ENHANCEMENT OF SPEECH RECOGNITION SYSTEM USING HIGHER ORDER LINEAR PREDICTIVE CODING OVER CONVOLUTIONAL NEURAL NETWORK

Sudha Sharma[1], Ruchi Singh[2], Astha Gautam[3]

[1,2,3]Department of Computer Science and Engineering, L.R.I.E.T. Solan, HPTU, (India)

ABSTRACT

This paper presents a high order linear predictive coding based speech recognition framework in comparison to the artificial neural networks (ANNs) with a multiple-layer deep architecture. In the proposed process, a large training set ensures a powerful modeling capability to estimate the complicated nonlinear mapping from observed noisy speech to desired clean signals. Acoustic context was found to improve the continuity of speech to be separated from the background noises successfully without the annoying sound artifact commonly observed in conventional speech recognition algorithms. A series of experiments were conducted under different sample training with iteration of simulated speech data, resulting in a good generalization capability even in mismatched testing conditions. When compared with the DNN-based algorithm approach, the proposed high order linear predictive coding algorithm tends to achieve significant improvements in terms of various objective quality measures. Furthermore, in an objective preference evaluation more number of instances was found to perform better speech recognition to those obtained with other conventional technique. The results are presented in support of the proposed method.

Keywords: Artificial Neural Networks (ANNs), Discriminative Likelihood Linear Transform (DLTR), Mean Opinion Score (MOS).

I. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential Information areas and remote access to computers. Due to the limited space, we will only test our system on a mall speech database. But one can have many database files for training the system; the more files one train/teach to the system, the more accuracy is achieved. Measuring speech quality constitutes an important task for evaluating many recent speech applications such as telephony, telephony over IP, coding, watermarking, speech enhancement, etc. Traditionally, user's opinions are measured using slow and costly subjective listening tests. In this test, listeners rate the speech they heard on a five-point opinion scale, ranging from 'bad' to 'excellent'. The ratings are unsigned integer scores ranging from 1 for ‘bad’ to 5 for ‘excellent’. Then an average of these scores is computed and defined as the well-known Mean Opinion Score (MOS). It is widely used to characterize the quality of the speech. As an alternative to subjective measurement, an automated ‘objective’ criterion provides a rapid and economical way to estimate user opinion and makes it possible to perform real-time speech quality measurement.

The key to automated machine understanding of audio data is description of its content. Audio descriptions are generally provided in terms of the distinct, identifiable sound units detected in the recordings. Traditionally, these units have been human-identifiable acoustic events, detectors for which are learned from annotated data. Needless to say, these events must be from a finite vocabulary of events for which it was possible to train detectors – one might anthropomorphize this to state that the automated system only detects the events it is familiar with. Contrast this with human (and possibly other animal) listeners. While we definitely do identify sound events that we are familiar with, we are often also able to detect the occurrence of sound events or acoustic “objects” that we have never encountered earlier, based only on how the phenomenon stands out against the background – an ability that greatly enables us to form our own vocabulary of sounds from repetitions of the detected novel phenomena [6]. Cognitively, it is argued by several researchers, there may be an underlying model of “objectness” that human (or animal) listeners subscribe to, and that we are able to detect the occurrence of acoustic phenomena that conform to this notion of objectness, even when we are unfamiliar with the event itself.

To reduce the effect of mismatch, various techniques have been proposed, which can be broadly categorized as:

1. Noise estimation and filtering that reconditions the speech signal or reconstruct speech features based on noise characteristics.
2. On-line model adaptation to reduce the effect of mismatch in training and testing environments
3. Extraction of speech features robust to noise, including features based on human auditory and perception modeling.
4. Normalization techniques to compensate for the channel effect and speech production variations including cepstral mean normalization and vocal track length normalization.
5. Adaptation of acoustic model parameters to a specific speaker based on some criteria, including MLLR, constrained MLLR, maximum a posteriori (MAP) speaker adaptation , and Discriminative Likelihood Linear Transform (DLTR).

