DOI: 10.21276/sjet.2016.4.7.3

Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2016; 4(7):308-311 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

Robust Phoneme Recognizer at Noise Corrupted Acoustic Environment

Bulbul Ahamed[1], RaselAhmed[2], Khaled Mahmud[3], Mohammad Nurul Huda[4]
[1]Northern University Bangladesh, Dhaka, Bangladesh
[2]Dhaka Residential Model College, Dhaka, Bangladesh
[3]University of Dhaka, Dhaka, Bangladesh
[4]United International University, Dhaka, Bangladesh

***Corresponding author**
Khaled Mahmud
Email: khaled@iba-du.edu

Abstract: This paper proposes a robust automatic phoneme recognizer for Japanese language in noise corrupted acoustic environments. From the previous studies it is found that some hidden factors such as speaking style, gender effects, and noisy acoustic environments degrade the performance of automatic phoneme recognizers (APRs). In this study, an APR is designed in noise corrupted acoustic environments resolving the noise effect. The proposed system comprises three stages. At first stage, a multilayer neural network (MLN) that outputs Distinctive Phonetic Features (DPFs) from the input acoustic features is incorporated, and then the Karhunen-Loeve Transformation (KLT) and the Gram-Schmidt (GS) algorithms are used at second stage to extract reduced feature vector. Finally, the output phoneme strings are generated by inserting the reduced features into a hidden Markov model (HMM) based classifier. It is observed from the experiments in clean and noisy acoustic environments that the proposed method provides higher recognition accuracy at lower Signal-to-Noise Ratios (SNRs).
Keywords: automatic speech recognition; local features; gender factor; phoneme recognizer; hidden Markov model.

INTRODUCTION

Various methods had been proposed to find an automatic phoneme recognizer [1-5]. However, most of these proposed methods embed only hidden Markov models (HMMs) in its architecture and need a higher computational cost to get a large scale performance. Besides, some of them incorporate acoustic features, which produce a narrow acoustic likelihood between two phonemes in noisy acoustic environments and then generate misclassifications.

Therefore, a more accurate phoneme recognizer needs a hybrid classifier with low computation, which incorporates distinctive phonetic features (DPFs) instead of acoustic features. A distinctive phonetic feature (DPF)-based system can model coarticulatory phenomena more easily [6]. In a previous work, a DPF-based feature extraction method was introduced [7], where a multi-layer neural network (MLN) was used to extract DPFs. In [7], a clean acoustic environment was considered for experiments, but no experiments were done in real environments.

This paper proposes an automatic phoneme recognizer in clean and noisy [8, 9] acoustic environments. The proposed system comprises three stages. Firstly, Distinctive Phonetic Features (DPFs) from acoustic features are extracted using a multilayer neural network (MLN). Secondly, a reduced decorrelated feature vector is obtained using the Karhunen-Loeve Transformation (KLT) and Gram-Schmidt (GS) algorithms. Finally, an HMM-based classifier is added at the end of the system to generate phoneme strings for each input speech. It is observed that the proposed phoneme recognition system incorporating KLT provides higher recognition accuracy at lower Signal-to-Noise Ratios (SNRs) on Japanese Newspaper Article Sentences (JNAS) in noisy and clean acoustic conditions.

This paper is organized as follows. Section II discusses the Japanese articulatory features and Section III outlines KLT procedure. Section IV explains the proposed KLT-based technique. Section V describes an experimental setup, and section VI analyzes experimental results. Finally, section VII concludes the paper with some future remarks.

DISTINCTIVE PHONETIC FEATURES

A phoneme can easily be identified by using its unique Distinctive Phonetic Features (DPFs) set [10,

11]. The paper [7] discusses the DPFs set for the Japanese language..

KURHONEN-LOEVE TRANSFORMATION

The KLT, which is closely related to the Principal Component Analysis (PCA) [16], is used here for reducing the dimensionality. The orthogonal basis using the KLT is calculated using the 15 dimensional DPF vector. Then the reduced feature vector with lower dimensionality is obtained by multiplying the input data matrix of size (total number of frames * 15) with orthogonal basis of size 15*11 dimensions. As a result, the size of feature vector is total number of data frames*11, which reduces the total feature vector dimensionality of total number of data frames*15.

PROPOSED KLT-BASED PHONEME RECOGNIZER

The KLT-based phoneme recognition method is proposed in Figure 1. For obtaining LFs as acoustic feature, we convert the input speech into time and frequency domain features [12]. Two LFs extracted by the procedure described in Section IV are used here. LFs are then entered into an MLN with four layers including two hidden layers after combining a current frame xt with the other two frames that are three points before and after the current frame (xt-3, xt+3). The MLN has 45 output units (15×3) corresponding to a set of triphones, or to a context-dependent DPF vector that comprises three DPF vectors (a preceding context DPF, a current DPF, and a following context DPF) with 15 dimensions each. The two hidden layers comprise 256 and 96 units, respectively. The MLN is trained using the standard back-propagation algorithm. This DPF extractor takes 75 (=25x3) LFs as input and outputs a 45-dimensional context-dependent DPF vector. After including KLT and GS algorithm, the system generates a 33-dimensional decorrelated DPF vector for the HMM-based classifier.

