International Journal of Research in Signal Processing, Computing & Communication System Design
5 (1), 2019, 49-53
http://www.publishingindia.com/ijrsp

Spoken English Digit Classification Using Supervised Learning

Maddimsetti Srinivas[1], Kasiprasad Mannepalli[2] and G. L. P. Ashok[3]

[1]Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur(Dt), Andhra Pradesh, India.
Email: maddimsetti34@kluniversity.in

[2]Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur(Dt), Andhra Pradesh, India.
Email: mkasiprasad@kluniversity.in

[3]Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur(Dt), Andhra Pradesh, India.
Email: glpashok@kluniversity.in

Abstract: Multiclass classification is a fundamental problem for many speech recognition systems. Spoken digit recognition is a multiclass problem of 10 classes. Present paper using Support Vector Machine (SVM) and K-Nearest-Neighbour (KNN) and Ensemble method i.e., Random Forest (RF) to English digit classification. Caffe speech dataset of 2400 input instances (15 speakers*16 repetitions*10 digits) used for experiments. Mel Frequency Cepstral Coefficients (MFCC) features are formed for all input instances. The dataset is divided into training set and testing set with 10%, 30% and 50% of dataset as testing set. Confusion matrices generated with all test cases for all classification methods. Performance of Ensemble method is high compared to SVM and KNN at different number of frames. The highest accuracy achieved by RF method is 97.50% by taking 10% testing data.

Keywords: Caffe, Ensemble methods, KNN, MFCC, Random Forest (RF), Spoken english digit, SVM.

I. Introduction

Spoken Digit Recognition (SDR) widely used in telephonic conversation, patient requirement analysis, business and general election surveys. Some of the works on SDR are Multiple Abductive Network Classifier Strategy [2], classification Bengali spoken digits using MFCC features and HMM classifier [3], KNN and MFCC to classify Pashto digits [4], Portuguese spoken digits classification using Line Spectral Frequencies Coefficients (LSFC) features [5], English digit recognition using SVM kernals [6], Hierarchical Temporal Memory [7], Self organizing maps and Perceptual Linear Predictive Coding (PLP) [8], Japanese digit recognition using Two Dimensional Mel Cepstrum (TDMC) features using Neural Networks [9], Arabic digit recognition using Tree distribution classier using MFCC and Vector Quantization (VQ) [10]. Present paper focuses classification of Spoken English digits using SVM, KNN and Bagging classification techniques.

In Section 2, we give an overview of classification algorithms. Section 3 describes the MFCC feature extraction process. In Section 4, we present description of dataset [1], Sections 5 and 6 describe the experimental results and the conclusion and future direction.

II. Supervised Classification Techniques

We compare widely used kernel-based supervised learning algorithms, Support Vector Machines (SVM), K-Nearest Neighbour (KNN) and Bagging for present work. These algorithms can used for other applications, music genre [13] and medical datasets [20].

A. Support Vector Machines

SVM is based on Structural Risk Minimization (SRM) [14] to avoid overfitting. SVM is a supervised machine learning technique developed by Vapnik [15]. Consider an input set X, and a feature space F, there exists a mapping function $\Phi : X \rightarrow F$. Widely used kernels [16] are linear, polynomial based on order and Radial Basis Function (RBF). Polynomial kernal is used for present work as the performance of polynomial kernal is better than other kernals in Natural Language Processing (NLP) applications.

B. K-Nearest Neighbours

K-NN classifier [4] [17] is based on the distance calculation. Different distances used to calculate the nearest neighbours. Cityblock distance is considered for comparison with other techniques.

Cityblock Distance defined as:

$$d_{i,j} = \sum_{k=1}^{n} |x_{ik} - x_{jk}| \quad (1)$$

where, i,j are i^{th},j^{th} feature vectors, n is the length of the feature vector. KNN is used for classification for medical datasets [18] and the performances of the distances are compared.

1. MFCC feature vectors are feeding to KNN classifier.
2. For each testing instance in the testing set, distance between all training instances is computed [17].
3. Computed distances are sorted in ascending order and first K instances are considered for voting.
4. Apply voting or means according to the application. The label which got the more votes is the correct label (class) for that instance.

