International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 11, November-2014

Recognition of Handwritten Marathi Vowels using Combination of Topological and Statistical Features

C. H. Patil
Department of Computer Science
Yashwantrao Mohite College,
Bharati Vidyapeeth Deemed University,
Pune

Abstract: **In this paper, combination of horizontal feature and normalized chain code feature are used to recognize Handwritten Marathi Vowels. Recognition of handwritten Marathi vowels is a challenging task due to their interclass structural similarities. This paper describes a method for recognition of handwritten Marathi Vowels. Since a benchmark database does not exist for handwritten Marathi vowels, as a part of work database of 2294 handwritten Marathi vowels was created. Pre-processing techniques are applied to remove noise, horizontal and chain code features are extracted. According to SVM and KNN classifier the recognition rate achieved is 94.79% and 90.83% respectively using only 33 features.**

Keywords: Handwritten Marathi Vowels Recognition; OCR; Horizontal feature; Chain Code feature; SVM classifier; KNN classifier;

I. INTRODUCTION

Marathi is an official language of Maharashtra and Marathi is written in devanagari script. Marathi is 15th most spoken language in the world. Marathi language consists of 12 vowels and 36 consonants making a total of 48 characters.

Recognizing handwritten Marathi characters is important because of its application in various fields like bank cheque automation, postal automation, form processing's, historical document preservation, etc [1-5, 16]. Recognition of Handwritten Marathi vowel is a difficult and challenging task due to interclass and intraclass similarities. Vowels are used as characters in the formation of Marathi words as well as are combined with consonants as modifiers. Modifier comes above header line, or at the bottom of character or in line.

Data collection, Pre-processing, Feature Extraction and Classification which are the major steps in OCR[8, 21] are shown in Fig. 1.

II. DATA COLLECTION

Since a standard database does not exist[1-29] for handwritten Marathi vowels, an attempt was made to develop a database of isolated Marathi handwritten vowels to enable experiments to be carried out. Specially designed A4 sheets are used for data collection. Twenty writers were chosen from different professions including students, clerks, teachers and were asked to write the Marathi vowels on the datasheets provided. No constraints were imposed on the use of ink or pen except that they have to write the characters in the boxes of the sheets provided. A sample sheet of handwritten Marathi vowels is shown in Figure 2.1.

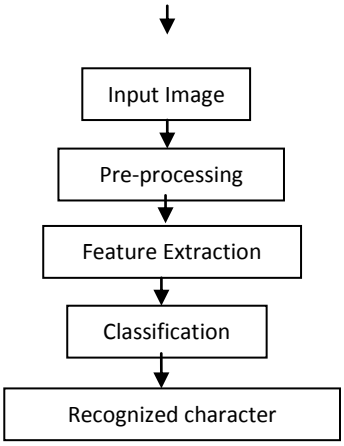

Fig. 1. Steps in isolated handwritten character recognition

The data sheets were scanned using a flat bed scanner at a resolution of 1200 dpi and stored as gray scale images. From the scanned gray scale image, the character images were cropped manually and stored in respective class folders. Figure 2.2 shows some characters cropped from the scanned image of a datasheet in gray scale.

III. PRE-PROCESSING

Pre-processing commonly involves in normalizing the intensity of the individual particles images by removing reflections, and masking portions of the images[5, 7, 19, 23]. Pre-processing enhances recognition rate of the images prior to feature extraction.

The raw input for the digitizer typically contains noise due to erratic hand movements and inaccuracies in digitization of the actual input. In order to reduce the blurring of character edges and suppress noise, the median filter is used. In median filtering, the idea is to replace the current point in the image by the median of the brightness in its neighbourhood. A 3×3 square neighbourhood is used to remove noise from the gray scale images[1-9, 19, 25-27].

Image binarization is performed on input image [1-7,19]. Histogram-shape based image thresholding suggested by Otsu's is used for converting gray scale image to binary image.

Fig. 2.1 Sheet for Handwritten vowel

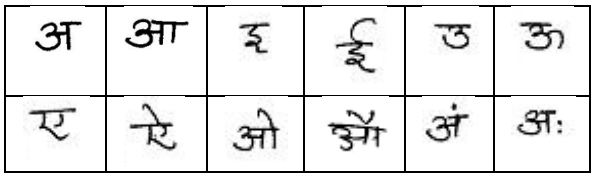

Fig. 2.2 Handwritten Marathi Vowels

The algorithm assumes that the image contains two classes of pixels (foreground and background) prior to thresholding and it calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal.

