International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC) 2011
Proceedings published by International Journal of Computer Applications® (IJCA)

An Approach to Noise Robust Speech Recognition using LPC-Cepstral Coefficient and MLP based Artificial Neural Network with respect to Assamese and Bodo Language

M.K.Deka
Dept of Instrumentation
Gauhati University
Guwahati-14, Assam

C.K.Nath
Dept of Instrumentation
Gauhati University
Guwahati-14, Assam

S.K.Sarma
Dept of Computer Sc.
Gauhati University
Guwahati-14, Assam

P.H. Talukdar
Dept of Instrumentation
Gauhati University
Guwahati-14 Assam

ABSTRACT
In this paper, a new simplified approach has been made for the design and implementation of a noise robust speech recognition using Multilayer Perceptron (MLP) based Artificial Neural Network and LPC-Cepstral Coefficient. Cepstral matrices obtained via Linear Prediction Coefficient are chosen as the eligible features. Here, MLP neural network based transformation method is studied for environmental mismatch compensation. MLP based neural network has been used by many researchers in conjunction with speech recognition, basically for the transformation of the speech feature vectors. In our current study, neural network (MLP) is used to compensate for the environmental mismatch either in feature domain, the model domain, or both. It has been observed that environmental mismatch is automatically compensated without particular knowledge of the environmental interference and retraining. This method can be applied to both linear and non-linear distortion of the speech signal, such as in noisy reverberant speech or telephone speech. Further it can be used for speaker adaptation. By using MLP based neural network, the adaptation processes would require small volume of training data. The Assamese and Bodo are two local languages of North-East India, and they are used as reference languages to carry out this study.

General Terms
Speech Recognition, Language, Computing machine

Keywords
Linear Predictive (LPC) Cepstral Coefficient, Artificial Neural Network, Multilayer Perceptron (MLP), Feature Vector, Assamese Language, Bodo Language

1. INTRODUCTION

Automatic speech recognition involves a number of disciplines such as physiology, acoustics, signal processing, pattern recognition and linguistics. A survey in the robustness issues associated with automatic speech recognition has been reported by several workers [2, 3]. In our present study, the difficulties occurring due to speaker variability and environmental factors are considered. A word may be uttered by the same user in different modes such as in different - emotional levels, health status, surrounding environment (noise/quietness) etc. Again, utterance of the same word varies due to gender, age, dialect, influence of other languages on the speaker etc. Another layer of variation is introduced by the acoustical environment where we use the speech recognizer. These variations are due to background noise, microphone, transmission channel, reverberation etc. In this paper, the problem of environmental variability is addressed, and an attempt has been made to develop speaker and environment independent speech recognition system for the recognition of Assamese and Bodo vowels. Also another attempt has been made to achieve the said objective taking into account other contingencies like – normalization of speech data, volume of training data set and other technical snag which might occur in the process of speech recognition. One simple approach for robust speech recognition in adverse environment is to collect a large volume of samples from all possible environments and train the speech recognizer with all these data. This method will add a little bit of robustness to the speech recognizer, but it's performance is worse than the recognizer that is trained with data collected from one typical environment [5]. Another approach to deal with the problem is retraining, i.e., training of the recognizer in the environment in which it will be used in future. Though the retraining method provides very high level of accuracy, the major drawback of this approach is that at each time whenever the environment

changes, a large volume of data has to be collected to train the recognizer, which is not feasible in practice.

In our study, Assamese and Bodo languages have been taken as the target languages. Assamese language has eight vowels and Bodo language has six vowels, whose chronological architecture changes with gender, age, educational qualification (due to the influence of other language) and locality of the speakers [10]. In this paper, an attempt has been made to develop a speaker and environment independent recognizer for the vowels of both the languages. LPC-Cepstral Coefficient is used as the feature vector of this recognizer [9].

2. THE SPEECH RECOGNITION PROCESS

The proposed speech recognition process has the following three steps:

(a) Digitizing the speech that is to be recognized.

(b) Compute features that represent the spectral-domain content of the speech.

(c) A Multi-layer perceptron (MLP) based Artificial Neural Network is used to classify each set of features corresponding to a vowel utterance into phonetic-based categories.

