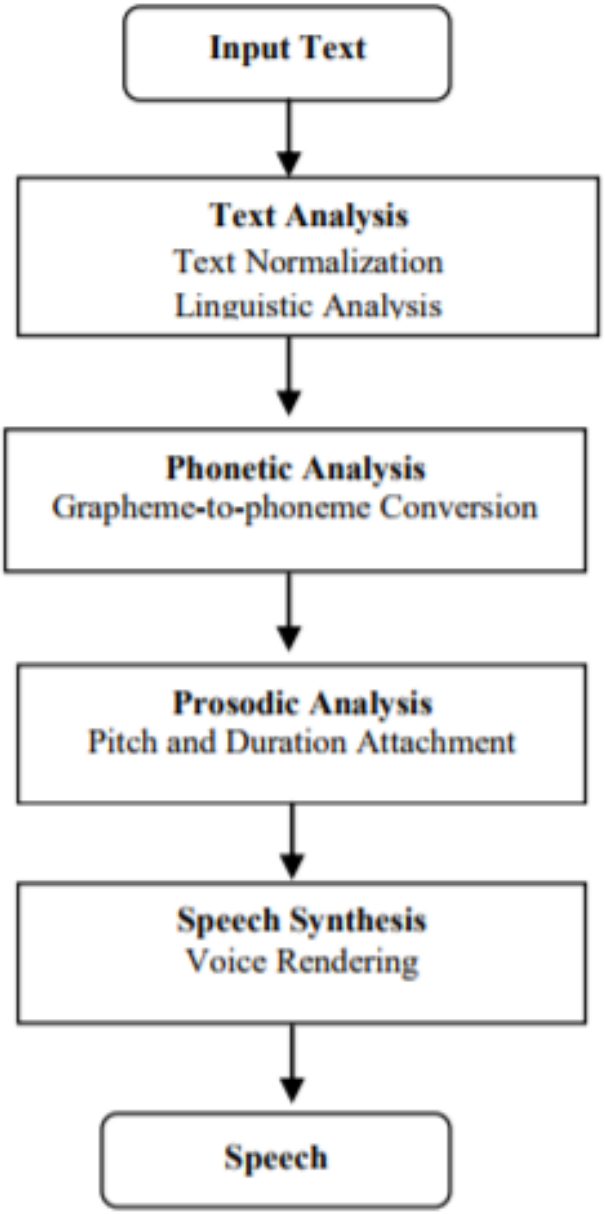


FIG.5 Block diagram for text to speech synthesis

2.1.1 TYPES OF SPEECH SYNTHESIS

The two of the min technologies that help in the generation of the waveforms of synthetic speech are the formant and the concatenative synthesis. Both of the technologies that have been used have their own strengths and weaknesses, and the planned uses of a synthesis system will characteristically conclude which approach is most viable to be used.

1. CONCATENATIVE SYNTHESIS

Concatenative synthesis is the synthesis that is based on the concatenation or we can also call it stringing together of segments of recorded speech. In any general case, concatenative synthesis helps in producing the most natural-sounding synthesized speech. However, in many cases we come across the differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms which in turn results in audible glitches in the output. There are three main sub-types of concatenative synthesis.

i. Unit Selection Synthesis
ii. Diphone Synthesis
iii. Domain Specific Synthesis

2. FORMANT SYNTHESIS

This method of synthesis is seldom called the synthesis based on rules. Although there are many concatenative systems which also include components based on rules. There are many systems that are based on the technology of formant synthesis which generate non-natural, robotic-sounding speech that would never in a million years be mistaken for the natural human speech. However, no one ever sets true naturalness as the goal of a speech synthesis system, and formant synthesis systems have in most cases an edge over concatenative systems. The speech that is synthesized using formant synthesis can be reliably comprehensible, even at extremely high speeds, they always tend to acoid the acoustic glitches that frequently infect the concatenative systems. Visually challenged people can easily use the high-speed synthesized speech to swiftly find the way through computers using a screen reader. Formant synthesizers are mostly the small programs than that of concatenative systems because they do not include with them large databases of speech samples. Therefore it is easy to use them in embedded systems where memory and microprocessor power are especially limited.

i. Articulatory Synthesis
ii. HMM Based Synthesis
iii. Sinewave Synthesis

CHAPTER 3
SYSTEM DEVELOPMENT

3.1 STAGES OF GENERIC OCR ALGORITHM

3.1.1 Extraction of character region form an image

- Extraction of character locales from a picture depends on utilizing auxiliary data thought about the picture to choose a picture property (or properties) that is (are) suciently dierent for the content locales and the foundation areas to be a premise of isolating one from the other.

3.1.2 Segmentation of image into text and background

- While some OCR calculations work with dim scale pictures, numerous change over their contribution to double pictures amid the beginning times of preparing. In spite of the fact that this is the situation most attempt to remove non-printed areas, for example, designs before they portion content from the foundation. Given that we have picture locales that contain content, regardless of whether single word districts or entire pieces of content, the objective of this stage is to distinguish picture pixels that have a place with content and those that have a place with the foundation.

3.1.3 Conditioning of the image

- Whatever procedures are utilized to separate content from its experience it is inescapable that the resultant picture sections will contain a few pixels identied as having a place with the wrong gathering. Molding the picture alludes to the strategies used to 'tidy up' the picture, to erase commotion.

- Morphological administrators and neighborhood tasks are the most famous for distinguishing clamor and erasing it. For the most part secluded pixels are effortlessly evacuated while districts of commotion adjoining content or foundation are more dicult to recognize and consequently expel.

3.1.4 Segmentation of characters

- OCR calculations divided their info picture into districts that contained individual characters. While this is regularly feasible for machine-printed content, it is more dicult for written by hand material and is maybe the significant errand experienced in perusing cursive contents. Numerous more current calculations, notwithstanding for machine-printed content, maintain a strategic distance from this stage and move into character acknowledgment without earlier character division. Actually character acknowledgment happens before character division.

- Morphology activities, Connected Component Analysis, and vertical projections are utilized to section characters. The majority of the strategies used to extricate character districts from a picture can be utilized to portion characters. Be that as it may, most calculations that utilize this stage expect that some joined characters will be portioned as one character and a few characters will be divided into in excess of one piece. Later phases of preparing may need to endeavor to part a district or go along with at least one to frame a solitary character.

