ISSN 2278-3091
Volume 12, No.5, September - October 2023
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse051252023.pdf
https://doi.org/10.30534/ijatcse/2023/051252023

Clustering of Feature Vectors and Recognition of Bodo Phoneme Using MLP Technique

Dr. Manoj Kumar Deka[1], Gwgwm Borgoyary[2]
[1]Deptt. of Computer Science & Technology, Bodoland University, Kokrajhar, Assam, India.
manoj.deka1@gmail.com
[2]Deptt. of Computer Science & Technology, Bodoland University, Kokrajhar, Assam, India.
gwgwm96@gmail.com

Received Date : September 01, 2023 Accepted Date: September 29, 2023 Published Date: October 06, 2023

ABSTRACT

The process through which a computer can identify spoken words is termed as speech recognition. After analysis and finding of features of the speech sound, one can go towards the recognition of the speech. The extraction of feature vector is known as the feature extraction process or the front-end process. This front-end process is considered as the 1st stage of speech recognition. Pattern matching process is the 2nd stage or final stage of speech recognition where actual search is carried out to decode the spoken utterances by matching the sequence of feature vectors against the acoustic and language models stored in the recognizer. To reduce this problem, clustering technique is used. Clustering makes it possible to look at properties of whole clusters instead of individual objects - a simplification that might be useful when handling large volume of data. Clustering is nothing but the assignment of a set of observations into subsets so that the observations in the same cluster are similar in some sense.

Key words: MLP, Clustering, Speech Recognizer, Feature Extraction.

1. INTRODUCTION

The process through which a computer can identify spoken words is termed as speech recognition. After analysis and finding of features of the speech sound, one can go towards the recognition of the speech. The extraction of feature vector is known as the feature extraction process or the front-end process. This front-end process is considered as the 1st stage of speech recognition. Pattern matching process is the 2nd stage or final stage of speech recognition where actual search is carried out to decode the spoken utterances by matching the sequence of feature vectors against the acoustic and language models stored in the recognizer.

In the field of speech recognition, the contributions of Artificial Neural Networks have been playing an important role. Here, through the present study, an attempt has been made to recognize the Bodo speech through the utterances of selected Bodo phoneme i.e. Bodo vowels by using **Multi-layer Perceptron Technique (MLP).** In this technique, there are three layers of neurons, namely, one input layer, one output layer and one or more hidden layer(s).

2. CLUSTER METHOD USED

The clustering methods have traditionally been used in order to find emerging patterns from data sets with unknown properties. Clustering is a ill-defined problem for which there exist numerous methods [1,2,3]. A unilateral proposition in the speech / speaker recognition using clustering techniques is to account the high dimensional data sets. In the present study, **K-means** clustering technique has used and tested for the Bodo phonemes, corresponding to both male and female Bodo informants.

K-means clustering can best be described as a partitioning method. K-means treats each observation in data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. The K-means clustering algorithm improves iteratively the codebook iteratively according to the algorithm as shown below:

Step 0: set the number of clusters K;
Step 1: initialize cluster centroids ($\mu_1, \mu_2, \mu_3, \text{------} \mu_k$)
Step 2: classify the samples according to the nearest μ_k;
Step 3: recomputed centroids ($\mu_1, \mu_2, \mu_3, \text{------} \mu_k$)
until there is no significant change.

In general, K-means does not achieve a global minimum over the assignments. In fact, since the algorithm uses discrete assignment rather than a set of continuous parameters, the minimum it reaches cannot be properly called a local minimum. In addition, results can greatly depend on initialization.

3. DATABASE

The database used for recognition consists of isolated utterances of Bodo phonemes from Bodo (male and female)

speakers. The database has been categories into following sub-categories: Bodo Male Data (BMD), Bodo Female Data (BFD). Sample recorded for 12 informants (6 male and 6 female) from different age groups. Using the wave signal corresponding to the selected Bodo phoneme i.e. Bodo vowels, feature vectors are extracted by following feature extraction procedure and then extracted features are subjected to the K-means clustering to reduce their dimension.

4. FEATURE EXTRACTION

In the present study, for extraction of features embedded in speech signal, Formant Frequency, Linear Predictive Cepstral coefficients (LPCC), Mel Frequency Cepstral Frequency (MFCC) are used. Cepstral weighted feature vector is obtained for every frame by block processing of continuous speech signals. Speech signal is first subjected to low-pass filter to prevent the aliasing effect in sampling and removing parts of the speech signal which is not important and may contain noise. The speech waveform is then sampled at 8 kHz and quantized by a 16 bit ADC (analog-to-digital converter). The unnecessary silence period at the beginning and at the end of the speech signal has been discarded by using end-point detection algorithm. To spectrally flatten the signal, the speech signal after discarding the silence period, has been subjected to the pre-emphasis procedure through a first order digital filter. The transfer function of the said filter is (1-0.96z-1). Consecutive speech signal of 30 microseconds are taken as a single frame. Frames are overlapped by 20 microseconds. To reduce the undesired effect of Gibbs phenomenon [4], the frames are multiplied by Hamming window, which is read as [4]:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad \text{--- (1)}$$

where N is the number of sample in a block

Like other Sino-Tibetan languages, Bodo is also a tonal language. A language is said to be tonal if a word of the same phonetic structure but different tone convey different meaning, since tone is associated with pitch and cepstral coefficients based feature vector discard the pitch related information, such a feature vector is not sufficient to recognize tonal words. Again, it has been observed that if pith related feature is used, the performance of the system degrades considerably for non-tonal languages. Therefore, in our present study, the cepstral features and the pitch related features are combined. The pitch related features which are considered in the present is the second order derivatives, namely Formant Frequency.