The binarized character image is mapped onto a standard plane (with predefined size) so as to give a representation of fixed dimensionality for classification. The goal of character normalization is to reduce the inter-class variation of the shapes of the characters in order to facilitate feature extraction process and improve their classification accuracy. We have used linear normalization method to standardize the character images. The standard plane is considered as a square of size 50 x 50. The width and height ratio of the character image is not disturbed due to normalization.

The goal of character thinning is to remove pixels so that an object without holes shrinks to a minimally connected stroke, and an object with holes shrinks to a ring halfway between the hold and outer boundary [1-5, 19].

IV. FEATURE EXTRACTION

A. HORIZONTAL FEATURE

To extract horizontal feature of the binary image representing the handwritten character is first preprocessed and is normalized to size of 50 x 50 pixels Fig 4(a). The size-normalized image is divided into 25 equal zones, each zone is of size 10 x 10 as shown in Fig. 4(b). Each zone has 10 horizontal lines, each horizontal line is summed to get a single sub feature and thus 10 sub-features are obtained from the each zone as shown in Fig. 4(c).

These 10 sub-features values are averaged to form a single feature value and assigned as horizontal feature to the corresponding zone. This procedure is sequentially repeated for the all the zones. Finally, 25 features are extracted for 25 zones for each character.

B. CHAIN CODE FEATURE

To extract Freeman chain codes first locate any boundary pixel, called as starting pixel, and then move along the boundary of character either clockwise or anticlockwise direction, find out next boundary pixel and allocate this new pixel a number

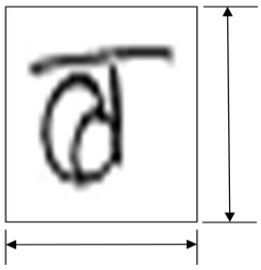

Fig.4(a) Pre-processed Sample Handwritten Marathi Vowel of size 50 x 50.

Z1	Z2	Z3	Z4	Z5
Z6	Z7	Z8	Z9	Z10
Z11	Z12	Z13	Z14	Z15
Z16	Z17	Z18	Z19	Z20
Z21	Z22	Z23	Z24	Z25

Fig. 4(b) Character image of size 50 x 50 divided into 25 zones.

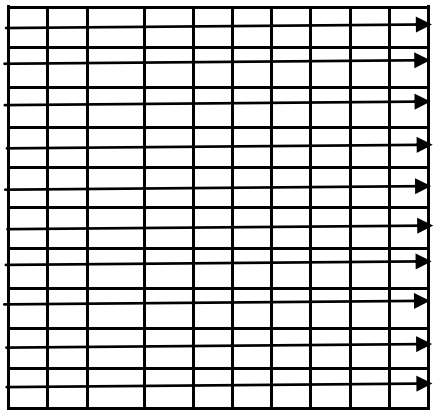

Fig. 4(c) Zone contains 10 horizontal lines which are summed up to get 10 subfeatures.

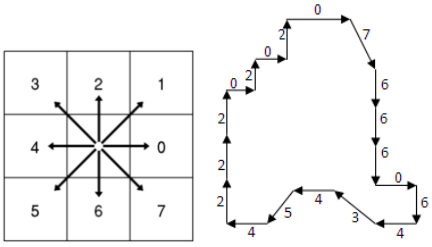

Fig. 5 Eight directional Chain code

depending upon its direction from the previous pixel is called code for that pixel. The process is repeated till starting pixel is not encountered. The codes may be 4-directional or 8-directional depending upon 4-connectivity or 8-connectivity of a pixel to its neighboring contour pixel. An 8-directional chain coded image is given in fig. 5.