3.2 Standardization of character measure

- After the picture is portioned into characters, it is regular to modify the measure of the character locale with the goal that the accompanying stages can accept a 'standard' character estimate. Obviously, characters that have a similar tallness fluctuate in width, so the angle proportion of the character locale is vital in the standardization of character estimate. Note that size standardization is typically just required when the strategies in the accompanying stage rely upon character estimate. Some character highlights, for example, topological ones are autonomous of size and therefore measure standardization isn't a pre-imperative.

- On the off chance that the span of the character area is N pushes by M sections and the 'standard' size is P pushes by Q segments at that point estimate standardization can be accomplished by shaping an extended character district of PN pushes by QM segments by pixel replication and afterward by examining the extended character locale to acquire a P pushes by Q segments standardized character district. Note that Q may need to change for each character locale with the goal that the angle proportion of the character area P by Q is the same as the perspective proportion of the N by M character district.

3.3 Feature Engineering

- The stages going before Feature Detection are regularly portrayed as preparatory handling. Highlight identification and classication are the core of OCR. Over the historical backdrop of OCR numerous dierent include recognition systems have been utilized. At first format coordinating was utilized to nd the entire character as an element, while later, subfeatures of the character were looked for. Calculation found the limit traces, the character skeleton or average hub, the Fourier or Wavelet coecients of the spatial example, different minutes both spatial and dark level, and topological properties, for example, the quantity of openings in an example. All have been utilized as highlights for classication of the character areas.

- In choosing character highlights to be identified, calculation fashioners were aware of the need to distinguish highlights that described the characters being perused autonomously of real textual style and size. What's more, they were pulled in to highlights that were invariant to picture varieties, for example, interpretation, turn, complexity, and shading and additionally being illustrative of the characters when parts might miss or have embellishments (especially for written by hand message).

3.4 Classification

- The part of classication is to allocate to a character district the character whose properties best match the properties put away in the element vector of the area. At first, OCR classiers had a tendency to be auxiliary classiers, that is, the originator devises an arrangement of tests in view of whether specific highlights exist in specic positions inside the character locale. For instance, one may test for a sharp corner in the lower left of a character district as a method for recognizing a 'B' and a '8'. The tests utilized depended on the creator's comprehension of character development and were a component of his preparation. The auxiliary approach has more accomplishment with machine printed content than it does with transcribed content where definite spatial highlights are less normal for the content than in machine printed shape.

- Later classiers have taken a factual instead of basic way to deal with character acknowledgment. The creator utilizes an arrangement of preparing illustrations, that is, an arrangement of character areas where the character exhibit in the district is known, and afterward utilizes factual systems to manufacture a more tasteful that matches the component vector of a yet to be identied character to the element vectors in the preparation set. The more tasteful doles out the character whose preparation vectors best match the component vector of the obscure character. This approach can be typified in Bayes Decision Rule strategies, Nearest Neighbor Matching methods, Decision Tree approaches, and in Neural Networks. Fundamentally all utilization the preparation set to manufacture an arrangement of choice decides that are then used to order.

- In an innovation like neural systems the preparation set decide the weights that encode the choice standards of the system. Late factual classiers have accomplished magnificent acknowledgment rates. Be that as it may, while the acknowledgment rates are in the high 90% territory, most have not accomplished rates more than 99%. For the most part, this is on account of a solitary more tasteful can be prepared to get close ideal outcome in constrained conditions, say with a solitary textual style, yet they lose their close impeccable conduct when they are prepared with an extensive variety of characters. Subsequently, present day classiers are in truth combinations of classiers combined with a component to restore the best classication from the best more tasteful (or set of classiers).

3.5 Verification

- The last stage in OCR is verication. In this stage, learning about the normal outcome is utilized to check if the perceived content is steady with what is normal. Such verication may incorporate conrming that perceived words are without a doubt words found in a lexicon, or that vehicle tags are recorded in the database of issued plates.

- Verication can't ensure that the perceived content is right however regularly it can distinguish mistakes. Blunders can be taken care of by re-trying the OCR yet this time getting the second best classication, and rehashing the verication. On the other hand, the more tasteful may give a requested rundown of conceivable classications when the sureness of right classication is beneath some limit.

- The verfier can utilize its data to decide the nal classication. Verication, while a critical stage, is particularly application subordinate and regularly actualized autonomously of the OCR motor.

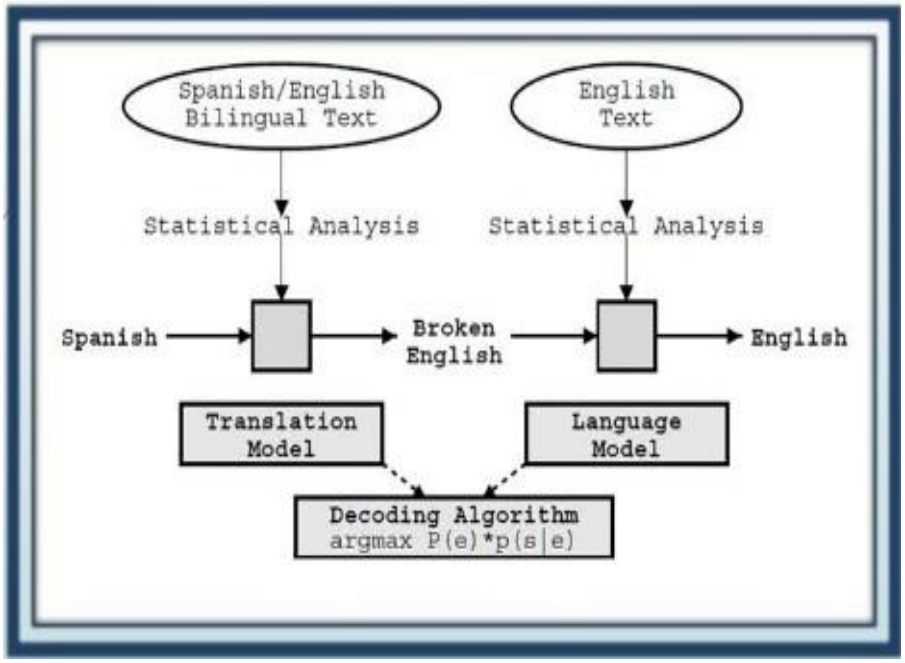


FIG.6

3.6 TTS SYSTEM'S MAIN PHASES

1. Text processing -

- A text-to-speech system the input text is first analyzed, normalized and transcribed into a phonetic or some other linguistic representation. Text processing components deals with low level processing issues such as sentence segmentation and word segmentation

- Document structure detection – It is an easy task to detect the document structure by interpreting, how the paragraphs are formatting and the punctuation marks as well.

