



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Spoken Keyword Based Speech Recognizer for Extracting Information

Sangharmitra Mohanty¹

Department of Computer Sc & Application

Utkal University

Vani Vihar, Bhubaneswar

INDIA

Basanta Kumar Swain²

Department of Computer Sc & Engineering

Government College of Engineering

Kalahandi, Bhawanipatna

INDIA

Abstract: This paper investigates multilayer feed forward back-propagation artificial neural network hybridized with ontology, a knowledge engineering tool for developing speech recognizer that will assist to retrieve Indian tourism information in hands free mode. Spoken key words are collected from 30 speakers of both genders in Indian tourism domain. This paper also used ontology technique a knowledge engineering approach in order to reduce the search space during information extraction. Thirteen Mel frequency cepstral coefficients have been applied to represent speech signal. Features from the selected frames are used to train multilayer FFBPNN. The same routine is applied to the speech signal during the recognition stage in such a way that unknown test patterns are classified to the nearest patterns and retrieved. The system can be said to be robust as average accuracy for clean data is 88.6% % while that for noisy data is 67.7 %.

Keywords: Automatic Speech Recognition (ASR); Ontology; Mel Frequency Cepstral Coefficient (MFCC); Feed-forward back-propagation; Neural networks

I. Introduction

Tourism has emerged as an instrument for employment generation, poverty alleviation and sustainable human development. Tourism promotes international understanding and gives support to local handicrafts and cultural activities. It is an important segment of the country's economy, especially in terms of its contribution towards foreign exchange earnings, generation of additional income and creation of employment opportunities. Apart from the religious places, beautiful temples and artistic sculpture, there are several other spots of scenic beauty that attract tourists from far and near. In spite of the presence of such picturesque tourist spots and places of historical importance, tourism in India has occupied a backseat. Tourist arrival in the state has kept a very low profile. There are many reasons for low foot steps to Indian tourism places but one of them is lack of information for tourists specially for physically challenged. Now we are doing research by utilizing extensively both information technology and knowledge engineering technique to avail the Odisha tourism information in hands free mode. Our work not only looks into the development of a tool that acts as an aid to tourism domain of India but also uses natural mode of communication called speech that skills learnt during early childhood, without instruction and rely on speech communication throughout our lives. We can demand that communications to machine are more comfortable via speech than primitive interfaces such as keyboards and pointing devices[1,2,3].

Our objective is to retrieve the India tourism information accurately by developing an automatic Odia speech recognizer, in another angle automatic speech recognizer (ASR) will be more flexible for tourists those are illiterate or physically challenged. Automatic speech recognizer performs as expert systems, a technology whereby computers can act as a substitute for a human expert [4]. An intelligent computer that acts, responds or thinks like a human being can be equipped with an automatic speech recognition module that enables it to process spoken information. The goal of this research paper is to develop a human-machine interface (HMI) that can classify spoken words under normal conditions and speech that is distorted in some manner due to noise. In this work we emphasize on natural language like Odia for Human-Computer interaction (HCI) by using Odia ASR that uses mel cepstral coefficients as feature vector and feed forward back propagation neural network as a classifier to aid information exchange. We have also tried to combine two technologies, a knowledge engineering technique called ontology with FFBPNN classifier for reducing the search time and retrieving more

accurate tourism information in short span of time. The speech corpus is collected from 30 speakers, some of them are native to Odisha and few are non-native but can speak Odia. The ongoing research work used two groups of people for accommodating the accent variations in Odia ASR. If accent variations attributes is accustomed to ASR, it can be used by different regions of tourists for extracting the information related to tourist places of India. Odia is an Indo-Aryan language spoken by about 31 million people mainly in the Indian state of Odisha, and also in West Bengal, Jharkhand, and Gujarat. Odia is closely related to Bengali and Assamese. This paper is organized as follows, Sect 2 describes tourism database used in our research work, Sect. 3 deals with feature extraction from recorded speech. In Sect. 4 research methodology is presented, Sect. 5 describes the results of our experiment. Finally Sect. 6 concludes the paper and outlines our future work.