II. RELATED WORK

Anis ben aicha et al. [1] In this paper, a new methodology to estimate MOS score of denoised speech is proposed. Experimental results show that the proposed method leads to more accurate estimation of the MOS score of the denoised speech. Statistical analyses of objective criteria are presented to select the more relevant ones for the specific case of denoised speech assessment. The novel proposed methodology to assess denoised speech consists of estimating the subjective MOS score using an Artificial Neural Network as a pattern recognition tool. For each assessed speech signal, scores obtained from selected objective criteria are used as a set of labels for the designed ANN.

Anurag Kumar, Rita Singh and Bhiksha Raj et al. [2] In this paper we explore the idea of defining sound objects and how they may be detected. Our definition tries to confirm to notions present in human auditory perception. Our experimental results are promising, and show that the idea of sound objects is worth pursuing and that it could give a new direction to semi- supervised or unsupervised learning of acoustic event detection mechanisms. Through very simple characterization techniques we are able to achieve a reasonable performance in detection of sound objects. Although using different window sizes in the current setup does not show any remarkable change in the performance we do believe that since we are dealing with a general concept viz. sound objects we need to check with different window sizes to see which is best suited. This might actually be visible in different characterization techniques. The best performance among all is achieved on window of 20ms. This might be attributed to the fact that small window will spread even very short lasting sound objects over multiple frames and thus help in detecting some signature characteristics. The most important conclusion drawn from this work is that we can consider the generic concept of sound objects as a complement to target class- driven acoustic event detection. Even though there was no instance of test acoustic events in training, we are still able to get a reasonable performance using the very simple strategy shown in this paper, validating this idea. This might, in turn enable us to build up vocabularies of sound objects or events for audio annotation – a process that may in fact mimic how we as humans ourselves gain our vocabularies. Semantics may be obtained through the similarity between detected objects and their association with information from other modalities. This could also be used in hierarchical analyses, or as a bootstrap for obtaining human labels, to reduce annotation costs.

Kehuang Li y Zhen Huang You-Chi Cheng Chin-Hui Lee et al. [3]Proposed a maximal figure-of-merit (MFoM) learning framework to directly maximize mean average precision (MAP) which is a key performance metric in many multi-class classification tasks. Tested on both automatic image annotation and audio event classification, the experimental results show consistent improvements of MAP on both datasets when compared with other state-of-the-art classifiers without using MMAP. In this paper, we have presented a technique to maximize the mean average precision of multi-class multi-label classifiers. By further embedding DNNs into the MAP objective function we gain a flexibility to use nonlinear classifiers to improve the discriminative power of conventional linear classifiers. Experiments on two different datasets show an up to 25% relative MAP improvement. In the future, we will study problems with large size training sets and ways to further improve the MMAP-DNN framework.

Y. Deng, X. Li, C. Kwan, B. Raj, and R. Stern [4] In this paper, we proposed a novel continuous acoustic feature adaptation algorithm for on-line accent and environmental adaptation. Implemented by incremental singular value decomposition (SVD), the algorithm captures local acoustic variation and runs in real-time. This feature-based adaptation method is then integrated with conventional model-based maximum likelihood linear regression (MLLR) algorithm. Extensive experiments have been performed on the NATO non-native speech corpus with baseline acoustic model trained on native American English. The proposed feature-based adaptation algorithm improved the average recognition accuracy by 15%, while the MLLR model based adaptation achieved 11% improvement. The corresponding word error rate (WER) reduction was 25.8% and 2.73%, as compared to that without adaptation. The combined adaptation achieved overall recognition accuracy improvement of 29.5%, and WER reduction of 31.8%, as compared to that without adaptation. In the paper, a feature based adaptation algorithm was proposed for unsupervised continuous speaker and environmental adaptation. Like MLLR, which modifies acoustic models to reduce the mismatch between training and test conditions, feature based adaptation also reduces the mismatch between the intra-phoneme spectral variations that occur as a result of non-nativity in the test data, as compared to those encountered in the training data. The feature based adaptation integrated with MLLR model based adaptation improved the performance even further.

Kilian Q. Weinberger, John Blitzer and Lawrence K. Saul [5] In this paper, Mahalanobis distance metrics for kNN classification by semi definite programming have presented. Our framework makes no assumptions about the structure or distribution of the data and scales naturally to large number of classes. Ongoing work is focused in three directions. First, we are working to apply LMNN classification to problems with hundreds or thousands of classes, where its advantages are most apparent.

