CHAPTER 5
POST PROCESSING

5.1 INTRODUCTION

The post processing techniques is applied to enhance the performance of text extraction system. It is an integral part of any OCR. This process is same in the case of Gurmukhi text extraction from natural scene images. The following post processing techniques are implemented:

i. Error detection and correction in character recognition.
ii. Gurmukhi word pronunciation based on Phoneme.

The post processing is applied immediately after the recognition process where, the output of the recognition process acts as input of post processing phase. The purpose of the post processing techniques is to validate recognized words or suggested alternatives words. The text present in the natural scene image is closer to the printing text which makes it easy to recognize. But there are various issues that are involved in the recognition process. These issues are as follows:

- Text in Gurmukhi script is split into three zones namely upper zone, middle zone and lower zone. The headline connects upper characters with middle zone characters. The head line is required to be removed so that upper and middle sub characters can be segmented. It happens sometime that some regions of a sub character image which touches upper zone may be stripped off while separating both the zones. For example ਰਾ , ਦੀ

- After trimming headline some of the Gurmukhi characters generate similar characters features which are difficult to distinguish. Table 5.1 shows similar characters features which are difficult to distinguish.

Table 5.1: Removal of headline cause confusion while recognising the character

With headline		After removal of headline	
 Figure(a)	 Figure(b)	 Figure(e)	 Figure(f)
 Figure(c)	 Figure(d)	 Figure(g)	Figure(h)
The figure(a) and figure(c) are with headlines and when these headlines are removed from that results into similar images, as shown in figure(b) and figure(d)		The figure (e) and figure (g) are with headlines and when these headlines are removed from that results to similar images, as shown in figure (h), and figure (h).	

- The binarisation along with thinning is used for various feature extraction of Gurmukhi characters. The skeleton of few Gurmukhi characters are similar and causes confusion in recognition. Following table 5.2 shows the list of few characters which have similar skeleton.

Table 5.2: Example of Gurmukhi characters having similar skeletons after thinning and binarisation

Similar skeletons		Similar skeletons	
 Figure(a)	 Figure(b)	 Figure(e)	 Figure(f)
 Figure(c)	 Figure(d)	 Figure(g)	Figure(h)
Figure (a) and Figure (b) have similar skeletons characters with very minor difference at the bottom of the character (Lower zone). Figure(c) and Figure (b) have difference of lower sub character only.		Figure (e) and Figure (f) have similar skeleton as there is only difference of lower characters. Figure (g) and Figure (h) have only difference in middle zone.	

Various post processing techniques are available in literature as listed below: Contextual processing

- Dictionary matching or word validation technique
- Shape similarity based processing
- N-grams based validation
- Language translation

The word validation technique based on dictionary lookup method is best suitable for word level post processing methods. In the methods, recognized word is searched in the pre existing databases of words. If word is found in the database then recognized word is accepted otherwise a list of recommended words based on some similarity rules is listed. The problem of this method is the storage and searching huge word databases. To overcome the problem of large databases and long searching time, the concept of N-gram is introduced. The N-gram is a technique to store data in string format such that it can represent different combination of characters. The word database is stored in string format which represents different combinations of words. In another approach, the word collection table [110] is used to store combination of similar words to enhance the performance in searching operation.

Sinha[111] developed post processing technique Devanagari text error correction which is based on rule-based contextual. Bansal and Sinha[112] developed a word dictionary for enhancing reading of Devanagari character strings. Yet another method of post-processing is based on visual similarity. Chanda [113] have developed an OCR error detection and correction technique for Bangla.The technique applied to separate input word into different parts such as lexicons of root, suffixes. The input words are tested against the grammatical agreements using lexicons of root and suffix pairs.

It is a challenging situation to identify the words which have structural similarities for example: ਚੋਰ (CHOR) or ਚੌਰ (CHAUR), ਸੇਵਾ (SEVA) or ਮੇਵਾ (MEVA). To handle such problems system has implemented consonants based special encoding technique.

5.2 ERROR DETECTION AND CORRECTION IN RECOGNITION

The Gurumukhi text recognition returns a sequence of characters to be verified and replaced with the correct one. The classification phase does not always provide accurate results. The results need to be refined using post processing techniques. Error detection and correction is one of the most commonly used methods to enhance the performance of recognition system.

The present scheme for error detection and correction is based on similarity of shape in which the alphabets of Gurmukhi script are divided into 8 subsets. The similarity of shape rules help to place alphabets in same subset which share same geometrical similarity.