- Text normalization – The task for the text normalization is to handle all the abbreviation and the acronyms in the text. The work that normalization does is to compare the given text e.g. the acronym Dr. is taken as doctor. Proper normalization always helps in making an output look good.

- Linguistic analysis - It takes into account a structural breakdown for the proper pronunciation of the words and syntactic analysis to smooth the progress of accenting and choice of words to handle ambiguity in written text input.

2. Speech generation -

- The basic work of the speech generation component is to process and generate the speech by making the use of different parameters as Phonetic analysis.

- It majorly focuses on the phone level that is there within each word of the input text. As we know each of the available phone is setup with the information about what and how the sound is to be produced which means the style and the emphasis of a particular language.

- Grapheme to phoneme conversion: It basically deals with the exactly correct pronunciation of each word of the sentences that have been input by the user.

- Homograph disambiguation: It figures out whether the input sentence is in a past tense or a future tense or a present tense; however its on the dictionary to identify the word tense system.

- Prosodic analysis – As we need to have some basics for the marking of prosodic effect and the utterance plans hence the analysis of the prosody is quite important i.e. phonological prosodic processing and later to arrive at suitable rendering strategies for the marked prosody i.e. phonetic prosodic processing. We can have two different approaches towards the prosody.

- We can either create an conceptual expressive system which characterizes explanation of the performance of the parameters of prosody within the acoustic signal (elementary frequency movement, intensity changes and duration movement) and promote the system to a symbolic phonological role.

- Or we can create a phonological system which can be easily used to input the process, which in turn result in an acoustic signal jugged by listeners to have a proper prosody.