George P. Kafentzis2, Theodora Yakoumaki1;2, Athanasios Mouchtaris1;2, Yannis Stylianouet al. [6]in this work, presented an application of an adaptive sinusoidal model, named eaQHM, on the problem of emotional speech analysis and classification. It was shown that different emotional speech styles can be effectively represented by the adaptivity mechanism of eaQHM, yielding very accurate AM-FM decomposition. This was demonstrated through resynthesis of the original speech signal from its AM-FM components and by evaluating the Signalto- Reconstruction Error (SRER).

Ville Hautam¨aki, IsmoK¨arkk¨ainen and PasiFr¨antiet al. [7] present an Outlier Detection using Indegree Number (ODIN) algorithm that utilizes k-nearest neighbour graph. To evaluate the performance of the proposed system, an objective evaluation using 50 locutions was performed. The system detected 72.41% of the segmentation marks, in which, 77.6% were detected with an error less or equal to 10 ms and 22.4% of the boundaries were found with an error between 10 and 20 ms. Digital processing techniques are present in several applications, such as voice mail, automatic voice systems, biometric identification, language identification, voice dialing, residential automation. Voice commands are used, as well as reading systems for the blind. Those applications show the unequivocal contributions of vocal communications between man and machine, and they include the voice recognition and text to speech conversion systems, for which the segmentation system performs and important task. This paper describes a speech segmentation system based on observation of a prosodic characteristic of the voice signal, the energy. To find the segmentation marks, the proposed method initially located the reference frontiers. Then, new reference boundaries are found using an energy encoder that operates in the locution region defined by reference boundaries initially detected. The proposed segmentation system was tested using 50 segmentation locutions. To verify its robustness, the obtained results were compared which segmentation marks achieved using manual segmentation. In general, the system is capable to detect 72.41% of the segmentations marks, in which 77.6% were detected with an error smaller than 10 ms, and 22.4% had and error between 10 and 20 ms, when compared with boundaries resulting from manual segmentation. Furthermore, the system presented, on average, 6, 74 false frontiers and 7, 6 boundaries that were not detected. The proposed segmentation algorithm presents a low complexity. The development does not depend on a robust database to train probabilistic models, as in the case HMM segmentation. Moreover, no prior knowledge of phonetic transcription is necessary for the segmentation. The proposed system has competitive results as compared to usual systems, without the use of a refinement of the obtained results.

III. METHODOLOGY

LPC Feature Extraction

The main motive is to calculate the set of LPC coefficient of filter from incoming speech signal. Some LPC coefficients like amplitude, pitch and filter coefficient are extracted. When some LPC coefficients are calculated for specific features of the signal, then error may be decreased. Linear predictive coding is very effective and it signifies the speech parameter at lower bit. The speech production model uses the vocal track with the sources for voiced and unvoiced sounds.

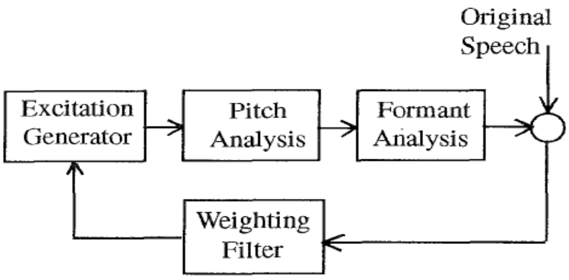


Fig. : LPC Coder

LPC is an approach for analysis of speech. It is widely used technique for encoding the speech quality at less bit rate. This analysis carry a specific speech sample at the present time may be estimated as a linear combination of past speech samples.

The LPC filter is inverse filter that helps to forecast the vocal track features. It is represented by A(z) . The LPC filter is utilized for getting Short Term Predictor (STP) residual and also it helps to extract the format. The FIR form is

$$A(z) = 1 - \sum_{i=1}^{M} a_i z^{-i}$$

Where a_i are M - order LPC coefficients with a_i=l.

It is Long Term Predictor (LTP) and can have ability to predict the feature of vocal chord. It is represented by P(z).

$$P(z) = 1 - \sum_{j=-S}^{S} b_i z^{-(T+i)}$$

Where,

T is long term delay corresponding to pitch period.

b is gain.