LPC coefficient has been determined using Levinson-Durbin algorithm [5]. Instead of using directly the LPC coefficient as feature vector, cepstral coefficients, based on LPC analysis, are used because of their superior recognition capabilities [6]. Further, cepstral coefficients of LPC give better recognition result if they are weighted appropriately [7,8]. The weighting function has been given as:

$$w(n) = 1 + \frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right), 1 \le m \le Q \quad --(2)$$

where Q is the order of the cepstral coefficient. Besides cepstral coefficients, the approximations of their time derivative are also used as feature vectors. And these time derivative approximations are used to account for speech signal in terms of the dynamic characteristics. The time derivative is approximated by a linear regression coefficient over a finite window, which is given by [7,8]:

$$\Delta_{c_t}^{\wedge}(m) = [\sum_{k=-k}^{k} c_{t-k}^{\wedge}(m)\, k]G, \quad 0 \le m \le Q \text{ ---(3)}$$

where $c_t^{\wedge}(m)$ is the m[th] weighted cepstral coefficient at time t and G is a constant used to make the variances of the derivative terms equal to those with original cepstral coefficient.

In the present study, the following typical values are used N=240, P=10, Q=12, K=2 and G=0.316 [9]. The weighted cepstral coefficients and the corresponding time-derivatives are concatenated, which results in a 24-dimensional observation vector. To reduce the computational cost some of the less useful cepstral features can be discarded. After discarding the less useful cepstral features, only 13 elements are considered for the recognition of the basic phonemes in the present study, which is given by:

$$U_t(t) = \left[\hat{C}_1 \hat{C}_2 \hat{C}_3 \ldots\ldots \hat{C}_8 \Delta\hat{C}_2 \Delta\hat{C}_3 \ldots\ldots \Delta\hat{C}_6\right] \quad \text{------(4)}$$

On the other hand, in signal processing in general and speech processing in particular, use of Mel Frequency Cepstral Coefficient (MFCC) analysis is very popular. Derivation of Fourier Transform of the audio clip is the result of MFCC. In MFCC technique, the frequency bands are positioned logarithmically. But in the Fourier Transform, the frequency bands are not positioned logarithmically. As, in MFCC, the frequency bands are positioned logarithmically, it approximates the human system response more closely than in any other system [10,11]. So, coefficients from MFCC show better processing of data.

5. MLP BASED SPEECH RECOGNIZER

In the present study Multilayer Perceptron (MLP) has been used to design the speech recognizer to recognize the phonemes of Bodo languages. The MLP consist of 72 input nodes, variable number of hidden nodes as well as layers and 6 output layers. To train the MLP, a modified version of well-known **Back Propagation Algorithm** [12] has been used. The algorithm used for training the network and testing the recognition result is organized in the following steps:

Step 1: Preparation of the Input Matrix: The feature vector extracted from feature extraction block is converted to a single column and is appended on the previous column of the input matrix.

Step2: Preparation of The Target Matrix: The target is prepared by observing the input matrix. We have assigned the same target for a single vowel sound. The target varies from vowel to vowel.

Step3: Training: A feed forward multi-layer perceptron network is formed with 72 neurons with 'logsig' transfar function in the input layer, different numbers of neurons with 'tansig' transfar function in the hidden layer. The number of neurons in the output layer for one person is 6, and it varies with the number of persons in the training set. The network is trained with 'traingdx' training function which uses gradient descent with momentum and adaptive learning rate back propagation. The adapt function is 'trains', init function is 'initlay', performance function 'mse'. We are setting the maximum training epoch to be 300.

Step4: Testing and Classification: We have arranged testing in three steps. One sound from each person is used in testing which is not used in training. Then testing is done using some vowel-sounds of unknown persons. Utterance from the training set is also used for testing.

6. OBSERVATIONS AND RESULTS

The architecture of the recognizer has been shown in the Table 1.The Table 2, Table 3 and Table 4 show the recognition accuracy when the system is tested K-means clustered data for Bodo phonemes using Formant Frequency, LPC Coefficient and MFCC as feature vectors respectively. It has been found that with the increasing number of hidden nodes and layer, the LPCC-based recognizer provides better recognition accuracy. But considering all bottlenecks, LPCC-with 3layers (2-hidden and 1-output) of 8th number type of recognizer given in the Table 1 has been found to be optimal for the recognition of Bodo phonemes.