II. Tourism Speech Database

In this research paper, we have focused on two tourism states of India i.e. Odisha and West Bengal. Speech data base developments depend on the type of application, nature of application and domain of natural language. This research work is based on isolated speech recognizer in Odia language. The speech data base for isolated ASR is key words in nature. Odia key words are collected from 24 males and 6 females. The key words are based on names of districts or head quarter of district of state Odisha as well as West Bengal and names of tourism places available in each district. We have considered two Indian tourism states and their related tourism places in each state. This research work deals with 1500 isolated key words. The speakers are persons aged between 25-35 years. Out of 30 speakers ten speakers are non-native to Odisha and remaining born in Odisha. Two different categories of speakers are adopted in tourism data base development because the system should be suitable for native and none native tourists at the end. The recordings use a sampling frequency of 16000 Hz, 16bits and mono channel [5]. The recording is done laboratory room with noise cancellation microphone and desktop computer.

III. Feature Extraction(Mel Frequency Cepstrum Coefficients)

In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The feature extraction of speech signal is to distinguish between two different words and treat as same word if one word spoken twice or more, similar to the perception of human being. The reason is that when same phones or sounds are produced by different speakers they have common features and have different features for different uttered phone or words. In our experiment we have used mel frequency cepstral coefficients (MFCC) because it is considered to be the best available approximation of human ear. We extract twelve Mel-frequency cepstral coefficients (MFCC coefficients) and one log energy feature (for a total of 13 features per frame) from the spectrum of the speech.

The first step is pre-emphasis which basically makes one sample in speech influence the next sample by a certain weight. Next the speech signal is divided in to frames. After this Hamming window is applied to each sample in the frame. This step is primarily to have a smooth transition between samples of a frame. Next total energy of each frame is calculated. This energy forms one of the values in a typical 13 feature vector of MFCC coefficients. The next step is to pass the data through Mel filters. Mel filters are triangular equidistant filters in Mel scale, which is a logarithmic scale. Taking logarithm of this we obtain Melspectrum coefficients. The final step in obtaining MFCC is performing discrete cosine transform (DCT) on the Melspectrum coefficients. The output of DCT is Melcepstral coefficients. Usually first twelve of these coefficients are used along with the energy coefficient to form the MFCC vector of length 13 each representing a 30 msec duration of speech at every 10 msec [3,6].

IV. Neural Network as Speech Recognition Methodology

Artificial neural network is recognized as distinct class of design technique for pattern recognition due to its discriminative training methods such as using minimization of squared error loss. Researchers use many types of neural network for broad applications. In this research study, we have used feed forward back propagation neural network (FFBPNN) for retrieving tourism information because they allow to take into account speech features information and to build complex decision regions [7].

A successful pattern classification methodology depends heavily on the particular choice of the features used by the classifier. The Back-Propagation is the best known and widely used learning algorithm in training multilayer feed forward neural networks. The feed forward neural net refer to the network consisting of a set of sensory units (source nodes) that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes. The input signal propagates through the network in a forward direction and on a layer-by-layer basis. Back propagation is a multi-layer feed forward, supervised learning network based on gradient descent learning rule. This FFBPNN provides a computationally efficient method for changing the

weights in feed forward network, with differentiable activation function units, to learn a training set of input-output data. Being a gradient descent method it minimizes the total squared error of the output computed by the network [8].

$$E = \frac{1}{2} \sum_{i=1}^p ||o_i - t_i||^2 \quad (1)$$

Where O_i is produced output and t_i is the target.

Back propagation training begins with setting node weights and offsets to small random values. Then a training example and desired outputs are presented to the net. The actual outputs of the net are calculated, error signals are formed, and an iterative algorithm is used to adapt the node weights - starting at the output nodes and working back to the first hidden layer. An example and desired outputs are again presented to the net and the entire process is repeated until weights converge [9,10].

$$w_{j,k}(new) = w_{j,k}(old) + \nabla w_{j,k} \quad (2)$$

The aim is to train the network to achieve a balance between the ability to respond correctly to the input patterns that are used for training and the ability to provide good response to the input that are similar, called test set. FFBPNN architecture is show in the Figure 1.

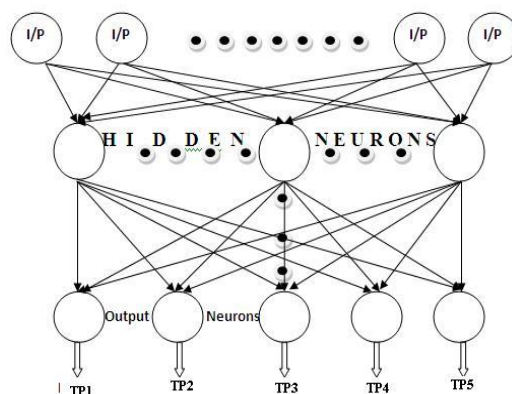


Figure 1. Architecture of FFBPNN

This research work is based on multiple processes. Individual process is carried out in sequential fashion to accomplish the goal. Odia key word based speech recognizer for information extraction in tourism domain is shown in the Figure 2.