C. Ensemble Methods

The idea behind the ensemble method is to consider the aggregation of individual predictor scores instead of individual score. In this ensemble method group of predictors is called ensemble. Predictors can be either classifiers or regressors. Ensemble method can use any of the classifiers as predictors. Dataset is divided into subsets and classification score is calculated for each subset. To classify a test instance, votes from all predictors is considered. If the classifier is decision tree then it is called Random Forest (RF). In this paper, decision tree is considered as classifier. Sampling is the process of selecting subsets from the training set. If it is performed with replacement (bootstrap) then it is called bootstrap aggregating also called bagging [19]. This sampling and training process [12] is represented in Fig. 1.

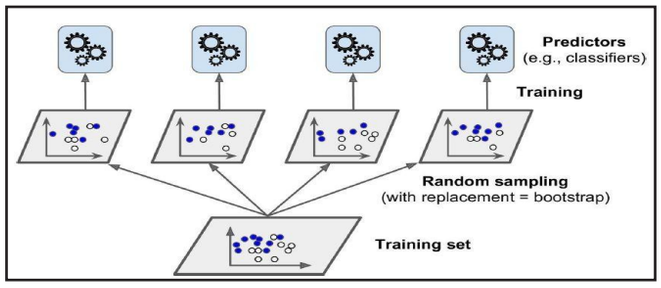

Fig. 1: Bagging Training Set Sampling and Training

III. MFCC Feature Extraction

MFCC features [11] are very successful in speech recognition for many applications. The block diagram is shown in Fig. 3.

Table I: Parameter List

Parameter	Value
Input Signal Length	1500
Frame Length	200
Cepstral Coefficients	13
Frame Shift	80
Sampling Rate	8KHz
Filterbank Channels	20

Table II: Input to Frame Splitting

Input Length	Frames
500	4
1000	11
1500	17
2000	23
2500	29
3000	36

1. Each instance of input is passed through low pass FIR filter for pre-emphasising.
2. Input instance X(n) is divided into frames. There is a overlapping between the frames. Assume X(n) as input instance and Xi(n) is frame representation.
3. To remove the discontinuities and edge effect hamming window is used. Hamming window is processed on input instance. Hamming window equation is shown below:

$$W(n) = 0.54 - 0.46\cos(2\pi/N - 1) \quad (2)$$

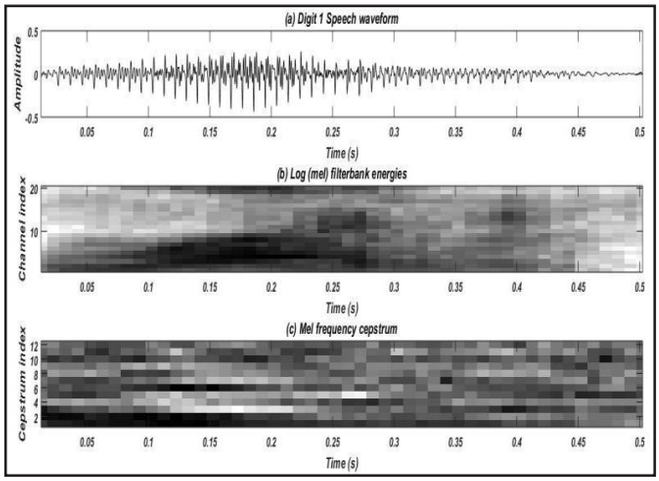

Fig. 2: (a) Digit '1' Wave Form (b) Filter Bank Energies (c) MFCCs

4. Spectral coefficients can be obtained by applying Short Time Fourier Transform (STFT) on each frame.
5. Amplitude spectrum of Yi(n) is multiplied 20 triangular filter banks to take log to resemble with perceptual system [11]. Mel frequency scaling is done as follows:

   $$mel(f) = 1125 * \ln(1 + f/700) \quad (3)$$

6. Discrete Cosine Transform (DCT) is applied on log magnitude spectrum to generate the 13 ceptral coefficients of MFCC for each frame.

   Mel Frequency Cepstral Coefficients are calculated above are used as feature vectors for the classifiers. The parameters are shown in Table I.

IV. Spoken Number Dataset

The dataset contains 2400 instances with 15 speakers. Each speaker utters a digit 16 times which leads to 15*16=240 instances for each digit. This audio / speech dataset downloaded from Caffe [1] maintained and developed by the Berkeley Vision and Learning Center (BVLC). Since the dataset instances of different sizes, padding is used to make the instances of fixed size. The dataset is divided into training data and testing data. For example, if 10% dataset is used as testing data 240 instances are used as testing instances. In the present work, 10%, 30% and 50% of dataset used as testing data.