S is the number of stages.

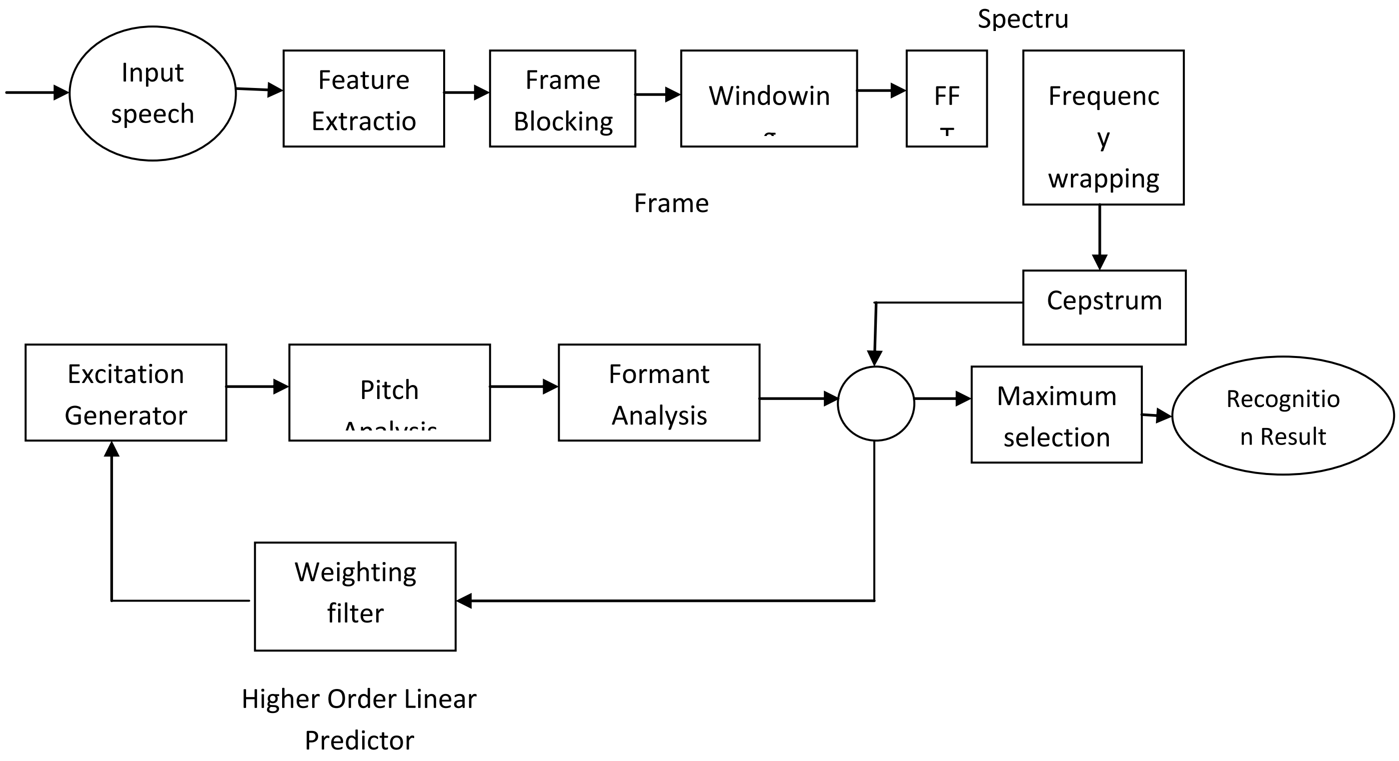
The parameters T & b can be obtained by closed loop method.

LPC analysis is a model which is depending on the human speech production. LP uses conventional source-filter model and in this filter vocal tract, glottal and transfer function are included into all-pole filter which simulates acoustics of the vocal tract. The principle on which LPC analysis works is that it reduces the sum of the squared differences between the new speech signal and the expected speech signal over a finite duration. This might be utilized to produce set of predictor coefficients. These coefficients are given by a_k. In time varying digital filter, the transfer function is given by

$$H(z) = \frac{G}{(1 - \sum akz - k)}$$

Where k lies to 1 to p and it is 10 for LPC-10 algorithm and 18 for proposed algorithm which is used in this work. Levinson-Durbin recursion is used to calculate the essential parameters for the auto-correlation method. In every frame, LPC analysis includes decision making process of voiced or unvoiced. a pitch detecting algorithm is developed to calculate the pitch period. It is significant to compute the gain, pitch and various coefficients which will vary with time from one frame to other frame. Moreover, it is noted that the predictor coefficient is altered to cepstral coefficients (set of parameters.)