2.1 Digitization of the speech signal

(a) Speech is first filtered to a bandwidth of 3.4 KHz and then digitized at 8 KHz sampling rate

(b) The digitized speech is then emphasized using a simple first order digital filter with transfer function H (z) = 1 – 0.95 z^{-1}.

(c) The pre-emphasized speech is then blocked into frames of 31.25 ms in length containing 250 samples.

2.2 Computing spectral features

(a) In order to remove the leakage effects and to smooth the edges, each frame is multiplied by a Hamming window as define by equation given below:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi\ n}{N-1}\right), \quad n \in [0, N-1] \quad and \quad N = 250$$

(b) The linear predictive coding (LPC) analysis is then performed on each frame using Levinson-Durbin recursive algorithm.

(c) LPC-cepstral coefficients are computed from the p^{th} order linear predictive coding using the following equations [10]

$$c[1] = a[1]$$

$$\mathbf{c[n]} = \mathbf{a[n]} + \sum_{\mathbf{m=1}}^{\mathbf{n-1}} \frac{\mathbf{m}}{\mathbf{n}}\ a[m].c[n-m], \quad 2 \le n \le p$$

$$\mathbf{c[n]} = \sum_{\mathbf{m=1}}^{\mathbf{n-1}} \frac{\mathbf{n-m}}{\mathbf{n}}\ a[m].c[n-m], \quad n > p$$

2.3 Feature Clustering using Self-Organizing Map

In order to reduce the volume of data without losing the topological information, we use self-organizing map (SOM) to cluster the feature vector into six clusters. Each cluster contains 20 LPC-Cepstral coefficients and the centroid of the clusters is dynamically detected.

2.4 Feature detection

Neural networks have been used in many aspects of speech recognition. They have been used for phonetic classification [11, 12], isolated word recognition [4] and as probability estimator for speech recognizer and as feature detector [6, 7]. A new layered approach is used in the process of designing the Neural Network and the structure of the Neural Node being independent of both the training process and recognition process. It consists of 120 input nodes, variable number of hidden nodes and 14 output nodes. Also, it has the adaptability of changing its transfer function i.e.

Sigmoid [$1+/(1+\exp(-A*value))$], and

Tanh Sigmoid [$a\tanh(b*value)$], Values for A=1, a=1.1759, b=0.6667.

Following the enhanced back propagation algorithm and making use of the sequence training method, the state of the recognizer is stored internally every time its weights are adjusted in order to avoid the inconsistency of weights. The training process is made automated so that both the test set as well as the training set, present to the layered network, stops the training process when the test set satisfies the condition checked at the end of presence of each epoch of training set. At the end of training, the whole state of the layered network is stored for retrieving the state for recognition. Heuristic techniques applied for the network is as follows:

(a) Learning rate has been reduced with the each epoch number

$$\eta(epochNumber) = \eta_0 \exp\left(\frac{-\ epochNumber}{100}\right)$$

(b) Use of Momentum for avoiding the oscillation at the local minima.

(c) Normalization of the input vector.

3. DATABASES

In order to measure the cepstral coefficients, two sets of vocabularies consisting of six Bodo vowels and eight Assamese vowels are used. The database has the following descriptions:

Database-I: 600 isolated utterances of six Bodo vowels spoken by 50 speakers of various age group and gender are taken. Each user uttered the word twice. The sampling frequency is 8 KHz.

Database-II: 800 isolated utterances of eight Assamese vowels spoken by 50 speakers of various age group and gender are taken. Each user uttered the word twice. The sample frequency is 8 KHz.

Database-III (Noisy version of the Database): To create a noisy version of databases, randomly generated Gaussian noises were digitally added to the clear speech of Database-I and Database-II. We create a noisy version of the database by producing 30 dB, 25 dB and 20 dB of noise levels and added digitally to make the database more noisy and robust.

