Pattern Recognition
CS669

ASSIGNMENT 3

Consonant Vowel Segment Dataset

Bayes Classifier using
K - Nearest Neighbor Method
&
Hidden Markov Models

Group Number 8

Aman Khandelwal	B16007
Amrendra Singh	B16010
Bharat Lodhi	B16015

Contents

Page

Contents **i**

List of Plots **ii**

List of Tables **1**

1. Problem Description **3**

2. Solution Approach **4**
1 K-Nearest Neighbor Method 4
2 Hidden Markov Model 4
2.1 Obtaining set of observations from feature vectors 4
2.2 Baum-Welch Algorithm 5
2.3 Classification Using Bayes Classifier 5

3. Results **6**
1 K - Nearest Neighbor Method 6
1.1 Confusion Matrix, Precision, Recall and F-measure 6
1.2 Accuracy vs K 7
1.3 Observations & Inferences 7
2 Hidden Markov Model 8
2.1 Confusion Matrix, Precision, Recall and F-measure 8
2.2 Variation of accuracy with N and M 13
2.3 Observations & Inferences 14

4. Conclusion **15**

List of Plots

3..1 Variation of accuracy with K 7
3..2 Variation of accuracy with M for each N 13
3..3 Variation of accuracy with N for each M 13

List of Tables

3..1 KNN - Confusion Matrix and Analysis: K = 4 6
3..2 KNN - Confusion Matrix and Analysis: K = 8 6
3..3 KNN - Confusion Matrix and Analysis: K = 16 6
3..4 KNN - Confusion Matrix and Analysis: K = 32 6
3..5 KNN - Results 7
3..6 Confusion Matrix and Analysis: N = 2 , M = 4 8
3..7 Confusion Matrix and Analysis: N = 2 , M = 8 8
3..8 Confusion Matrix and Analysis: N = 2 , M = 16 8
3..9 Confusion Matrix and Analysis: N = 2 , M = 32 8
3..10 HMM Results : N = 2 9
3..11 Confusion Matrix and Analysis: N = 3 , M = 4 9
3..12 Confusion Matrix and Analysis: N = 3 , M = 8 9
3..13 Confusion Matrix and Analysis: N = 3 , M = 16 9
3..14 Confusion Matrix and Analysis: N = 3 , M = 32 10
3..15 HMM Results : N = 3 10
3..16 Confusion Matrix and Analysis: N = 4 , M = 4 10
3..17 Confusion Matrix and Analysis: N = 4 , M = 8 10
3..18 Confusion Matrix and Analysis: N = 4 , M = 16 11
3..19 Confusion Matrix and Analysis: N = 4 , M = 32 11
3..20 HMM Results : N = 4 11
3..21 Confusion Matrix and Analysis: N = 5 , M = 4 11
3..22 Confusion Matrix and Analysis: N = 5 , M = 8 12
3..23 Confusion Matrix and Analysis: N = 5 , M = 16 12
3..24 Confusion Matrix and Analysis: N = 5 , M = 32 12
3..25 HMM Results : N = 5 12

List of Symbols and Abbreviations

- All in bold symbols are matrices.
- λ - Hidden Markov Model O - Observation sequence
- **A** - State transition matrix, $\mathbf{A} = [a_{ij}]$, a n x n matrix, where a_{ij} denotes the probability associated with the transition from state i to state j.
- **B** - State observation probability matrix, $\mathbf{B} = [b_j(v_k)]$, a n x m matrix, where $b_j(v_k)$ denotes that being in state j what is the probability of observing kth symbol v_k.
- $\pi = [\pi_i]$, a $n - length$ vector denoting the probability of coming state i at t $= 1$.
- $\xi_t(i, j)$ - Probability of being in state i at time t and in state j at time $t + 1$ given O and λ.
- $\gamma_t(i)$ - Probability of transition from state i at time t.
- $\alpha_t(i)$ - Probability of observing partial observation sequence $o_1, o_2, ..., o_t$ until time t and being at state i at the time t, given model λ.
- $\beta_t(i)$ - Probability of observing partial observation sequence $o_1, o_2, ..., o_t$ given the state at time t and model λ.

1. Problem Description

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs on the basis of a training set of data containing observations (or instances) whose category membership is known.