Flowchart



IV. RESULTS

The proposed method is compared with the traditional approach using CNN (Convolutional Neural Network) methods. The comparison shows that proposed system attains improvements in terms of recognition. The results are presented below:

- **Input is V**

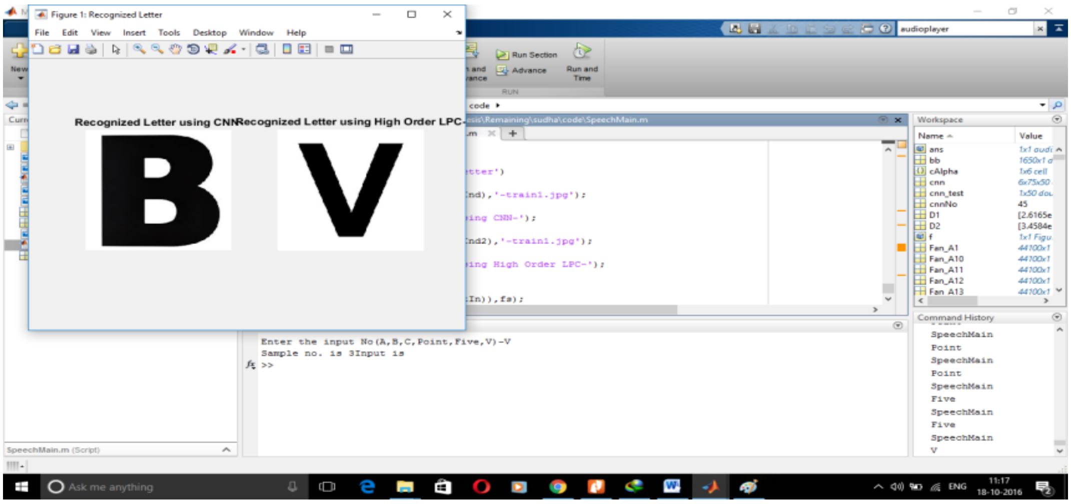
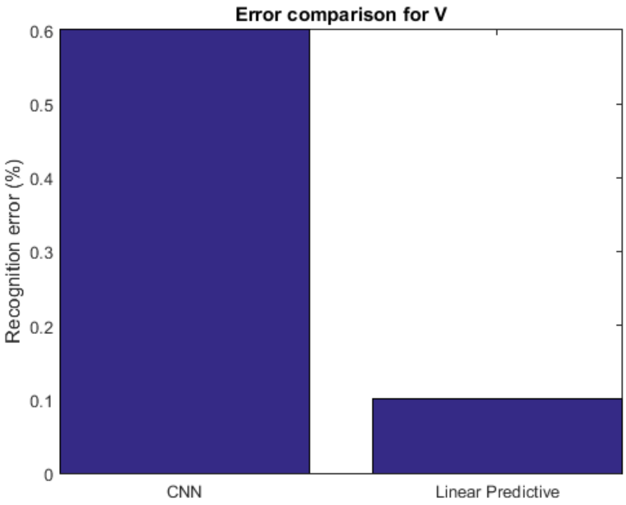


Figure 1: Recognized Letter
File Edit View Insert Tools Desktop Window Help
Recognized Letter using CNN
Recognized Letter using High Order LPC-
B
V
Run Section
Advance
Run and Time
audioplayer
Workspace
Command History

```
Enter the input No(A,B,C,Point,Five,V)-V
Sample no. is 3Input is
>>
```