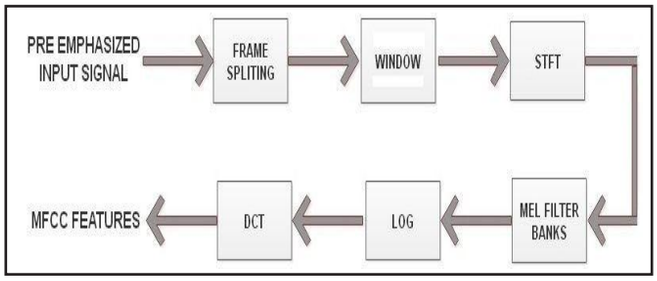

Fig. 3: MFCC Block Diagram

V. Experimental Results

Performance of classifiers is analysed by applying MFCC features vectors of training and testing instances. MFCC parameters used for experimentation are shown in Table I. Each input instance is divided into frames and MFCC coefficients are calculated for each frame. The standard frame duration is 25ms is taken and 10ms is the overlap time with previous frame. Number of samples for one frame depends on sampling rate. For the present work, frame time is 25ms which results 200 discrete samples for a sampling rate is 8KHz. Windowing is the technique to remove the discontinuities in the input instance. Hamming window is applied on each frame to reduce the edge effect. Amplitude spectrum for each frame calculated from Short Time Fourier Transform (STFT) and it is multiplied with the triangular filter bank channels. Log operation is applied on amplitude spectrum to mimic human auditory system. Final step is to get MFCC coefficients by operating DCT on the obtained log values. Each frame gives 13 MFCC features so the feature vector depends on number of frames. For example, if the length of the input is 1500 implies number of frames are 17. Length of the feature vector is 17*13=221. Finally the One-Dimensional feature vector of size 221 represents an a training or testing instance. In a similar way, different input lengths produce the different feature vectors. Comparison of classification techniques is shown in Fig. 4, 5, 6 with 10%, 30% and 50% dataset as testing set. SVM polynomial kernel, KNN cityblock distance and Bagging with decision trees are used as classification techniques. The graphs are between accuracy and number of frames. It can be observed from the Fig. 4, Bagging method using decision trees is called Random Forest (RF) outperformed other two techniques with reasonable margin. During experimentation, SVM poly 4 performed better than other poly orders as shown in Fig. 7. Polynomial kernal accuracy is better than linear and RBF for NLP applications as shown in Fig. 7. KNN cityblock accuracy is higher than other distance metrics. RF method using multiple predictors with subsets of training set depends on multiple predictor votes.

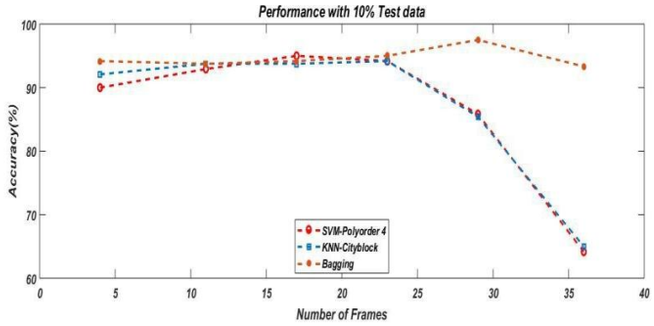

Fig. 4: Accuracy with 10% Test Data

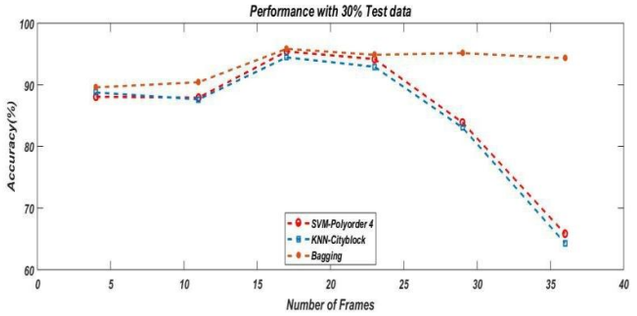

Fig. 5: Accuracy with 30% Test Data

Accuracy of RF method is better even the number of frames are increased where as accuracy of SVM and KNN are reduced steadily.

VI. Conclusion and Future Direction

In the present paper performance of different supervised ML techniques is presented. Experiments carried out with 10%, 20% and 50% of dataset as testing data. It can be concluded that with more training data i.e., with 10% testing data all classifiers achieved high accuracy. From the figures it can concluded that RF method works better as compared to the other two techniques.

In future, present work can be extended using hard voting classifiers to get better accuracy. Present work can guide to solve other speech problems.