The chain code extracted from above process is different for different characters as length of each chain code depends on the size of the handwritten characters. Example shows Chain code extracted for the image shown in Fig. 5.
Chain code: [07666064345422202 02], V1= [07666064345422202 02].
Compute the frequency of the codes 0, 1, 2…..7. For vector V1 frequency vector V2= [4 0 5 1 3 1 3 1]
The normalized frequency will be computer using the formula V3=V2/|V1| where |V1|=ΣV2.
For example considered above, we have V3= [0.22 0 0.27 0.05 0.16 0.05 0.16 0.05]
Finally, V3 will be the required feature vector of size 8.

C. ALGORITHM TO CALCULATE FEATURE VECTOR:

Steps to calculate feature vector for classification
Algorithm:

1. Binarize the image by applying Otsu's algorithm to obtain the binary image with character representing binary 1 and background 0
2. To bring uniformity among the characters, character image is cropped and resize it to a size of 50 x 50 pixels.
3. Divide the input image into 25 equal zones of size 10 x 10; calculate Horizontal density feature for 25 zones. Store 25 features in feature vector.
4. Extract the boundary of the character image and resample the boundary in order to obtain a uniform resampling along the running arc length of the boundary.
5. Trace the boundary in counterclockwise direction and generate 8 dimensional chain codes 0 to 7.
6. Compute the frequency of the codes 0 to 7.
7. Divide frequency of each code by sum of the frequencies.
8. Store eight features in feature vector.
9. Combine feature vector of 25 features and feature vector of 8 features to get final feature vector of length 33.

V. CLASSIFICATION

As discussed in section 4 feature vector is created for every image. Using KNN and SVM classifiers experiments are carried out and class labels are assigned to images.

A. K-NN

The k-Nearest Neighbor (k-NN) classifies an unknown sample based on the known classification of its neighbors [13, 16, 20, 24]. Suppose that a set of samples with known classification is available, the so-called training set. Intuitively, each sample should be classified similarly to its surrounding samples. Therefore, if the classification of a sample is unknown, then it could be predicted by considering the classification of its nearest neighbor samples. Given an unknown sample and a training set, all the distances between the unknown sample and all the samples in the training set can be computed. The distance with the smallest value corresponds to the sample in the training set closest to the unknown sample. Therefore, the unknown sample may be classified based on the classification of this nearest neighbor. k- NN is an instance-based learning type classifier, or lazy learning where the function is only approximated locally and all computation is deferred until classification. Euclidean distance is used.

B. SVM

Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification [3, 4, 3]. SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

VI. RESULTS

Classification of test samples is carried out by using SVM and KNN classifiers. Out of total samples 80% samples are chosen for training and remaining for testing. Experiments are carried out for size normalized as 40 x 40, 50 x 50 and 60 x 60. Features extracted for 40 x 40, 50 x 50 and 60 x 60 size are 24, 33 and 44 features respectively. As shown in Table 1 Recognition rate for size 50 x 50 is better than 40 x 40 and 60x 60 using SVM and KNN classifier.

TABLE 1 RESULT FOR IMAGE SIZED NORMALIZED TO DIFFERENT SIZE

Size	No. of features	SVM	KNN
40x40	24	90.83	89.17
50x50	33	94.79	90.83
60x60	44	92.5	86.25

VII. CONCLUSION

This paper describes a simple and efficient method to extract Horizontal and Chain code features. It may be observed that pre-processing techniques are used to improve recognition rate. Experimental results show that combination of Horizontal and Chain code features sufficiently improve recognition rate. Recognition rate when image size normalized to 50 x 50 is better than when it is 40 x 40 or 60 x 60. The proposed algorithm exhibits seems to be insensitive writing style, ink, size, noise and character slant. Using SVM classifier gives better recognition rate than KNN classifier. In this work recognition rate for handwritten Marathi vowels is improved to 94.79% while reducing number of features to 33. Future work will focus on reducing pre-processing and the number of features used for recognition.

VIII. ACKNOWLEDGMENT

The author is grateful to Dr. M. S. Prasad and Dr. S. M. Mali, for their helpful discussion and encouragement during this work.