Data-sets:

- Speech Data: ka, kA and kha

Classifiers to be built:

- Bayes classifier using K-nearest neighbour method for class-conditional density estimation using DTW distance.
- Bayes classifier using Discrete Hidden Markov Model (DHMM)

2. Solution Approach

1 K-Nearest Neighbor Method

We use the KNN classifier with the DTW as the distance measure between two data samples.

Procedure

For every data point in the test set:

1. Calculate its DTW distance from all the training samples
2. Sort the distance array obtained in increasing order
3. Consider the first K values in this sorted list. Calculate the frequency of values from each of the classes in these K selected values
4. The class from which maximum number of values occur from these K selected, the test data point it classified to that class

2 Hidden Markov Model

2.1 Obtaining set of observations from feature vectors

First we need to convert the given 39 dimensional feature vectors into a set of M observations.

Procedure

1. Take all the feature vectors of training as well as test data-set and cluster them into M clusters using K-Means clustering.
2. Every cluster center is now assigned a number between $1...M$. This acts as the observation symbol for the feature vector.
3. For every feature vector of every class in test as well as train data assign it a number between $1...M$ depending on its minimum euclidean distance from the cluster centers.

Now, we have represented every feature vector as an observation. Hence every sample file is represented as set of observations. Now we shall do the DHMM analysis on these set of observations.

2.2 Baum-Welch Algorithm

We use the expectation-maximization based Baum-Welch algorithm for Hidden Markov Models to estimate the density of incoming data distribution.

1. **Initialization** First step is to initialize the $\mathbf{A}$, $\mathbf{B}$ and π. The way initialization done was to divide the given observation sequence into the n equal parts (first part may be larger if number of observation symbols in a sequence is not perfectly divisible by n) and then assign the n parts to taken n states. Now according to the definition of $\mathbf{A}$, $\mathbf{B}$ and π, we calculate them because we have both observation sequence and state sequence.

2. Evaluate $\alpha_t(i)$ and $\beta_t(i)$ using the Forward Procedure and the Backward Procedure respectively.

   E-Step Evaluate $\xi_t(i,j)$ and $\gamma_t(i)$.

   M-Step Re-estimate $\mathbf{A}$, $\mathbf{B}$ and π from $\xi_t(i,j)$ and $\gamma_t(i)$.

   Repeat the above second step till some convergence criteria. Here the convergence criteria was to repeat above step till the difference between two consecutive total data likelihood is greater than or equal to some threshold.

2.3 Classification Using Bayes Classifier

After obtaining the $\mathbf{A}$, $\mathbf{B}$ and π using the Baum-Welch algorithm. We use these to obtain the probabilities of each test sample belonging to each of the classes and assign the class with the maximum probability obtained.

3. Results

1 K - Nearest Neighbor Method

1.1 Confusion Matrix, Precision, Recall and F-measure

	ka	kA	kha
ka	66	29	1
kA	13	114	0
kha	3	11	1

(a) Confusion Matrix

	ka	kA	kha
Precision	0.8048	0.7402	0.5
Recall	0.6875	0.8976	0.0666
F-Measure	0.7415	0.8113	0.1176

(b) Analysis

Table 3..1. KNN - Confusion Matrix and Analysis: K = 4

	ka	kA	kha
ka	55	41	0
kA	11	116	0
kha	3	10	2

(a) Confusion Matrix

	ka	kA	kha
Precision	0.7971	0.6946	1.0
Recall	0.5729	0.9133	0.1333
F-Measure	0.6666	0.7891	0.2352

(b) Analysis

Table 3..2. KNN - Confusion Matrix and Analysis: K = 8

	ka	kA	kha
ka	43	53	0
kA	6	121	0
kha	4	11	0

(a) Confusion Matrix

	ka	kA	kha
Precision	0.8113	0.6540	0
Recall	0.4479	0.9527	0
F-Measure	0.5771	0.7756	0

(b) Analysis

Table 3..3. KNN - Confusion Matrix and Analysis: K = 16

	ka	kA	kha
ka	27	69	0
kA	4	123	0
kha	2	13	0

(a) Confusion Matrix

	ka	kA	kha
Precision	0.8181	0.6	0
Recall	0.2812	0.9685	0
F-Measure	0.4186	0.7409	0