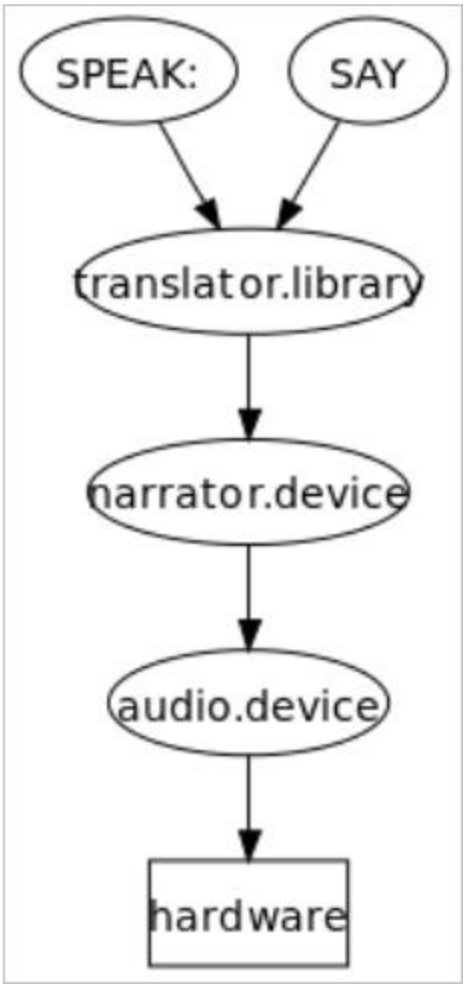


FIG.7 Basic algorithm for text to speech translation

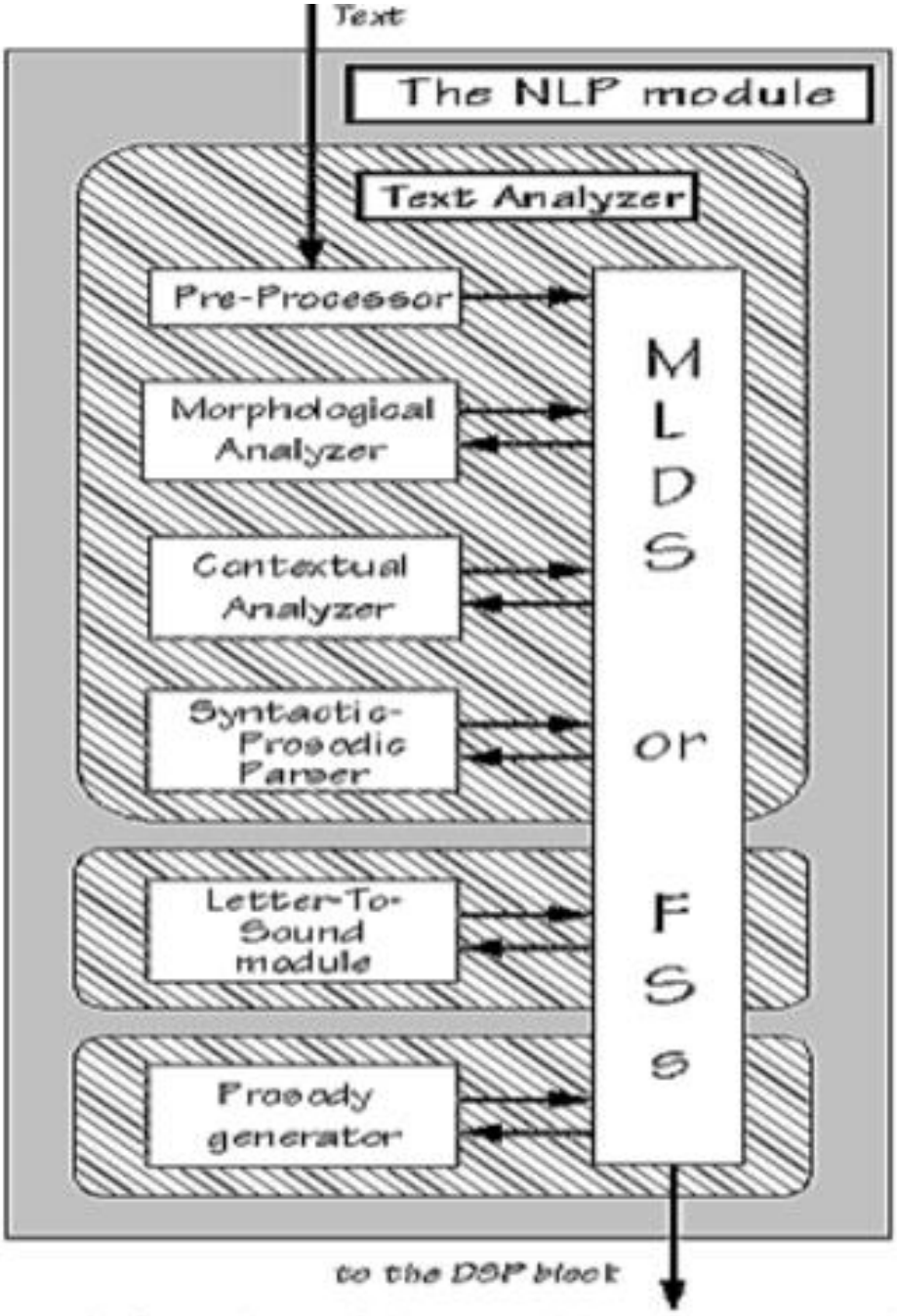


FIG.8 Operations of the natural language processing module of a tts synthesizer

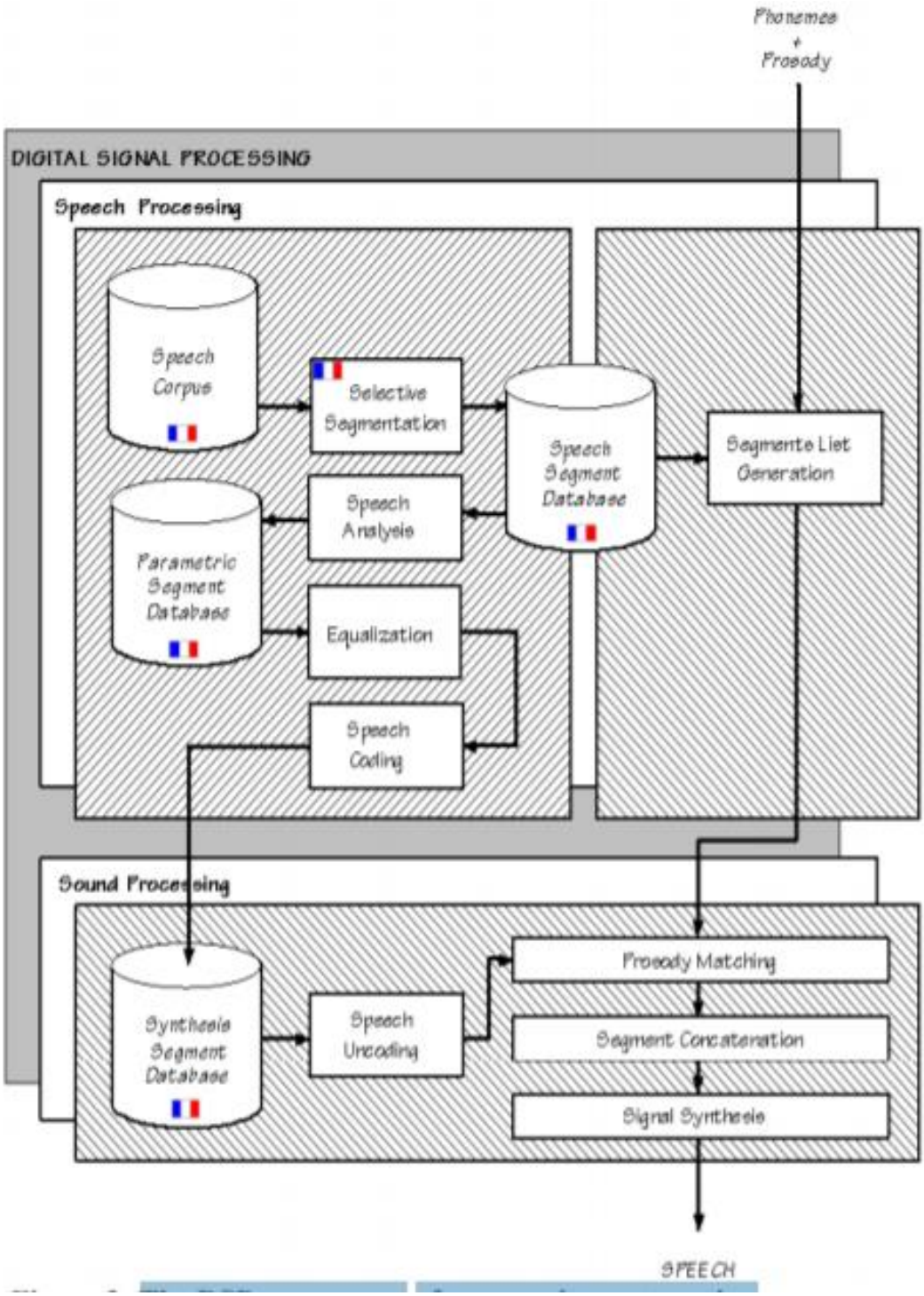


FIG.9 The dsp component of a general concatenation based synthesizer

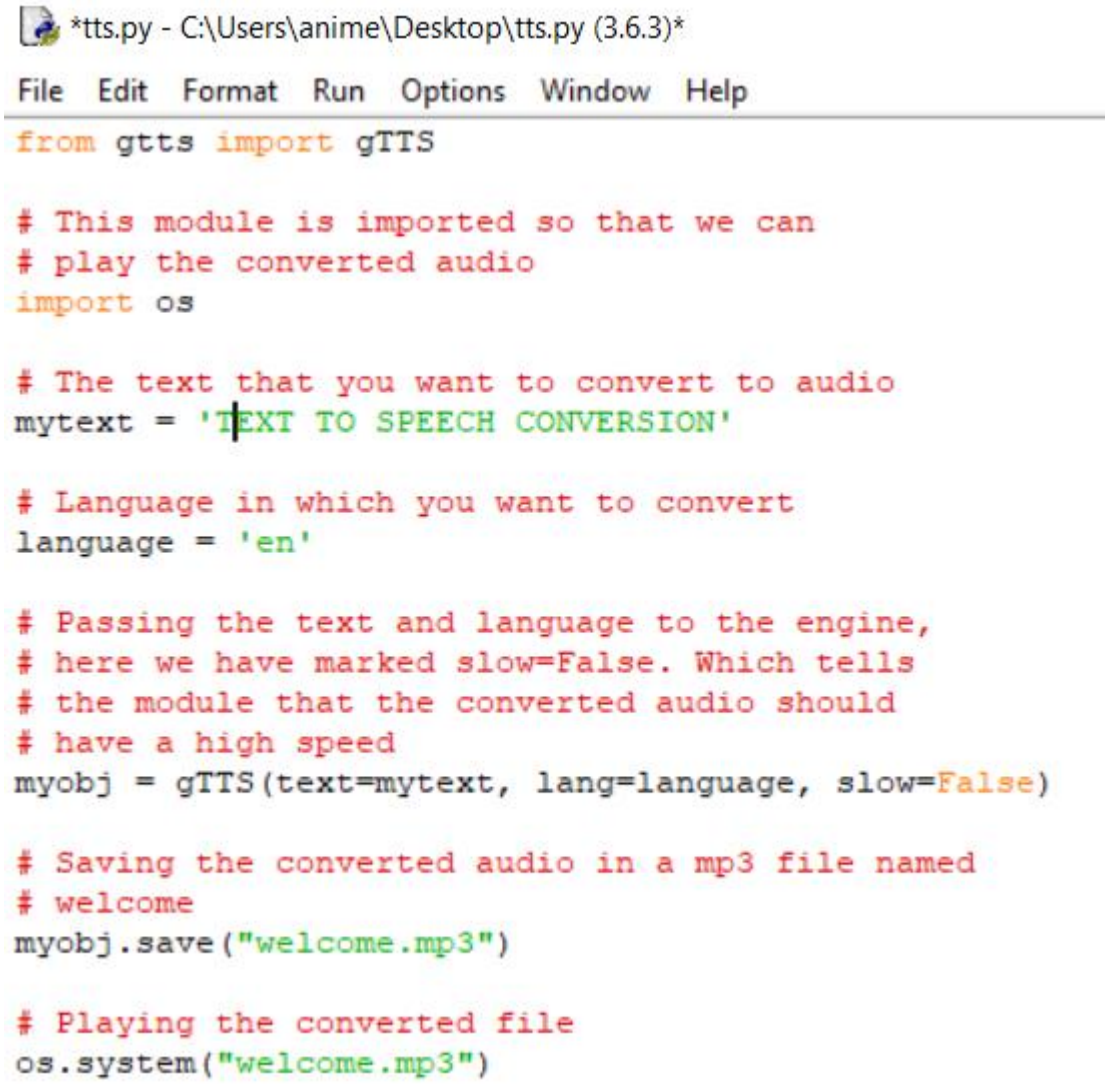


FIG.10

- Achieved the normal Text To Speech conversion using the above python's gTTs library and got the desired output to whatever text mentioned in the specified landuage.

CHAPTER 4
PERFORMANCE ANALYSIS

4.1 Classiers Used:

- Four classiers that were used in this research will be discussed as examples:

- Naïve Bayes, SMO a classier that implements John Platt's sequential minimal optimization for training a support vector classier and a classier generating a pruned C4.5 decision tree .

- **NaïveBayes** classier is based on Bayes' rule of conditional probability and naïvely assumes independence.The suspicion that traits are free (given the class) may sound shortsighted be that as it may, as has appeared, the quantity of conveyances for which the extra punishment term is extensive goes down exponentially quick (that is, all dispersions are near the dissemination accepting restrictive Independence).

- **ComplementNaïveBayes** is a variation of the multi-nominal naïve Bayes classier, optimized to work with text.

- **SMO** trains a support vector classier which uses linear models to implement nonlinear class boundaries by transforming the input using a nonlinear mapping (transforming the instance space into a new space).

- **C4.5** creates a decision tree using a divide-and-conquer algorithm. The tree is pruned to remove unnecessary subtrees, make the searches faster and avoid overtting. The complexity of training a decision tree of this type is $O(mn\log n) + O(n\ (\log n)2)$.

4.2 Classier Evaluation:

- To assess the execution of a more tasteful, a few measures should be set. Each more tasteful should be prepared on one dataset and tried on another (called, individually, the preparation set and the test set). Due to the restricted sizes of datasets in inquire about, a standout amongst the most normally utilized techniques n keeps running of k-crease crossvalidation, which partitions the dataset into k arbitrary bits of a similar size, at that point takes each piece to be the preparation set and tests on the rest of the k − 1 pieces.This process is repeated *n*times to obtain an average result.A number of measures are used to evaluate the success of a classier :
- Precision is he proportion of correctly classied instances to all the instances classier:

$$\boldsymbol{precision = |R \cap D| / |D|}$$

- where R is the set of correctly classied instances, while D is the complete set of classied instances. Recall is the proportion of correctly classied instances, out of all classied instances:

$$\boldsymbol{recall = |R \cap D| / |R|}$$

- where R and D are again the set of correctly classied instances and the complete of instances classied. The F-measure is the weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

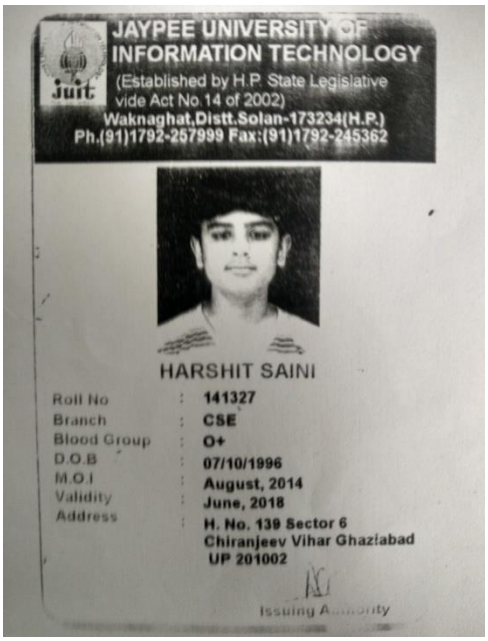
$$\boldsymbol{F = 2 \cdot precision \cdot recall / precision + recall}$$

- This is also known as the *F*1 measure, because recall and precision are evenly weighted. The general formula for non-negative real α is:

$$\boldsymbol{F = (1 + \alpha) \cdot precision \cdot recall / \alpha \cdot precision + recall}$$

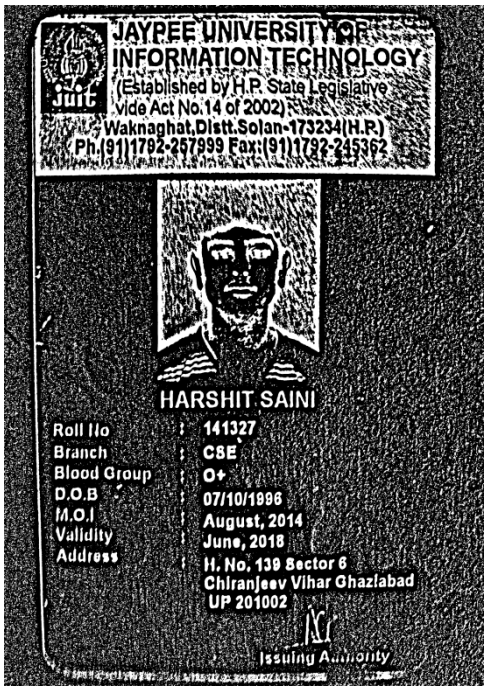
- Two other generally utilized F measures are the F2 measure, which weights review twice as much as exactness, and the F0.5 measure, which weights accuracy twice as much as review. Another measure that can be watched is the measure of connection between's consequences of the classication and the dataset, called the κ−statistic. A κ−statstic esteem more than 0.7 demonstrates a measurably signicant connection. At long last, we can see the consequences of the classication as a perplexity network which demonstrates to every one of us conceivable classes for the classication and how the occasions of each class were classied.