IX. REFERENCES

[1] Ajmire P.E. And Warkhede S.E. Handwritten Marathi Character (Vowel) Recognition. Advances In Information Mining, Issn: 0975–3265, Volume 2, Issue 2, 2010, Pp-11-13

[2] Anilkumar N. Holambe And Dr.Ravinder.C.Thool. Comparative Study Of Different Classifiers For Devanagari Handwritten Character Recognition. International Journal Of Engineering Science And Technology Vol. 2 (7), 2010, 2681-2689

[3] Holambe A.N., Thool R.C., Shinde U.B. And Holambe S.N. Brief Review Of Research On Devanagari Script. International Journal Of Computational Intelligence Techniques, ISSN: 0976–0466, Volume 1, Issue 2, 2010, Pp-06-09

[4] Jyotsna Vaid And Ashum Gupta. Exploring Word Recognition In A Semi-Alphabetic Script: The Case Of Devanagari. 2002 Elsevier Science (Usa) 0093-934x/02

[5] Latesh Malik And Dr. P.S. Deshpande. Recognition Of Printed And Handwritten Devanagari Characters With Regular Expression In Finite State Models. V. Sn_A_Sel (Ed.): Digital Technology Journal 2009, Vol. 2, Pp. 1{7}.

[6] M C Padma And P A Vijaya. Identification Of Telugu, Devanagari And English Scripts Using Discriminating Features. International Journal Of Computer Science & Information Technology (IJCSIT), Vol 1, No 2, November 2009

[7] M. Hanmandlu, O.V. Ramana Murthy And Vamsi Krishna Madasu. Fuzzy Model Based Recognition Of Handwritten Hindi Characters. Digital Image Computing Techniques And Applications. IEEE Computer Society0-76954-3067-2/07-2007.

[8] Mohit Mehta, Member, Iacsit, Rupesh Sanchati And Ajay Marchya. Automatic Cheque Processing System. International Journal Of Computer And Electrical Engineering, Vol. 2, No. 4, August, 2010

[9] Mrs.Vinaya. S. Tapkir And Mrs.Sushma.D.Shelke. Ocr For Handwritten Marathi Script. International Journal Of Scientific & Engineering Research Volume 3, Issue 8, August-2012 ISSN 2229-5518.

[10] Naresh Kumar Garg, Lakhwinder Kaur And M. K. Jindal. Segmentation Of Handwritten Hindi Text. 2010 International Journal Of Computer Applications (0975 – 8887) Volume 1 – No. 4

[11] P. B. Khanale And S.D. Chitnis. Handwritten Devanagari Character Recognition Using Artificial Neural Network. Journal Of Artificial Intelligence 4(1):55-62, 2011

[12] P. S. Deshpande, Latesh Malik And Sandhya Arora. Fine Classification & Recognition Of Hand Written Devnagari Characters With Regular Expressions & Minimum Edit Distance Method. Journal Of Computers, Vol. 3, No. 5, May 2008

[13] P. S. Deshpande, Mrs. Latesh Malik And Mrs. Sandhya Arora. Recognition Of Hand Written Devnagari Characters With Percentage Component Regular Expression Matching And Classification Tree. I-4244-1272-2/07 2007 IEEE

[14] P.B. Khanale. Recognition Of Marathi Numerals Using Artificial Neural Network. Journal Of Arificial Intelligence 3(3): 135-140, 2010

[15] Prachi Mukherji And Priti P. Rege. Fuzzy Stroke Analysis Of Devnagari Handwritten Characters. WSEAS Transactions On Computers Issue 5, Volume 7, May 2008

[16] Prachi Mukherji And Priti P. Rege. Shape Feature And Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition. Journal Of Pattern Recognition Research 4 (2009) 52-68

[17] R. J. Ramteke. Invariant Moments Based Feature Extraction For Handwritten Devanagari Vowels Recognition. 2010 International Journal Of Computer Applications (0975 - 8887) Volume 1 – No. 18

[18] Raghuraj Singh, C. S. Yadav, Prabhat Verma And Vibhash Yadav. Optical Character Recognition (OCR) For Printed Devnagari Script Using Artificial Neural Network. International Journal Of Computer Science & Communicationvol. 1, No. 1, January-June 2010, Pp. 91-95

[19] S. Arora, D. Bhattacharjee, M. Nasipuri , D.K. Basu, M.Kundu, L.Malik. Study Of Different Features On Handwritten Devnagari Character. Second International Conference On Emerging Trends In Engineering And Technology, ICETET-09