Figure 2. Block diagram of Odia key word based speech recognizer

V. Results

Odia key word based speech recognizer for information retrieval in Indian tourism sector is based on feature vector as MFCC and pattern classifier as FFBPNN. The end user has to utter an isolated word that represents a tourist state of India for which s/he is interested to retrieve important information related to intended location. The key word based Odia ASR takes the isolated speech as input and converts the acoustic signal to thirteen MFCC feature vectors. The first coefficients give information regarding the energy concentration in each band. Feature vectors were extracted on 30 ms speech segments in every 10ms .Before passing the thirteen MFCC features values to FFBPNN block of Odia ASR, it is passed through the another sub-block of classifier called ontology which acts as a knowledge engineering tool. Ontology technique is used in this research article because as we know that, it is easy to discriminate among a small set of words, but error rates naturally increase as the vocabulary size grows. We have 1500 isolated key words in our Odia speech database and if we apply the pattern matching algorithm (FFBPNN) directly over speech corpus then it is going to consume maximum time as the search space is very huge. Therefore, the proposed solution is to apply FFBPNN with knowledge based approach called ontology. Ontology specifies a conceptualization of a domain in terms of concepts and their relationships. It (ontology) is recognized as a fundamental infrastructure for approaches to knowledge engineering. Ontology is not only reduces the search time but also very helpful in eliminating ambiguity in domain of information retrieval. We have represented the India tourism domain knowledge in terms of ontology as a tree structure. The root of tree is assumes as Indian Tourism and internal nodes as different states of India.

The next level nodes represent the districts/head quarters as tourist location of corresponding state. In this article we have considered two states of India i.e. Odisha and West Bengal. Total ten 10 districts of Odisha are used as tourists locations in this research article namely Bhubaneswar(BBS), Cuttack(CTC),Balasore(BLS), Baripada(BPD),Dhenkanal(DKL),Phulbani(PBN),Bhawanipatna(BPT),Sambalpur(SBP),Keonjhar(KNJ),Puri(PUR) etc. The sub nodes of internal nodes represent different tourist places of a particular district. Similarly, we have considered tourist places of West Bengal. The tree structure ontology is shown in Figure 3.

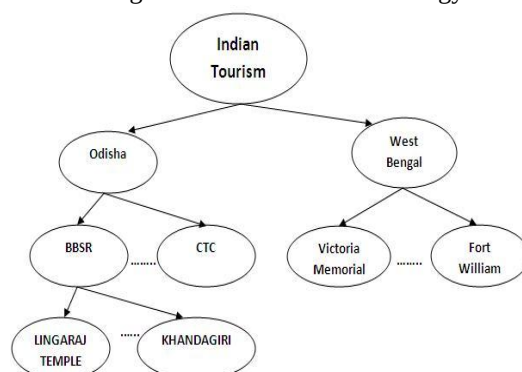


Figure 3. Ontology as tree structure

A user who is interested for retrieving information related to Indian tourism has to pass through the two phases. In first phase the end user has to utter the name of states. Odia ASR will take the key word as state name and extracts the MFCC feature vectors. After that pattern recognition technique, FFBPNN will be applied over the first level of ontology and discarding pattern matching over higher levels keys at that instance. Once the classifier found the a suitable matching, ASR retrieves the list of tourist places belong to that uttered state name. In second phase user has to again utter another key word displayed to him in order to retrieve the refined information of a tourist place. This stage gives emphasize on the pattern searching of the children nodes of found pattern in the first phase and also completely avoids unnecessary pattern matching with siblings of pattern found in first stage. Therefore, the ASR based on classifier as FFBPNN and knowledge engineering tool as ontology is helpful in retrieving the Indian tourism information in less search time as well as unambiguously. The performance evaluation of our Odia key words based ASR in tourism domain at normal condition is shown in the Table 1.

Table1. Accuracy rate(%) in normal condition

TP	BBS	CTC	BPD	BLS	DKL	PBN	BPT	SBP	KNJ	PUR
BBS	87	0	0	8	0	0	5	0	0	0
CTC	0	92	0	0	2	0	0	0	6	0
BPD	3	0	84	8	0	0	5	0	0	0
BLS	7	0	7	86	0	0	0	0	0	0
DKL	0	0	0	4	93	0	0	0	3	0
PBN	0	0	0	0	2	88	0	0	2	8
BPT	5	0	4	0	4	0	87	0	0	0
SBP	0	0	0	0	0	0	0	93	0	7
KNJ	0	6	0	0	3	7	0	0	84	0
PUR	0	0	2	0	0	6	0	0	0	92

The average performance rate of our developed system in normal condition is found to be as 88.6%.