(b) Analysis

Table 3..4. KNN - Confusion Matrix and Analysis: K = 32

	K = 4	K = 8	K = 16	K = 32
Accuracy	76.05%	72.68%	68.90%	63.02%
Mean Precision	0.6817	0.8305	0.4884	0.472
Mean Recall	0.5506	0.5398	0.4668	0.4165
Mean F-Measure	0.5568	0.5636	0.4509	0.3865

Table 3..5. KNN - Results

1.2 Accuracy vs K

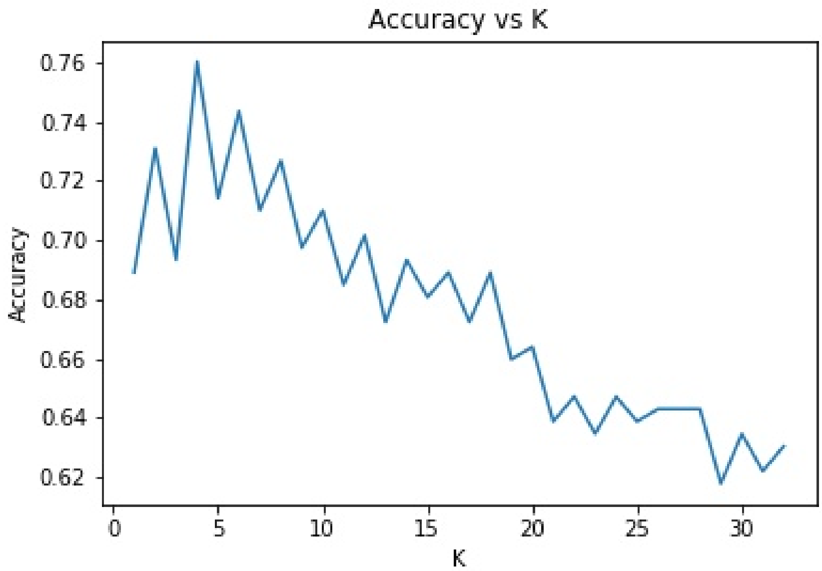

Figure 3..1. Variation of accuracy with K

1.3 Observations & Inferences

1. Maximum accuracy is observed at K = 4, 5, 6.
2. For a very low value of K (1 or 2), noise can easily influence the decision and hence the low accuracy obtained.
3. Recall and precision for class kA is highest while the same for class kha is the lowest.
4. We are taking K nearest neighbors of our test data in the train set. We are not fixing our area and finding points of every class in that area. Hence the number of data points of the train data in each class severely influences our results. For classes ka, kA and kha we have 383, 510 and 61 training samples respectively and even if k is large, then also for kha class due to the fact that the number of training samples for it are less, the accuracy will be less. Hence K-nearest neighbor method is not very effective for kha class as we can see from the results.

2 Hidden Markov Model

2.1 Confusion Matrix, Precision, Recall and F-measure

N = 2

	ka	kA	kha
ka	22	36	38
kA	41	57	29
kha	4	2	9

(a) Confusion Matrix

	ka	kA	kha
Precision	0.3283	0.6	0.118
Recall	0.2291	0.4488	0.6
F-Measure	0.2699	0.5135	0.1978

(b) Analysis

Table 3..6. Confusion Matrix and Analysis: N = 2 , M = 4

	ka	kA	kha
ka	29	33	34
kA	25	75	27
kha	4	7	4

(a) Confusion Matrix

	ka	kA	kha
Precision	0.5	0.6521	0.0615
Recall	0.3020	0.5905	0.2666
F-Measure	0.3766	0.6198	0.1

(b) Analysis

Table 3..7. Confusion Matrix and Analysis: N = 2 , M = 8

	ka	kA	kha
ka	34	37	25
kA	28	71	28
kha	6	6	3

(a) Confusion Matrix

	ka	kA	kha
Precision	0.5	0.6228	0.0535
Recall	0.5541	0.5590	0.2
F-Measure	0.4146	0.5862	0.0845

(b) Analysis

Table 3..8. Confusion Matrix and Analysis:
N = 2 , M = 16

	ka	kA	kha
ka	41	38	17
kA	39	69	19
kha	5	6	4

(a) Confusion Matrix

	ka	kA	kha
Precision	0.4823	0.6106	0.1
Recall	0.4270	0.5433	0.2666
F-Measure	0.4530	0.575	0.1454