[20] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu And Mahantapas Kundu. Combining Multiple Feature Extraction Techniques For Handwritten Devnagari Character Recognition. 2008 IEEE Region 10 Colloquium And The Third ICIIS, Kharagpur, India December 8-10

[21] Sandhya Arora. Debotosh Bhattacharjee, Mita Nasipuri, L. Malik , M. Kundu And D. K. Basu. Performance Comparison Of SVM And ANN For Handwritten Devnagari Character Recognition. IJCSI International Journal Of Computer Science Issues, Vol. 7, Issue 3, No 6, May 2010

[22] Sushama Shelke And Shaila Apte. A Multistage Handwritten Marathi Compound Character Recognition Scheme Using Neural Networks And Wavelet Features. International Journal Of Signal Processing, Image Processing And Pattern Recognition Vol. 4, No. 1, March 2011

[23] U. Pal, T. Wakabayashi And F. Kimura. Comparative Study Of Devnagari Handwritten Character Recognition Using Different Feature And Classifiers. 2009 10th International Conference On Document Analysis And Recognition

[24] Vandana M. Ladwani And Mrs. Latesh Malik. Survey Of Various Approaches Towards Handwritten Devnagari Word Recognition. International Journal On Computer Engineering And Information Technology 2010

[25] Veena Bansal And R. M. K. Sinha. Integrating Knowledge Sources In Devanagari Text Recognition System. IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 30, No. 4, July 2000

[26] Vikas J Dongre And Vijay H Mankar. A Review Of Research On Devnagari Character Recognition. International Journal Of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010

[27] Vandana M. Ladwani And Mrs. Latesh Malik. Survey Of Various Approaches Towards Handwritten Devnagari Word Recognition. International Journal On Computer Engineering And Information Technology 2010

[28] Veena Bansal And R. M. K. Sinha. Integrating Knowledge Sources In Devanagari Text Recognition System. IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 30, No. 4, July 2000

[29] Vikas J Dongre And Vijay H Mankar. A Review Of Research On Devnagari Character Recognition. International Journal Of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010

TABLE 2. CONFUSION MATRIX FOR SIZE 40 X 40 USING SVM CLASSIFIER

CM	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	RR
अ	31	1	4	0	0	2	0	0	0	0	1	1	77.5
आ	3	36	0	0	0	0	0	0	0	0	0	1	90
इ	0	0	37	1	1	0	1	0	0	0	0	0	92.5
ई	2	0	0	37	0	0	0	0	0	1	0	0	92.5
उ	0	0	0	0	39	1	0	0	0	0	0	0	97.5
ऊ	1	0	2	0	0	36	0	0	0	0	0	1	90
ए	0	0	4	0	0	0	36	0	0	0	0	0	90
ऐ	1	0	0	1	0	0	0	38	0	0	0	0	95
ओ	0	0	0	0	0	0	0	0	39	1	0	0	97.5
औ	0	1	0	0	0	0	0	0	5	34	0	0	85
अं	1	1	0	0	0	0	0	0	2	0	36	0	90
अः	0	0	0	1	0	0	1	1	0	0	0	37	92.5

TABLE 3. CONFUSION MATRIX FOR SIZE 40 X 40 USING KNN CLASSIFIER

CM	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	RR
अ	32	2	1	0	0	0	1	0	0	0	1	3	80
आ	3	35	0	0	0	1	0	0	0	0	0	1	87.5
इ	1	0	34	1	2	0	2	0	0	0	0	0	85
ई	2	0	0	36	0	0	0	1	0	0	0	1	90
उ	0	0	0	0	40	0	0	0	0	0	0	0	100
ऊ	0	0	1	0	3	35	0	0	0	0	0	1	87.5
ए	0	0	1	0	1	0	38	0	0	0	0	0	95
ऐ	2	0	0	0	1	0	0	36	1	0	0	0	90
ओ	0	0	0	0	0	0	0	1	37	2	0	0	92.5
औ	0	0	0	0	0	0	0	0	6	34	0	0	85
अं	0	1	0	0	0	0	0	1	0	0	38	0	95
अः	3	1	1	0	0	0	2	0	0	0	0	33	82.5