The performance evaluation of Odia key words based speech recognizer for information retrieval in tourism domain at noisy condition is shown in the Table 2.

The system performance in noisy condition is found as 67.7% in average.

Table 2. Accuracy rate(%) in noisy condition

TP	BBS	CTC	BPD	BLS	DKL	PBN	BPT	SBP	KNJ	PUR
BBS	65	0	10	6	0	6	8	5	0	0
CTC	0	71	0	0	8	0	12	0	9	0
BPD	8	0	55	14	0	8	5	7	0	3
BLS	12	0	9	65	0	5	2	7	0	0
DKL	0	6	0	8	80	3	0	0	2	1
PBN	5	0	6	9	2	66	0	0	2	12
BPT	10	0	8	4	4	3	71	0	0	0
SBP	4	0	7	3	0	4	3	67	0	12
KNJ	0	14	0	0	8	9	0	3	63	3
PUR	0	0	6	0	0	10	2	8	0	74

We have experimented the impact of knowledge engineering tool called ontology over FFBPNN. Fig.4 depicts that the average classification accuracy rate of FFBPNN with ontology is feasibly high in comparison to only FFBPNN.

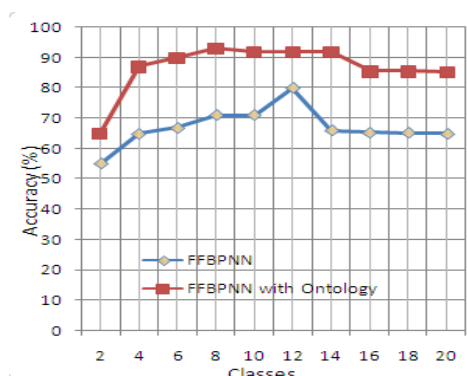


Figure 4. Recognition accuracy rate using FFBPNN and FFBPNN with ontology.

VI. Conclusion and Future work

In this research paper, spoken key word based speech recognizer for Indian tourism sector information retrieval is developed using thirteen mel frequency cepstral coefficient as feature vector and feed forward back propagation neural network as classifier. The computational complexity and search space is reduced by incorporating a knowledge engineering tool called ontology as sub module of FFBPNN pattern recognizer. The performance of the system in normal condition is achieved as 88.6%. But the performance of the system is found as 67.7% in noisy condition. The performance of system can be increased by applying filtering technique as preprocessing block of Odia ASR. The neural network classifier which is used in this experiment provides good accuracy rate. We are planning to extend the work of this study by selecting alternative classifier like Support Vector Machine, Genetic Algorithm, Fuzzy set approaches etc in near future.

VII. References

- [1] Quatieri, T.F., "Discrete-Time Speech Signal Processing Principles and Practice", Pearson Education, Third Impression 2007.
- [2] Juang B. H, "Pattern Recognition in Speech and Language Processing", CRC Press, 2003.
- [3] Rabiner, L.R., Schafer, R.W, Digital Processing of Speech Signals, Pearson education, 1st Edition, 2004.
- [4] Lawrence Rabiner, Biing-Hwang Juang. Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall, pages 333-352 and 434-450, 1993.
- [5] X. Z. Zhang, C.C. Broun, R. M. Mersereau, M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces". *EURASIP Journal on Applied Signal Processing*, Vol. 11, 2002, pp. 1228-1247.
- [6] Mohanty, S, Swain, B.K, "Continuous Oriya Digit Recognition using Bakis Model of HMM" *International Journal of Computer Information Systems*, Vol. 2, No. 1, 2011.
- [7] Beale M H, Hagan M T, Demuth H B, Neural Network Toolbox User's Guide.
- [8] Marques de Sa, J.P, "Pattern Recognition Concepts, Methods and Applications", Springer 2001.
- [9] M. Nakamura, K. Tsuda, J. Aoe, "A New Approach to Phoneme Recognition by Phoneme Filter Neural Networks". *Information Sciences Elsevier*, Vol. 90, 1996, pp. 109-119.
- [10] Frankel, K. Richmond, S. King, P. Taylor, "An Automatic Speech Recognition System Using Neural Networks and Linear Dynamic Models to Recover and Model Articulatory Traces", in *Proceeding ICSLP*, University of Edinburgh, 2000.