Table 1: Configuration of MLP-based Speech Recognizer

Type	Input Node	Output Node	Hidden Node at 1st Layer	Hidden Node at 2nd Layer	Hidden Node at 3rd Layer
1st Type	72	6	6	------	------
2nd Type	72	6	20	-------	------
3rd Type	72	6	50	--------	-------
4th Type	72	6	20	6	-------
5th Type	72	6	6	20	
6th Type	72	6	50	20	--------
7th Type	72	6	50	20	6
8th Type	72	6	50	20	10

Table 2: By using feature vector Formant Frequency after clustering, reorganization of Bodo phonemes with different types of MLP-based speech recognizer.
(Average of 100 experiments for each phoneme)

Recognizer Type	1	2	3	4	5	6	7	8
Efficiency of Recognization (Testing with the same Dataset already used as Training Dataset)	52	54	56	59	60	69	77	83
Efficiency of Recognization (Testing with another Dataset Taken from The Same Informants Already whose Sounds used in training)	40	40	46	48	48	58	66	71
Efficiency of Recognization (Testing with Unknown Informants)	35	40	46	48	46	52	58	61

Table 3: By using feature vector LPC Coefficient after clustering, reorganization of Bodo phonemes with different types of MLP-based speech recognizer.
(Average of 100 experiments for each phoneme)

Recognizer Type	1	2	3	4	5	6	7	8
Efficiency of Recognization (Testing with the same dataset already used as training data set)	64	68	77	81	85	90	91	94
Efficiency of Recognization (Testing with another dataset taken from the same informants already whose sounds used in training)	55	59	62	62	69	73	79	85
Efficiency of Recognization (Testing with unknown informants)	42	46	47	55	63	66	69	74

Table 4 : By using feature vector MFCC after clustering, reorganization of Bodo phonemes with different types of MLP-based speech recognizer.
(Average of 100 experiments for each phoneme)

Recognizer Type	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Efficiency of recognization (Testing with the same data set already used as training dataset)	60	63	74	80	83	86	88	90
Efficiency of recognization (Testing with another dataset taken from the same informants already whose sounds used in training)	50	53	58	60	63	67	72	78
Efficiency of Recognization (Testing with unknown informants)	40	43	44	47	53	56	61	66

The present study reveals the following typical features of the recognizer:

(i) The testing arrangement is done in three steps:
one is testing with the same data set through which data set training of the network was done, another one is the data set taken or collected from the those informants whose datas are already given to the network for training purpose, and other one is totally new data taken from new informants. Out of these three datasets, highest accuracy is gained for the 1^{st} type of data set.

(ii) The 8^{th} type recognizer is the best recognizer if the input of the ANN is given through Formant Frequency, LPC coefficient and MFC cepstral coefficient.

(iii) The higher recognition efficiency is more pronounced in case of LPC-cepstral feature vector as input than the Formant Frequency and MFC- cepstral feature vector as input.

REFERENCES

[1] V. Estivill-Castro. "Why so many clustering algorithms: a position paper". SIGKDD Explorations, 4(1):65-75, 2002
[2] Meila, M., and Hecherman, D., "An experimental comparison of modelbased clustering methods". Microsoft Research, Machine Learning, 42:9 – 29, 2001
[3] Halkidi,M.,Batistakis,Y.andVazirgiannis,M.,"Clustering algorithms and validity measures". IEEE Transactions on pattern analysis and machine intelligence, 24 (12), 2002
[4] Proakis,J.G. and Manolakis, D.G., "Digital Signal Processing Principles, Algorithms and Applications', Pearson Education, Third Indian Reprint **2004**
[5] Rabiner, L. and Juang, B., "Fundamentals of Speech Recognitions", Prentice Hall, New Jersey, ISBN 0–13–015157–2,1993
[6] Rabiner, L.R.,Pan,K.C. and Soong,F.K., 'On the performanceof isolated word speech recognition using vector quantization and temporal energy contours",AT& Bell Lab.,Tech.J.Vol63,pp-1245-1260,September,1984
[7] Tohkura, Y., "A weighted cepstral distance measure for speech recognition", Proceeding of IEEE intl. Conf. Acoustics, Speech and Signal processing', Tokyo, Japan, pp.761-764, April, 1986
[8] Furui, S., "Speaker–independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoustics, Speech and Audio processing, Vol. 34, pp. 52-59, Feb., 1986
[9] Hongwu, Y., Lianhong CAI, "Perceptually Weighted Mel–CepstrumAnalysis of Speech based on Psychoacoustic Model", IEICE TRANS. INF. & SYST., Vol.E89–D, NO.12 December 2006
[10] Deka, M.K., Talukdar, PH,Singha,M.K., 'LPCC based sex identification through the utterance of Bodo Phonemes' , j.Of Assam Sc.Society ,Vol.49,December 2008,pp 42-48
[11] Fukada, T., Tokuda, K., Kobayashi, T. and Imai, S., "An adaptive algorithm for mel-cepstral analysis of speech," In Proc. ICASSP-92, pp. I-137-I-140, 1992
[12] Choudhury, S.: "A Statistical Analysis of Speech Dependent/Independent Pattern Congruity of Assamese and Bodo Phonemes", a Ph.D. Thesis