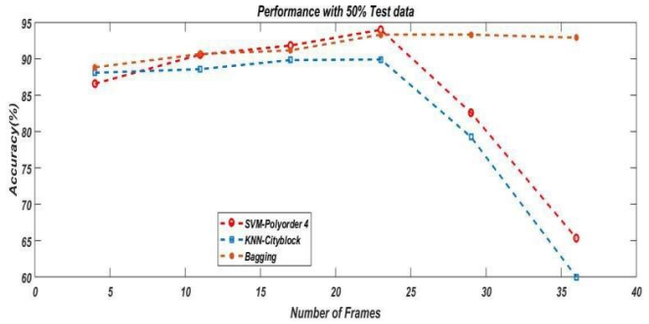

Fig. 6. Accuracy with 50% Test Data

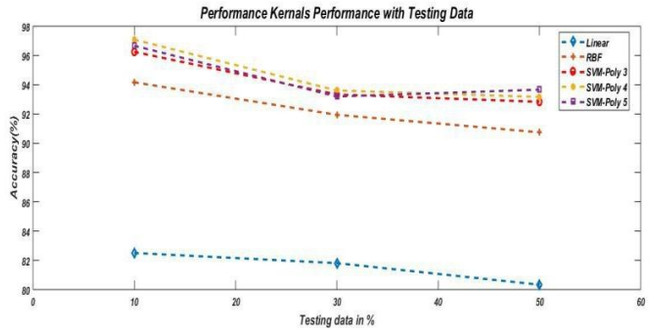

Fig. 7: Testing Set Analysis for SVM Kernals

References

[1] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *Proceedings of the 22nd ACM International Conference on Multimedi*a, pp. 675-678, ACM, 2014.

[2] I. A. Lawal, "Spoken character classification using abductive network," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 881-890, 2017.

[3] G. Muhammad, Y. A. Alotaibi, and M. N. Huda, "Automatic speech recognition for Bangla digits," *2009 12th International Conference on Computers and Information Technology (ICCIT'09)*, IEEE, 2009.

[4] Z. Ali, A. W. Abbas, T. M. Thasleema, B. Uddin, T. Raaz, and S. A. R. Abid, "Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 271-275, 2015.

[5] D. F. Silva, V. M. A. de Souza, G. E. A. P. A. Batista, and R. Giusti, "Spoken digit recognition in Portuguese using line spectral frequencies," *Ibero-American Conference on Artificial Intelligence*, Springer, Berlin, Heidelberg, 2012.

[6] I. Bazzi, and D. Katabi, "Using support vector machines for spoken digit recognition," *Sixth International Conference on Spoken Language Processing*, 2000.

[7] J. V. Doremalen, and L. Boves, "Spoken digit recognition using a hierarchical temporal memory," *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[8] F. Diaz, J. M Ferrández, P. Gomez, and V. Rodellar, "Spoken-digit recognition using self-organizing maps with perceptual pre-processing," *International Work-Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg, 1997.

[9] T. Kitamura, S. Ando, and E. Hayahara, "Speaker-independent spoken digit recognition in noisy environments using dynamic spectral features and neural networks," *Second International Conference on Spoken Language Processing*, 1992.

[10] N. Hammami, and M. Sellam, "Tree distribution classifier for automatic spoken Arabic digit recognition," *2009 International Conference for Internet Technology and Secured Transactions (ICITST 2009)*, IEEE, 2009.

[11] B. Logan, "Mel frequency cepstral coefficients for music modeling," *ISMIR*, vol. 270, 2000.

[12] A. Geron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, Inc, 2017.

[13] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133-141, March 2006.

[14] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[15] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science and Business Media, 2013.

[16] B. Scholkopf, C. J. C. Burges, and A. J. Smola, "Advances in kernel methods: Support vector machines," 1998.

[17] Z. Jan, M. Abrar, S. Bashir, and A. M. Mirza, "Seasonal to inter-annual climate prediction using data mining KNN technique," *International Multi Topic Conference*, Springer, Berlin, Heidelberg, 2008.

[18] L.-Y. Hu, and M.-W. Huang "The distance function effect on k-nearest neighbor classification for medical datasets," *SpringerPlus*, vol. 5, no. 1, p. 1304, 2016.

[19] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[20] M. Khondoker, R. Dobson, C. Skirrow, A. Simmons, and D. Stahl, "A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies," *Statistical Methods in Medical Research*, vol. 25, no. 5, pp. 1804-1823, 2016.