RESULTS AND ANALYSIS



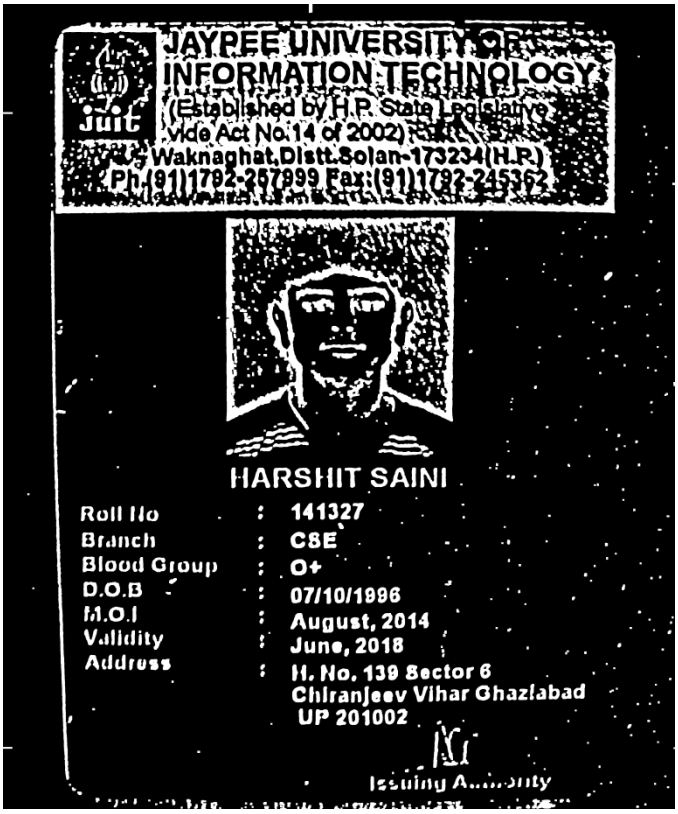
ORIGINAL IMAGE

FIG.11



THRESHOLDED IMAGE

FIG.12



OPENED IMAGE

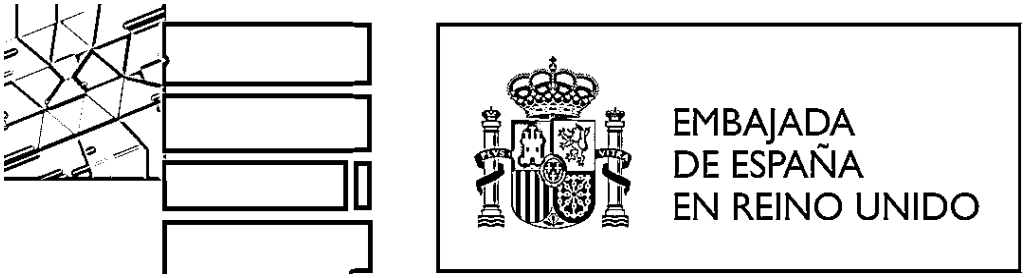
FIG.13

PRECISION -: 0.5471698

RECALL-: 0.4142857

F-MEASURE:- 0.4715447

ACCURACY-: 58.795562%



ORIGINAL IMAGE

FIG.14

THRESHOLDED IMAGE

FIG.15

PRECISION -: 0.706896

RECALL-: 0.718298

F-MEASURE:- 0.713043

ACCURACY-: 89.96%

ORIGINAL IMAGE

FIG.16

THRESHOLDED IMAGE

FIG.17

PRECISION -: 0.898305

RECALL-: 0.697368

F-MEASURE:- 0.785185

ACCURACY-: 91.39465%

4.3 Terminology Used

- To stay away from perplexity and long clarifications in the investigation of the outcomes and mistakes, extremely imperative wording utilized as a part of this part segments will be clarified in this short glossary.

- Because of the idea of the pictures utilized, a scanty vector was picked as the best portrayal. Inadequate vectors will be vectors where, rather than retaining each component of the vector, we just remember the non-zero ones and their positions, sparing both space and time.

- An eigenvector of a change is a non-invalid vector whose bearing is unaltered by that change. The factor by which the greatness is scaled is known as the eigenvalue of that vector Structure tensors are a framework portrayal of fractional subsidiary data.

- In the eld of picture handling and PC vision, it is regularly used to speak to the slope or edge data. It additionally has an all the more intense depiction of nearby examples rather than the directional subsidiary through its intelligence measure.

- Formulated structure tensor matrix can be represented as:

$$S = \begin{bmatrix} I_x^2 & I_X I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$

FIG.18

An Eigen based decomposition can be applied to the presented structure tensor matrix inorder to synthesize the eigenvalues and eigenvectors. A precise description of the local gradient characteristics can be formed using new delta features synthesized.

- The Hessian matrix represents delta of a gradient matrix, which also is an offset of a scalar-valued function. If we are presented with a real-valued function then assuming that all of its partial second derivatives of exist, then the Hessian matrix of the corresponding real-valued function will be the matrix $H(f)ij(x)$ which is exactly equal to $DiDjf(x)$, where x takes up value in the set $(x1, x2, \ldots, xn)$.

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

FIG.19 Hessian Matrix which represents delta of a gradient matrix

4.4 Datasets

- The dataset used contains 42000 images with dimensions of 28x28 pixels. In case of KNN and SVM classifications, the images were set to binary to remove noise for final classification, while in case of DCNN , the convolution and max-pooling layers in the neural network were self-sufficient to capture the important structural features for every character to be classified.