4. EXPERIMENT

A recognizer is built using the clear speech databases, as defined in the Database-I and Database-II. We implement the neural network using Java. Initially we create a neural network with one hidden layer. After training with the clear databases, we evaluate its performance with noisy database as defined in Database-III. In the next level, we increase the number of node in the hidden layer and after training; we evaluate its performance again with Database-III. The same experiment is done with two and three layers of hidden neurons. Again, observations are made with different numbers of neurons in each layer. The configuration of the neural networks and their performance is summarized below:

Learning rate = 0.1
Momentum constant = 0.9
Transfer function = $1/(1+exp(-value))$

Table 1: Configuration of the Neural Network

Type	Input Node	Output Node	Hidden node in the first layer	Hidden node in the second layer	Hidden node in the third layer
Type-1	120	14	21	N/A	N/A
Type-2	120	14	42	N/A	N/A
Type-3	120	14	21	42	N/A
Type-4	120	14	42	21	N/A
Type-5	120	14	21	42	84
Type-6	120	14	84	42	21

Performance of the networks with different configuration is given in Table – (2).

Table 2: Result of the experiment for the recognition of Bodo and Assamese vowels with different types of MLP (200 numbers of experiments are done for recognition of each vowel)

Language	Vowel	Accuracy of Speech Recogniser (%)					
		Type-1	Type-2	Type-3	Type-4	Type-5	Type-6
Assamese	/a/	36	57	71	88	89	89
	/aa/	67	68	15	23	23	96
	/e/	68	07	31	02	24	90
	/ee/	32	22	90	87	84	75
	/i/	74	47	99	04	98	89
	/o/	19	55	76	08	90	90
	/u/	12	64	35	27	40	91
	/w/	03	55	63	26	79	22
Bodo	/a/	34	22	77	81	42	90
	/e/	56	35	31	34	90	90
	/i/	48	31	42	59	53	95
	/o/	34	12	43	35	69	86
	/u/	26	15	64	75	73	88
	/w/	17	38	02	09	67	91

5. RESULTS AND DISCUSSION

From the above experiment, it is clear that neural networks are inherently fault tolerant. This fault tolerant capability of the neural network can be enhanced by proper configuration of the network. The properly configured neural network then can be used as a robust speech recognizer. It is found that in multi-layer perceptron, by increasing the number of hidden layers and nodes, we can get better performance in noisy environment. Since the increase in the number of layers also increases the time requirement for training as well as complexity of the training algorithm, so the proposed scheme sets a limit of maximum number of hidden layers (=3), after which both the training and recognition process become slow.

6. REFERENCES

[1] Dongsuk, Y.: Robust Speech Recognition Using Neural Network and Hidden Markov Models, *Ph. D. Dissertation*, 1999.

[2] Gong, Y.: Speech Recognition in Noisy Environments: A Survey, *Speech Communication*, 16(3):261-291., April 1995.

[3] Junqua, J. and Haton, J.: *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1996.

[4] Lang, k. and Waibel, A.: A time-delay neural network architecture for isolated word recognition, *Neural Networks*, 3:23–43, 1990.

[5] Lippmann, R.; Martin, E. and Paul, D.: Multi-Style Training for Robust Isolated-Word Speech Recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 705-708, April 1987.

[6] Morgan, N. and Bourlard, H.: Neural networks for statistical recognition of continuous speech, *Proceedings of the IEEE*, 83(5):742–770, May 1995.

[7] Renals, S.; Morgan, N.; Bourlard, H; Cohen, M. and Franco, H.: Connectionist probability estimators in HMM speech recognition, *IEEE Transactions on Speech and Audio Processing*, 2(1):161–173, January 1994.

[8] Sorensen, H.: A Cepstral Noise Reduction Multi-Layer Neural Network, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1:933-936, May 1991.

[9] Talukdar, P.H.; Bhattacharjee, U.; Goswami, C.K. and Barman, J.: Cepstral Measure of Bodo Vowels Through

LPC-Analysis, *Journal of the CSI*, Vol. 34 No-1, Jan-Mar, 2004.

[10] Talukdar, P.H.; Bhattacharjee, U.; Goswami, C.K. and Barman, J.: Formants Analysis of Bodo/Boro Vowels with References to Japanese Vowels, *Prajna* Vol. XII, 56-64, 2002-2003.

[11] Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K. and Lang, K.: Phoneme recognition using time-delay neural networks, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, March 1989.

[12] Yuk, D.: A study on Korean phoneme recognition, *Master's thesis*, Korea University, 1993.