Recognition errors ratio for V



- **Input is point**

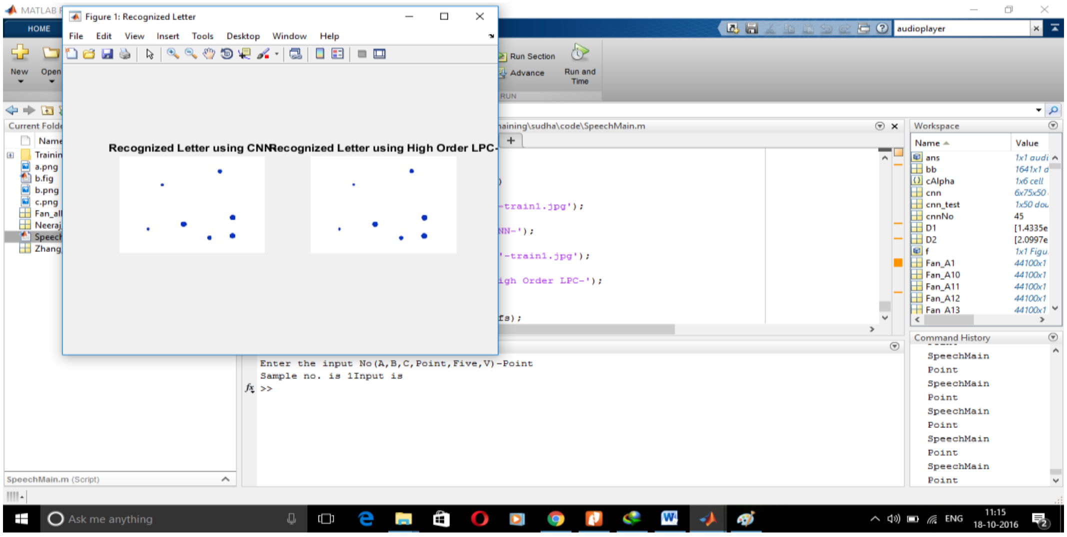
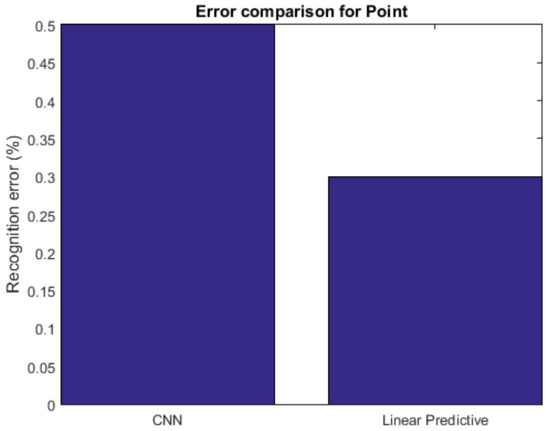


Figure 1: Recognized Letter
File Edit View Insert Tools Desktop Window Help
Recognized Letter using CNN
Recognized Letter using High Order LPC-
Run Section
Advance
Run and Time
audioplayer
Workspace
Command History

```
Enter the input No(A,B,C,Point,Five,V)-Point
Sample no. is 1Input is
>>
```

Recognition error ratio for point :