- Data was tried for reduction with use of flattening and application of Principal Component Analysis (PCA) for reducing features by significant amount and pertaining important explained variances so that computation complexity could be reduced significantly in order to reduce overall runtime of the training and classification program.

- Images obtained for classification purpose were briefly a result of segmentation procedure which was carried with help of Open CV library of python. The segmentation procedure included mainly 3 steps for text block identification, text line segmentation and word-character segmentation.

- For achieving the purpose of segmentation, structural & morphological operations were used as a tool to process and condense whole documented images into words and characters.

- For segmentation of different text blocks, morphological dilation was used to bleed out lines into blocks of text. For each block, lines were subjected to significant iterations of morphological horizontal dilation which led to discretion between different lines present in a single block. Same operations are applicable to the lines of text to extract out words and characters for further segmentation and preparing them for application of the pipelined machine learning algorithm.

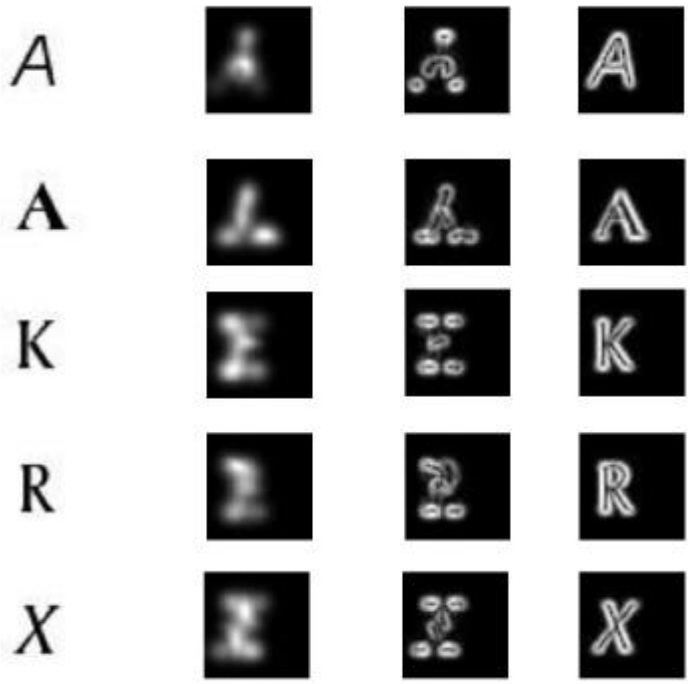


FIG.20

First Column refers to the text based on ground truth
2nd column refers to minimal eigenvalues of structure tensors
3rd column refers to minimal eigenvalues of Hessian Matrix
4th column refers to result of morphological thinning & skeletonization

4.5 Results and Errors

For evaluation of our main algorithms, we used 4 basic performance metric which includes acccuracy, precision , recall and F-measure. Looking at accuracy and recall, for the given dataset, accuracy comes out to be around ~80% while the recall ratio is quite high ~97% in comparison to the accuracy measure. This is due to presence of noise which wasn't successfully removed during image conditioning.Thus, presence of this noise led us to Type II error relating to False negatives which can only be encountered by using appropriate denoising or image segmentation techniques.

4.6 Comparison of Algorithms

The SVM classifier gave results for classification with 93.84% accuracy. On the other hand , application of KNN after considerate hyperparameter tuning and cross-validation, the accuracy comes out to be 96.35. At the end, the classifier which worked best comes out to be Deep Convolution Neual Network which gave aroun d 98.1% accuracy. After embedding data augmentation and dropout regularization, DCNN produces classifications with 99.61% accuracy.

CHAPTER 5

CONCLUSIONS

5.1 Conclusions

- For experimentation and training purpose, a dataset with 42000 images were used. Partitions were created according to the requirement of cross validations carried out. Dataset is generally divided into 3 partitions for training , testing and validation purposes.
- The datasets were classed using various classification algorithms namely KNN(K-Nearest Neighbours), SVM(Support Vector Machine), and CNN(Convolution Neural Network). If we take MNIST handwritten character data into consideration, KNN algorithm with use of Genetic algorithm beats Deep Convoluion Neural Network with an accuracy of ~99.7.
- Considering all algorithms, Naive Bayes gave the worst performance on original dataset, while SVM seems to perform better than Naive Bayes with promising results.SVM classier uses the concept of Sequential Minimal Optimizer(SMO) for which every classification has a high correlation with the related dataset. Drawback of SVM classifier was that it consumes a lot of time to train the SMO model.
- For exrtremely large datasets , Deep Convolution Network works best for classification. The only drawback associated with the use of DCNN is that it takes a lot of time to train and update weights and bias informations over its layers. This propogation procedure can be terminated using limited number of epochs or by applying a good optimizer with reduce on plateau condition for learning rate reduction
- Most noticible errors which occured with use most of the classifiers were induced due to overfitting. For reduction of these type of errors, data augmentation and regularization techniques were used. In case of CNN classifier, new mutated image data was generated for training using image transformation techniques like decentering, skewing and zooming. For regularization and generalization, dropout regularization induced after pooled layer to allow layers to learn some new information which might be helpful in reducing overfitting.
- Text-to-Speech synthesizer has continuously been worked upon and developed a great deal to take its current shape over the past decade. Speech synthesis is basically based on three forms of synthesis which are Articulator, Formant and Concatenative synthesis used in various synthesizers. Each day many new applications are developed all over the world, but simplicity and unambiguousness of synthetic speech has not reached the level that it should have been over the years.

5.1 POSSIBILITIES FOR IMPROVEMENT AND FUTURE PROJECTS

- The feature engineering process used was trivial and straight-forward. Use of more complex feature synthesis might either lead to increase in accuracy but might also lead to signinficant increase in computational complexity. Thus, more complex feature model must be avoided for adaptation as it might also lead to overfitting as increasing model complexity can decrease biasness but doesn't guarantee to decrease overall generalized error.

- It is very much interpretable that classification accuracy is dealt to increase with increase in the size of the dataset used. Looking at the size of the overall dataset provided , it is important to choose appropriate sizes of train , test and validation partitions in order to perform required cross validations inorder to leviate overall generalized error.

- By upraising the count of algorithms in the hybrid model, adding just a single layer or a hyperparameter has the capability to genreate completeness in the results. Though standalone models are highly capable of dedcuing significant results, hybrid models are preferred with complex algorithms , accompanied with data augmentations techniques to generate more generalized and complete results over new prediction set.