TABLE 4. CONFUSION MATRIX FOR SIZE 50 X 50 USING SVM CLASSIFIER

CM	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	RR
अ	34	1	0	0	2	0	0	0	0	0	2	1	85
आ	2	38	0	0	0	0	0	0	0	0	0	0	95
इ	0	0	40	0	0	0	0	0	0	0	0	0	100
ई	0	0	0	39	0	0	0	0	0	1	0	0	97.5
उ	0	0	0	0	40	0	0	0	0	0	0	0	100
ऊ	0	0	0	0	0	40	0	0	0	0	0	0	100
ए	0	0	5	0	4	0	31	0	0	0	0	0	77.5
ऐ	0	0	0	1	0	1	0	38	0	0	0	0	95
ओ	0	0	0	0	0	0	0	0	40	0	0	0	100
औ	0	0	0	0	0	0	0	0	0	40	0	0	100
अं	0	0	0	0	0	0	0	1	0	0	39	0	97.5
अः	1	0	0	0	0	1	0	1	0	0	1	36	90

TABLE 5. CONFUSION MATRIX FOR SIZE 50 X 50 USING KNN CLASSIFIER

CM	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	RR
अ	32	2	0	0	2	1	0	0	0	0	2	1	**80.00**
आ	2	36	1	0	0	0	0	0	0	0	0	1	**90.00**
इ	0	0	39	0	1	0	0	0	0	0	0	0	**97.50**
ई	0	0	1	38	0	0	0	0	1	0	0	0	**95.00**
उ	0	0	0	0	40	0	0	0	0	0	0	0	**100.00**
ऊ	1	0	0	0	2	37	0	0	0	0	0	0	**92.50**
ए	0	0	2	0	4	0	34	0	0	0	0	0	**85.00**
ऐ	0	0	0	1	1	0	0	38	0	0	0	0	**95.00**
ओ	0	0	0	0	0	0	0	0	39	1	0	0	**97.50**
औ	0	0	0	0	0	0	0	0	4	34	2	0	**85.00**
अं	0	0	0	0	0	0	0	1	1	0	38	0	**95.00**
अः	7	0	0	0	0	0	2	0	0	0	0	31	**77.50**

TABLE 6. CONFUSION MATRIX FOR SIZE 60 X 60 USING SVM CLASSIFIER

CM	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	RR
अ	28	0	1	0	1	0	0	0	0	0	2	8	70
आ	1	39	0	0	0	0	0	0	0	0	0	0	97.5
इ	0	0	39	1	0	0	0	0	0	0	0	0	97.5
ई	0	0	0	40	0	0	0	0	0	0	0	0	100
उ	0	0	0	0	40	0	0	0	0	0	0	0	100
ऊ	0	0	1	0	1	38	0	0	0	0	0	0	95
ए	0	0	2	1	3	0	34	0	0	0	0	0	85
ऐ	1	0	0	0	0	1	0	38	0	0	0	0	95
ओ	0	0	0	0	0	0	0	0	40	0	0	0	100
औ	0	0	0	0	0	0	0	0	9	31	0	0	77.5
अं	0	0	0	1	0	0	0	0	1	0	38	0	95
अः	0	0	0	1	0	0	0	0	0	0	0	39	97.5

TABLE 7. CONFUSION MATRIX FOR SIZE 60 X 60 USING KNN CLASSIFIER

CM	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	RR
अ	27	0	2	0	2	1	0	0	0	0	2	6	67.5
आ	1	37	1	0	0	0	0	0	0	0	0	1	92.5
इ	0	0	39	0	1	0	0	0	0	0	0	0	97.5
ई	0	0	0	38	0	0	0	0	0	0	2	0	95
उ	0	0	0	0	40	0	0	0	0	0	0	0	100
ऊ	2	0	1	0	2	34	0	0	0	0	0	1	85
ए	0	0	0	1	4	0	35	0	0	0	0	0	87.5
ऐ	1	0	0	0	0	1	0	35	3	0	0	0	87.5
ओ	0	0	0	0	0	0	0	0	38	2	0	0	95
औ	0	0	0	0	0	0	0	0	14	24	2	0	60
अं	0	0	0	0	0	0	0	0	3	0	37	0	92.5
अः	4	2	0	0	2	0	0	0	0	0	2	30	75