- **Input is C**

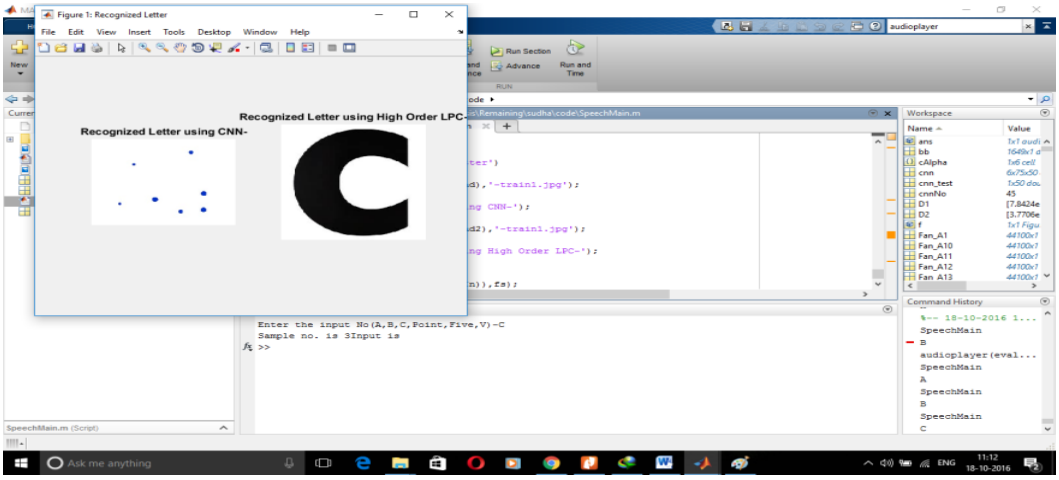
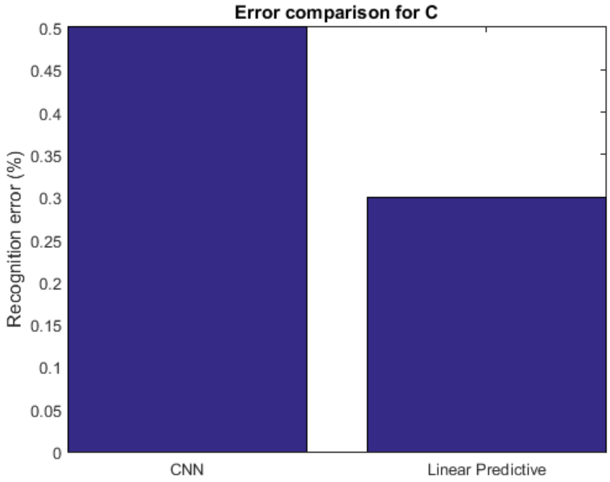


Figure 1: Recognized Letter
File Edit View Insert Tools Desktop Window Help
Recognized Letter using CNN-
Recognized Letter using High Order LPC-
Run Section
Advance
Run and Time
audioplayer
Workspace
Command History
Ask me anything
ENG
11:12
18-10-2016

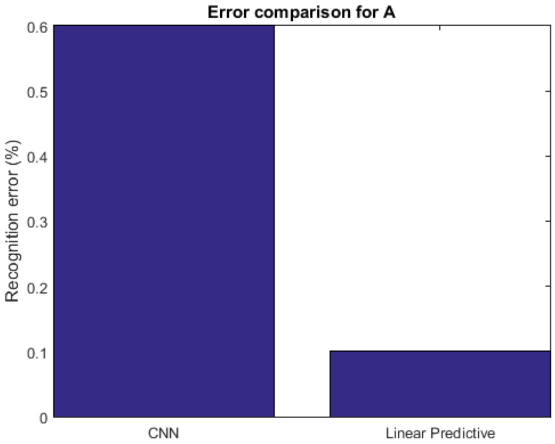
```
Enter the input No(A,B,C,Point,Five,V)-C
Sample no. is 3Input is
>>
```