(b) Analysis

Table 3..9. Confusion Matrix and Analysis:
N = 2 , M = 32

	M = 4	M = 8	M = 16	M = 32
Accuracy	36.97%	45.37%	45.37%	47.89%
Mean Precision	0.3489	0.4045	0.3921	0.3976
Mean Recall	0.4259	0.3864	0.3710	0.4123
Mean F-Measure	0.3270	0.3654	0.3627	0.3911

Table 3..10. HMM Results : N = 2

N = 3

	ka	kA	kha
ka	26	46	24
kA	51	63	13
kha	5	2	8

(a) Confusion Matrix

	ka	kA	kha
Precision	0.3170	0.567	0.1777
Recall	0.2708	0.4960	0.5333
F-Measure	0.2921	0.5294	0.2666

(b) Analysis

Table 3..11. Confusion Matrix and Analysis:
N = 3 , M = 4

	ka	kA	kha
ka	35	44	17
kA	22	82	23
kha	5	8	2

(a) Confusion Matrix

	ka	kA	kha
Precision	0.5645	0.6119	0.0476
Recall	0.3645	0.6456	0.1333
F-Measure	0.443	0.628	0.0701

(b) Analysis

Table 3..12. Confusion Matrix and Analysis:
N = 3 , M = 8

	ka	kA	kha
ka	32	43	21
kA	26	74	27
kha	9	3	3

(a) Confusion Matrix

	ka	kA	kha
Precision	0.4776	0.6166	0.0588
Recall	0.3333	0.5826	0.2
F-Measure	0.3926	0.5991	0.0909

(b) Analysis

Table 3..13. Confusion Matrix and Analysis:
N = 3 , M = 16

	ka	kA	kha
ka	40	41	15
kA	38	77	12
kha	4	8	3

(a) Confusion Matrix

	ka	kA	kha
Precision	0.4878	0.6111	0.1
Recall	0.4166	0.6062	0.2
F-Measure	0.4494	0.6086	0.1333

(b) Analysis

Table 3..14. Confusion Matrix and Analysis:
N = 3 , M = 32

	M = 4	M = 8	M = 16	M = 32
Accuracy	40.75%	50%	45.79%	50.42%
Mean Precision	0.3541	0.4080	0.3843	0.3996
Mean Recall	0.4334	0.3811	0.3720	0.4076
Mean F-Measure	0.3627	0.3805	0.3609	0.3971

Table 3..15. HMM Results : N = 3

N = 4

	ka	kA	kha
ka	21	47	28
kA	48	60	19
kha	3	1	11

(a) Confusion Matrix

	ka	kA	kha
Precision	0.2916	0.5555	0.1896
Recall	0.2187	0.4724	0.7333
F-Measure	0.25	0.5106	0.3013

(b) Analysis

Table 3..16. Confusion Matrix and Analysis:
N = 4 , M = 4

	ka	kA	kha
ka	29	49	18
kA	21	82	24
kha	4	10	1

(a) Confusion Matrix

	ka	kA	kha
Precision	0.5370	0.5815	0.0232
Recall	0.3020	0.6456	0.0666
F-Measure	0.3866	0.6119	0.0344

(b) Analysis

Table 3..17. Confusion Matrix and Analysis:
N = 4 , M = 8

	ka	kA	kha
ka	29	51	16
kA	27	77	23
kha	6	5	4

(a) Confusion Matrix

	ka	kA	kha
Precision	0.4677	0.5789	0.0930
Recall	0.3020	0.6062	0.26666
F-Measure	0.3670	0.5923	0.1379

(b) Analysis

Table 3..18. Confusion Matrix and Analysis:
N = 4 , M = 16

	ka	kA	kha
ka	34	51	11
kA	35	79	13
kha	3	8	4

(a) Confusion Matrix

	ka	kA	kha
Precision	0.4722	0.5724	0.1428
Recall	0.3541	0.6220	0.2666
F-Measure	0.4047	0.5962	0.1860