For example, Gurmukhi characters ਮ and ਸ, are similar but they have difference of upper headline. So, if any one character out of these two is misspelled then alternative should be from these two. Similarly, list of such combination will help to generate most appropriate correction alternatives.

Initially, The Gurmukhi script is classified into character set of 8 subsets based on their shape similarities. For example sub-code 1 represents ੳ, 2 represents ਚ, ਟ, ਰ, ਹ, ਢ, ਦ, ਫ, ਫ਼, ੲ, ਝ, ਵ, ਛ, similarly other characters are coded. Table 5.5 shows various subsets of Gurmukhi text. Each word has been given a code depending upon number in subset table. Each code has assigned a maximum length of size 10.

In the second step, the database dictionary has been created where 74,205 words of street name, city name, frequently used wording in banners, notice board are included. The third step deals with validation based on the special code assigned to consonants.

In third step, each recognized word from reorganization phase is matched with consonant of dictionary stored in memory as AVL tree. If the recognized word is found in the word list of word code, then the source word could be a correctly recognize the word, otherwise, highly ranked word will be suggested from the list.

5.2.1 Creating Dictionary

The dictionary contains most frequent common nouns of Person and place. It also contains most frequent words used in caution board. Some of them are listed in table 5.3

Table 5.3: Examples of most frequent word used in notice boards, caution boards

Sr. No.	Word	Sr. No.	Word	Sr. No.	Word	Sr. No.	Word
1	ਖ਼ਤਰਾ	7	ਰੂਕੋ	13	ਹੌਲੀ	19	ਗਤੀ
2	ਰਫ਼ਤਾਰ	8	ਧਿਆਨ	14	ਵਾਹਨ	20	ਪੁਲਿਸ
3	ਪੈਟਰੋਲ	9	ਪੰਪ	15	ਹੋਟਲ	21	ਕਾਲਜ
4	ਹਸਪਤਾਲ	10	ਸਵਾਗਤ	16	ਰੇਲਵੇ	22	ਯੂਨੀਵਰਸਿਟੀ
5	ਖੱਬੇ	11	ਮੋੜ	17	ਸਰਵਿਸ	23	ਸਟੇਸ਼ਨ
6	ਇਜਾਜ਼ਤ	12	ਤਸਦੀਕ	18	ਬੱਸ	24	ਅੱਡੇ

The dictionary also includes separate entries for few names of villages, cities, districts and names of states.

The 10 digits code is assigned to each word according to the character set given in Table 5.5. Each code is of 10 digits. The words are stored with their respective codes in the dictionary. The AVL tree data structure has been used to create dictionary. It has lowest search complexities ($\log_2 N$). The node in AVL tree has different fields shown in table 5.4.

Table 5.4: Different fields in AVL tree with balancing factor

Sr. No.	Name of the field	Purpose
1.	Code	10-digit Code of the Word
2	WordsCount	Represents No. of words in the words list
3	LeftPtr	left pointer
4	RightPtr	Right pointer
5	BF	Repesents Balancing Factor to balance the Tree

List of common words and most frequently used words found in natural scene images are typed in Unicode typing editor. This file is processed by AVL tree to store into memory. Every word is read from the text file and assigned a 10-digit code as per the codes given in Table 5.5.

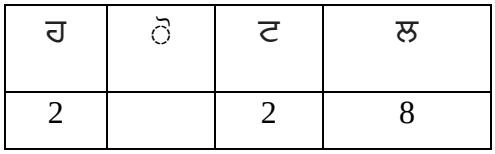
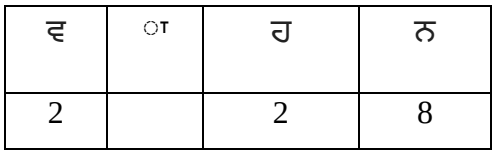
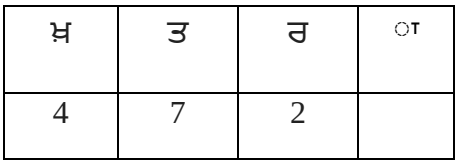
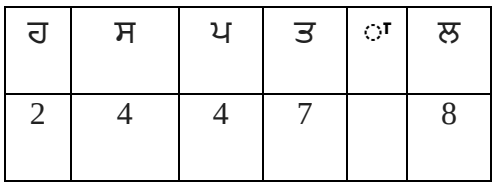
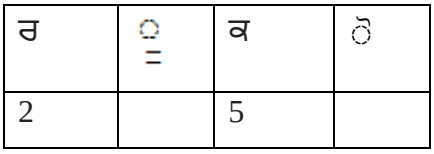
Table 5.5: Shape similarity based coding scheme of Gurmukhi characters