Recognition error ratio for C



- **Input is A**

Recognition error ratio for A



- **Input is 5**

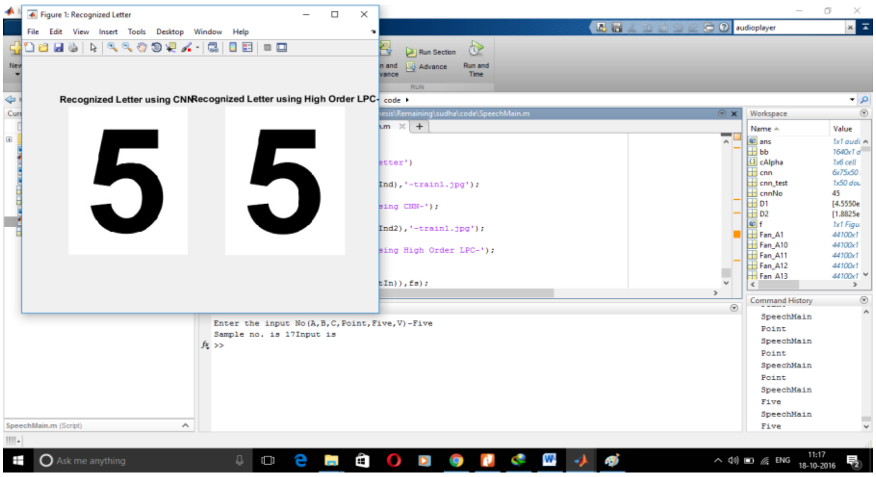


Figure 1: Recognized Letter
File Edit View Insert Tools Desktop Window Help
Recognized Letter using CNN
Recognized Letter using High Order LPC
5
5
Run Section
Advance
Run and Time
RUN
audioplayer
code
esis\Remaining\sudha\code\SpeechMain.m
SpeechMain.m (Script)
Ask me anything
ENG
11:17
18-10-2016

```
etter')
Ind),'-train1.jpg');
sing CNN-');
Ind2),'-train1.jpg');
sing High Order LPC-');
tIn)),fs);
```

```
Enter the input No(A,B,C,Point,Five,V)-Five
Sample no. is 17Input is
>>
```

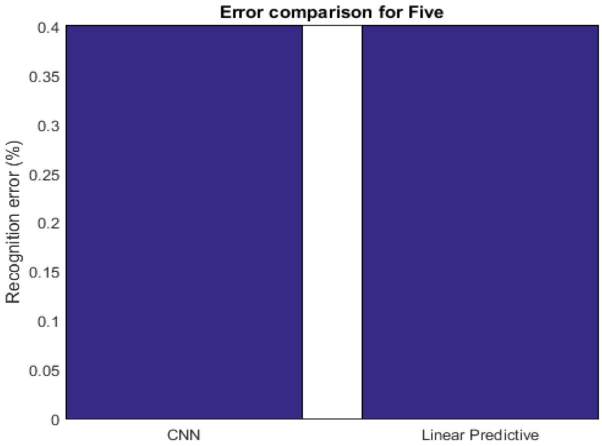
Workspace

Name	Value
ans	1x1 audi
bb	1640x1 d
cAlpha	1x6 cell
cnn	6x75x50
cnn_test	1x50 dou
cnnNo	45
D1	[4.5550e
D2	[1.8825e
f	1x1 Figu
Fan_A1	44100x1
Fan_A10	44100x1
Fan_A11	44100x1
Fan_A12	44100x1
Fan_A13	44100x1

Command History

```
SpeechMain
Point
SpeechMain
Point
SpeechMain
Point
SpeechMain
Five
SpeechMain
Five
```

Recognition Error ratio for five



V. CONCLUSION

In this work, distant conversational speech recognition has been presented. By using the high order LPC. Convolutional neural network acoustic model provides reduced recognition as can be observed from the results. Moreover, it is noted that reduction in recognition is obtained by utilizing a convolution layer within DNN architecture. The activation functions also being used instead of sigmoids. Specifically, the proposed system clearly attempts to optimize the acoustic model for similar syllable sounds. Furthermore, speech enhancement architecture depending on the DNN presented has not been efficient in terms of processing time. A pre-training strategy has been developed in order to initialize the higher order LPC. By utilizing the more acoustic context information, the performance of proposed system is improved.

REFERENCES

[1] Ben Aicha, Anis,On the use of artificial neural network to predict denoised speech quality, In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, pp. 1227-1231.IEEE, 2014.

[2] Kumar, Anurag, PranayDighe, Rita Singh, Sourish Chaudhuri, and Bhiksha Raj, Audio event detection from acoustic unit occurrence patterns, In ICASSP, pp. 489-492. 2012.

[3] Li, Kehuang, Zhen Huang, You-Chi Cheng, and Chin-Hui Lee., A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers, In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 4503-4507. IEEE, 2014.

[4] Deng, Yunbin, Xiaokun Li, Chiman Kwan, B. Raj, and R. Stern, Continuous feature adaptation for non-native speech recognition, International Journal of Signal Processing 3, no. 1 (2006).

[5] Weinberger, Kilian Q., John Blitzer, and Lawrence K. Saul, Distance metric learning for large margin nearest neighbor classification, In Advances in neural information processing systems, pp. 1473-1480. 2005.

[6] Athanasios Mouchtaris, Kafentzis, George P., Theodora Yakoumaki, and YannisStylianou, Analysis of emotional speech using an adaptive sinusoidal model, In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, pp. 1492-1496.IEEE, 2014.

[7] Hautamäki, Ville, IsmoKärkkäinen, and PasiFränti, Outlier Detection Using k-Nearest Neighbour Graph, In ICPR (3), pp. 430-433. 2004.