(b) Analysis

Table 3..19. Confusion Matrix and Analysis:
N = 4 , M = 32

	M = 4	M = 8	M = 16	M = 32
Accuracy	38.65%	47.05%	46.21%	49.15%
Mean Precision	0.3456	0.3806	0.3799	0.3958
Mean Recall	0.4748	0.3381	0.3916	0.4142
Mean F-Measure	0.3540	0.3443	0.3957	0.3956

Table 3..20. HMM Results : N = 4

N = 5

	ka	kA	kha
ka	20	46	30
kA	40	60	27
kha	2	3	10

(a) Confusion Matrix

	ka	kA	kha
Precision	0.3225	0.5504	0.1492
Recall	0.2083	0.4724	0.6666
F-Measure	0.2531	0.5084	0.2439

(b) Analysis

Table 3..21. Confusion Matrix and Analysis:
N = 5 , M = 4

	ka	kA	kha
ka	31	45	20
kA	23	78	26
kha	4	10	1

(a) Confusion Matrix

	ka	kA	kha
Precision	0.5344	0.5864	0.0212
Recall	0.3229	0.6141	0.0666
F-Measure	0.4025	0.6	0.0322

(b) Analysis

Table 3..22. Confusion Matrix and Analysis:
N = 5 , M = 8

	ka	kA	kha
ka	26	55	15
kA	23	85	19
kha	6	4	5

(a) Confusion Matrix

	ka	kA	kha
Precision	0.4727	0.5902	0.1282
Recall	0.2708	0.6692	0.3333
F-Measure	0.3443	0.6273	0.1851

(b) Analysis

Table 3..23. Confusion Matrix and Analysis:
N = 5 , M = 16

	ka	kA	kha
ka	39	46	11
kA	40	75	12
kha	4	8	3

(a) Confusion Matrix

	ka	kA	kha
Precision	0.4698	0.5813	0.1153
Recall	0.4062	0.5905	0.2
F-Measure	0.4357	0.5859	0.1463

(b) Analysis

Table 3..24. Confusion Matrix and Analysis:
N = 5 , M = 32

	M = 4	M = 8	M = 16	M = 32
Accuracy	37.81%	46.21%	48.73%	49.15%
Mean Precision	0.3407	0.3807	0.3970	0.3888
Mean Recall	0.4491	0.3345	0.4244	0.3989
Mean F-Measure	0.3351	0.3449	0.3856	0.3893