Code No.	Sets of Consonants
1	ਓ ਔ ਉ ਊ
2	ਚ ਟ ਣ ਰ ਹ ਢ ਦ ਫ ਫ਼ ਏ ਏ ਇ ਈ ਝ ਵ ਛ
3	ਜ ਜ਼
4	ਖ ਖ਼ ਬ ਧ ਥ ਪ ਯ ਮ ਸ ਸ਼ ਗ ਗ਼
5	ਕ ਝ
6	ਅ ਆ ਐ ਐ ਘ
7	ਤ ਡ ਭ ੜ
8	ਠ ਨ ਲ ਲ਼

For smaller words, the code is padded with 0s to make it of 10 digits. For example word ਹੋਟਲ have code: 2280000000. Table 5.6 shows examples of 10 Digit presentations of some Gurmukhi words.

Table 5.6: Example of 10 Digit code

Word(s)	Value of each character in the word					Final Code
ਕਾਲਜ		ਕ	ਾ	ਲ	ਜ	5830000000
		5		8	3	

ਯੂਨੀਵਰਸਿਟੀ		1822242000
ਹੋਟਲ		2280000000
ਵਾਹਨ		2280000000
ਖ਼ਤਰਾ		4720000000
ਹਸਪਤਾਲ		2447800000
ਰੂਕੋ		2500000000

5.2.2 Search word using Dictionary Look up Method

The word is searched in the AVL tree, where, left and right values represent the left and right sub-tree respectively. The balancing factor (BF) is used to maintain

the balance of the tree and it is computed by subtracting the height of the right sub-tree from height of the left sub-tree. The following procedure is carried out to search the word in Dictionary.

i. Check the word in the dictionary.
   a) If the word is present in the dictionary that means that word is correct and no change is required there.
   b) If the word is not found in the dictionary, and then the word might be misspelled.

Move to step 2.

ii. Algorithm is executed to generate a list of suggestions to replace the misspelled word.
   a) The procedure is executed on the list of generated suggestions of words.
   b) Highest ranked word is suggested and replaced with the misspelled word.

5.2.3 Results and Discussions

Error detection and correction have improved the recognition rate by an average 2.1 % for Gurmukhi alphabets, 3.1% for digits. Table 5.7 shows rate of improvement in the recognition.

Table 5.7: Improvement after post processing

Fields	Before Post processing	After post processing	Improvement
Alphabets	90.13	92.23	2.1
Digits	95.3	98.4	3.1
Overall	92.715	95.315	2.6

The overall result of recognition has improved up to 95.3 % for machine printed Gurmukhi text.

5.3 WORD PRONUNCIATION BASED ON PHONEME

Text to speech synthesis is a procedure to convert text into speech. In this procedure, first of all, the input text is analyzed and then transcribed into a phonemes representation. Where, a Punjabi phoneme can be defined as a minimum

sound unit of a language by which the meaning of the character can be conveyed and can be differentiated from others. Secondly, database of speech waveform is created according to synthesis rules.

The phoneme based system is developed to pronounce the word in Gurmukhi. In literature [114], the three major types of database speech synthesis are reported: phoneme based synthesis, domain specific and unit selection synthesis. Htun et al. [115] introduced Phoneme based speech synthesis method which is concatenation of phonetic units to form word. The Phonemes are considered as the synthesis units which are stored in memory. Ahmed et al. [116] proposed text-to-speech synthesis system based on phonetic concatenation. According to which, sound rules, input text transcripts into phonetics. The speech synthesis is based on the rules of Time Domain Pitch Synchronous OverlapAdd (TD-PSOLA). The system selects the recorded phoneme units from database and modifies the duration according (TD-PSOLA).

5.3.1 Gurmukhi Script Phoneme

Gurmukhi Phonemes are called segmented phoneme which include twenty vowels and thirty eight consonants. Out of twenty vowels ten are non-nasalized shown in table 5.8. Ten are nasalized listed in Table 5.9. Out of thirty eight consonants five are nasalized which are listed in table 5.10.