5.3 FUTURE SCOPES

- The Text to speech synthesis is an aspect of computer technology that is growing at a very great pace and it is playing a very crucial role in the way that the user interacts with the virtual system and interfaces across a variety of platforms. Identification of various operations and processes that are related to text to speech synthesis have been made.

- With the use of the code and the library mentioned above in python we have achieved the normal text to speech conversion in the specified language. In future there is a huge scope in this field. The user can train a dataset for a different voice using speech recognition and use it as voice for output of the speech that we receive for the input text that we give.

REFERENCES

[1] Archana Balyan, S.S. Agrwal and Amita Dev, Speech Synthesis: Review, IJERT, ISSN 2278-0181 Vol. 2 (2013) p. 57 – 75.

[2] D.D. Pande, M. Praveen Kumar, A Smart Device for People with Disabilities using ARM7, IJERT, ISSN 2278-0181 Vol.3(2014) p. 614 – 618.

[3] J.O. Onaolap, F.E. Idachaba, J. Badejo, T. Odu and O.I. Adu, in Proc. of the World Congress on Engineering, (London, UK. 2014).

[4] Alistair Conkie, Thomas Okken, Yeon-Jun Kim, Giuseppe Di Fabbrizio, Building Text-To-Speech Voices in the Cloud, in Proc. AT&T Labs Research, Park Avenue, Florham Park, NJ- USA).

[5] Mark Tatham and Katherine Morton, Developments in Speech Synthesis (John Wiley & Sons, Ltd. ISBN: 0-470-85538-X, 2005).

[6] A. Indumati and Dr. E. Chandra, Speech processing –An Overview, Int. J. of Engg. Sci. and Tech., Vol. 4, (2012) p. 2853-2860.

[7] Mattingly I. G., Speech Synthesis for Phonetic and Phonological Models, T.A. Sebeok (Ed.) Current Trends in Linguistics, Vol. 12, (1974) p. 2451-2487

[8] Klatt Dennis, Review of Text-to-Speech Conversion for English, J. of the Acoustical Soc. of America, Vol. 3, (1987) p. 737-793.

[9] Schroeder M., A Brief History of Synthetic Speech, J. Speech Communication, Vol. 13, (1993) p. 231-237.

[10] Allen, John, Hunnicutt, Sharon, and Dennis Klatt, Text To Speech, The MITTALK System (Cambridge: Cambridge University Press, 1987).

[11] A. S. Sawant, "Script Independent Text Pre-processing and Segmentation for OCR," *Int. Conf. Electr. Electron. Signals, Commun. Optim. - 2015*, pp. 1–5, 2015.

[12] V. Kieu, F. Cloppet, and N. Vincent, "OCR Accuracy Prediction Method Based on Blur Estimation," *2016 12th IAPR Work. Doc. Anal. Syst.*, pp. 317–322, 2016.

[13] J. B. Pedersen, K. Nasrollahi, and T. B. Moeslund, "Quality Inspection of Printed Texts," *IWSSP 2016- 23rd Int. Conf. Syst. Image Process. 23-25 May 2016, Bratislava, Slovakia*, pp. 6–9, 2016.

[14] A. F. Mollah, N. Majumder, S. Basu, and M. Nasipuri, "Design of an Optical Character Recognition System for Camera- based Handheld Devices," *IJCSI*, vol. 8, no. 4, pp. 283–289, 2011.

[15] B. Jain and M. Borah, "A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical," *IJSRP*, vol. 4, no. 6, pp. 4–7, 2014.

[16] E. N. Bhatia, "Optical Character Recognition Techniques : A Review," *IJARCSSE*, vol. 4, no. 5, pp. 1219–1223, 2014.

[17] M. Shen, "Improving OCR Performance with Background Image Elimination," *2015 12th Int. Conf. Fuzzy Syst. Knowl. Discov.*, pp. 1566–1570, 2015.

[18] A. Coates *et al.*, "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning."

[19] P. Road, "Confidence Guided Progressive Search and Fast Match Techniques for High Performance ChineseEnglish OCR *," *IEEE*, pp. 89–92, 2002.

[20] H. Wang and J. Kangas, "Character-Like Region Verification for Extracting Text in Scene Images," no. 11, 2001.

[21] I. Kastelan, S. Kukolj, V. Pekovic, V. Marinkovic, and Z. Marceta, "Extraction of Text on TV Screen using Optical Character Recognition," *IEEE*, pp. 153–156, 2012.

[22] J. Diaz-escobar, "Optical Character Recognition based on phase features," *IEEE*, 2015.

[23] A. Thilagavathy, K. Aarthi, and A. Chilambuchelvan, "A Hybrid Approach to Extract Scene Text from Videos," *ICCEET,* pp. 1017–1022, 2012.

[24] S. Goyal, "Optical Character Recognition," *IJARCSSE,* vol. 3, no. 11, pp. 982–985, 2013.

[25] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents," pp. 525–532, 2008.

[26] L. S. Yaeger, B. J. Webb, and R. F. Lyon, "Search for Online , Printed Handwriting N EWTON," *Am. Assoc. Artif. Intell.,* vol. 19, no. 1, pp. 73–90, 1998.

[27] J. Hu, S. G. Lim, and M. K. Brown, "Writer independent on-line handwriting recognition using an HMM approach," *J. PATTERN Recognit. Soc.,* vol. 33, pp. 133–147, 2000.

[28] A. Funada, D. Muramatsu, and T. Matsumoto, "The Reduction of Memory and the Improvement of Recognition Rate for HMM On-line Handwriting Recognition," *IEEE,* pp. 0–5, 2004.

[29] J. r´ı Matas, "Real-Time Scene Text Localization and Recognition," *IEEE,* pp. 3538–3545, 2012.

[30] H. Lin and C. Hsu, "Optical Character Recognition with Fast Training Neural Network," *IEEE,* pp. 1458–1461, 2016.

[31] C. N. E. Anagnostopoulos, I. E. Anagnostopoulos, V. Loumos, and E. Kayafas, "A License Plate-Recognition Algorithm for Intelligent Transportation System Applications," *IEEE,* vol. 7, no. 3, pp. 377–392, 2006.

[32] Y. J. Zhang, "A survey on evaluation methods for image segmentation," pp. 1–13.

[33] A. Singh, K. Bacchuwar, and A. Bhasin, "A Survey of OCR Applications," *Int. J. Mach. Learn. Comput.,* vol. 2, no. 3, pp. 314–318, 2012 .