Table 3..25. HMM Results : N = 5

2.2 Variation of accuracy with N and M

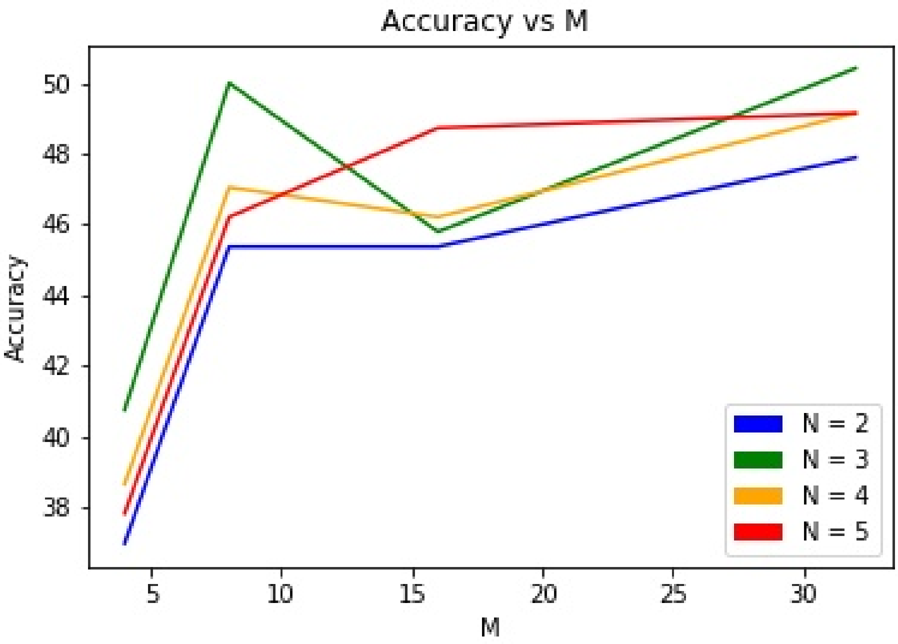

Figure 3..2. Variation of accuracy with M for each N

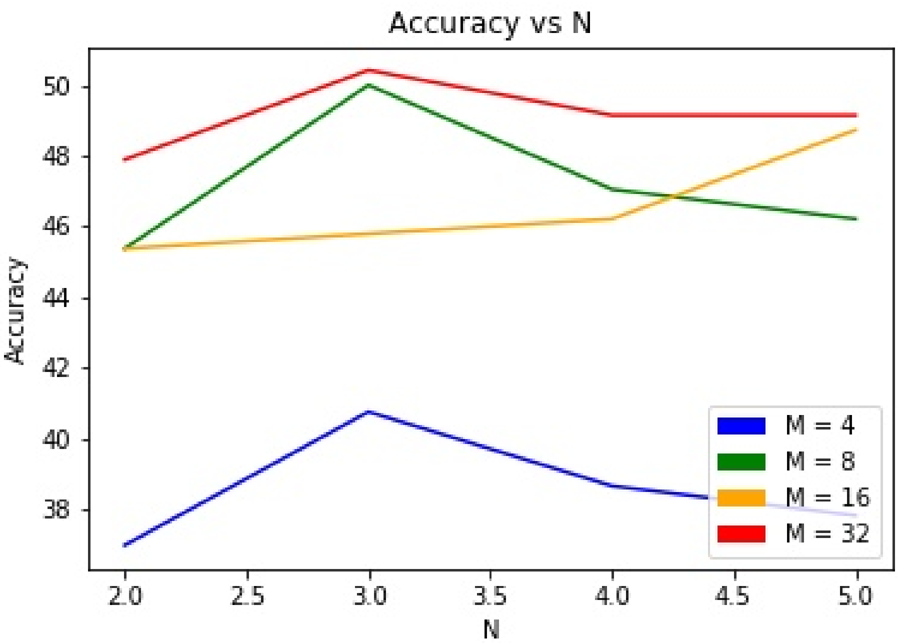

Figure 3..3. Variation of accuracy with N for each M

2.3 Observations & Inferences

1. In general, it is observed that for given value of N, accuracy increases with increasing value of M.

2. In case of k-nn the precision for third class was very poor, but for DHMM this is not the case. The reason for poor result of k-nn is due to the fact that in speech data there is dependency between current and previous states but k-nn is unable to capture this, but markovian model is essentially built upon the dependency of states.

3. Miss-classification between classes ka and kA is very high.

4. In class ka and class kA, we know that both the sounds are very similar and are different mostly in terms of the duration of the aa sound. This can be the reason that our Hidden Markov Model isn't very robust in differentiation between the classes ka and kA.

5. Accuracy for $M = 4$ is lower compared to other values of M. Accuracy for $M = 8, 16, 32$ is quite similar. This can be due to the fact that our data is not sufficiently represented with only 4 observation symbols, but representation is efficient enough for $M >= 8$.

6. Variation of accuracy with change of N is not too high, hence we can claim that our incoming data is sufficiently represented with just 2 states.

4. Conclusion

1. Although the accuracy in case of k-nn is better than HMM but the differentiation between different sounds (ka, kA, kha) is better done by DHMM.

2. Accuracy is decreasing with value of K in k-nn Classifier .

3. Accuracy is increasing with increase in the value of M in case of DHMM.

Bibliography

[1] K-Means clustering
https://en.wikipedia.org/wiki/K-means_clustering
https://www.geeksforgeeks.org/k-means-clustering-introduction/
https://towardsdatascience.com/understanding-k-means-clustering-\
in-machine-learning-6a6e67336aa1

[2] Bayes Classifier
https://en.wikipedia.org/wiki/Bayes_classifier

[3] Statistical Classification
https://en.wikipedia.org/wiki/Statistical_classification

[4] Naive Bayes Classifier
https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[5] Matplotlib Contours
https://matplotlib.org/api/_as_gen/matplotlib.pyplot.contour.html

[6] Stack Overflow
https://stackoverflow.com

[7] K-Nearest Neighbor Method https://stats.stackexchange.com/questions/252852/k-value-vs-accuracy-in-knn

[8] Baum Welch Algorithm https://en.wikipedia.org/wiki/Baum\OT1\textendashWelch_algorithm