Table 5.8: List of Non-nasalized vowels

ੳ	ੁ	ੂ	ਅ	ੲ	ਐ	ਇ	ੀ	ਏ

Table 5.9: List of Non-nasalized vowels

ਇੰ	ੀਂ	ਏਂ	ਅੰ	ਆਂ	ਐਂ	ੁੰ	ੂੰ	ੋਂ

Table 5.10: List of Non-nasalized vowels

ਙ	ਞ	ਣ	ਮ	ਨ

Punjabi vowels can be classified on the basis of opening of mouth, position of tongue tip and rounding of the tongue, whereas Punjabi consonants can be classified on the basis of place of co-articulation and manner of articulation.

5.3.2 Procedure to Implement Text To Speech Using Phonemes

Word pronunciation system is based on phonemes in which first of all, the phonemes dictionary is created. Phoneme sound recording is carried out to represent different phonemes. The input word is segmented into number of phonemes and is compared with database for sound entries. The sounds are then played as per sequence of phonemes of the word. Figure 5.1 shows working procedure of text to speech system

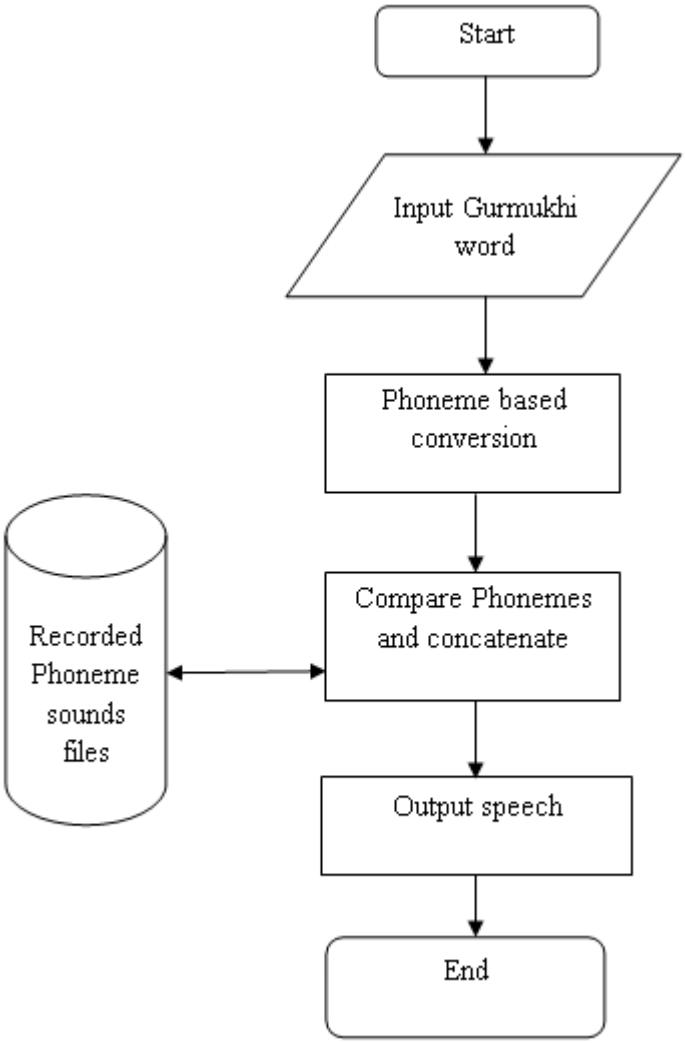


Figure 5.1 Procedure to Text-to-Speech-Phoneme based synthesis

Following are the steps which are followed to implement the system.

i. Create dictionary of Phonemes sounds which will represent the corpus of different phonemes.
ii. The input word in Gurmukhi is divided into different phonemes.

iii. Individual phoneme is searched into sound database and is played one after another

The corpus of limited Phonemes of Gurmukhi script is created. The labeling is carried to assign unique names to phoneme sound files. The recognized word in Gurumukhi script is divided into independent unit consonant which are pronounced using phoneme corpus.

The word pronunciation is the final outcome of the present system. The present system is aimed to extract the Gurmukhi text from natural scene images and make the information available in the form of voice message. Initially, the text detection and localization techniques are applied on input natural scene image. The segmentation technique is applied on detected text area. The individual characters are recognized using feature extraction and classification techniques. Finally, once the text is recognized, then present text to speech technique is applied. The performance of the present text to speech system can be enhanced by using efficient data structure and large corpus of phonemes.