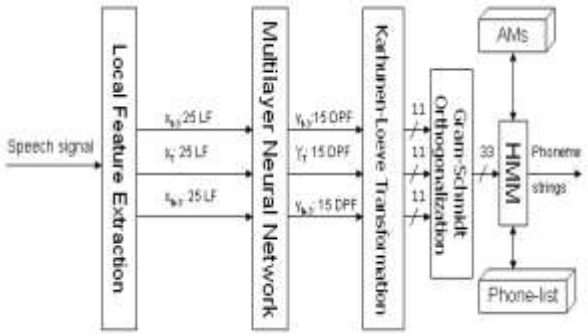

Fig-1: The proposed KLT-based Phoneme Recognizer

EXPERIMENTS

Experimental Database

The following three speech data sets are used in our experiments.

Training data set: A subset of the Acoustic Society of Japan (ASJ) Continuous Speech Database comprising 4503 sentences uttered by 30 different male speakers (16 kHz, 16 bit) is used [13].

Test data set: This test data set comprises 2379 JNAS [14] sentences uttered by 16 different male speakers (16 kHz, 16 bit).

Noisy test data set: Two thousand three hundred seventy nine utterances from JNAS [14] continuous speech sentences uttered by 16 male speakers are used as test data. Test utterances are noise corrupted (car noise) speech. Noise from Japan Electronic Industries Development Association (JEIDA) Noise Database [15] is added to the clean JNAS dataset D2 at different SNR (0 dB, 5 dB, 10 dB, 20 dB) conditions. For each SNR (0 dB, 5 dB, 10 dB and 20 dB), there are 2379 utterances. Sampling rate is 16 kHz.

Experimental Setup

LFs comprised of 25 dimensional (12 Δt, 12 Δf, and ΔP, where P stands for the log power of a raw speech signal) feature vectors.

Since our goal is to design a more accurate phoneme recognizer, phoneme correct rate (PCR) for D2 and D3 data set are evaluated using an HMM-based classifier. The D1 data set is used to design 38 Japanese mono-phoneme HMMs with five states, three loops, and left-to-right models. Input features for the classifier are orthogonalized DPFs. In the HMMs, the output probabilities are represented in the form of Gaussian mixtures, and diagonal matrices are used.

The mixture components are set to 1, 2, 4, 8, and 16. The experiments in clean acoustic environments are given below

(i) DPF(MLN+GS,dim:45)
(ii) DPF(MLN+KLT+GS,dim:33)

To observe PCR in car-noisy environment for different signal-to-noise ratios (SNR=0 dB, 5 dB, 10 dB, 20 dB), we have carried out some experiments using the D3 test data set for the following methods

(i) Car.DPF(MLN+GS,dim:45)
(ii) Car.DPF(MLN+KLT+GS, dim:33)

EXPERIMENTAL RESULT AND ANALYSIS

Table 1: phoneme correct rates in clean acoustic

Methods	Phoneme Correct Rate (%)				
	Mix 1	***Mix 2***	***Mix 4***	***Mix 8***	***Mix 16***
DPF(MLN+GS, dim:45)	77.80	78.06	78.30	78.70	79.19
DPF(MLN+KLT+GS, dim: 33)	79.48	79.12	78.27	78.71	79.29

The phoneme recognition performance in clean environment after applying the GS orthogonalization using the methods DPF(MLN+GS,dim:45) and DPF(MLN+KLT+GS,dim:33) are given in Table 1. From the table it is observed that the phoneme correct rate is increased by the proposed method at mixture components 1 and 2.

Figures 2, 3, 4, and 5 exhibit the phoneme recognition performance in car-noisy environment for the methods DPF(MLN+KLT+GS, dim:33) and DPF(MLN+GS, dim:45) using the SNRs 0dB, 5dB, 10dB, and 20dB, respectively. For 0dB in the Fig. 2, the DPF(MLN+KLT+GS, dim:33) provides a higher recognition performance for all mixture components over the other investigated methods. For example, at mixture component one, the DPF(MLN+KLT+GS, dim:33) gives 46.03% phoneme recognition performance, while the corresponding performance for the method DPF(MLN+GS, dim:45) is 43.42%. For the remaining investigated SNRs (5 dB, 10 dB, and 20 dB), the DPF(MLN+KLT+GS, dim:33) shows a significantly higher phoneme correct rate for lower mixture components with lower SNRs. These improving results are obtained due to the KLT procedure.

The SNR-wise phoneme correct rate for the methods DPF(MLN+KLT+GS, dim:33) and DPF(MLN+GS, dim:45) is shown in Fig. 6 using the mixture component one. For all the investigated SNRs, the method DPF(MLN+KLT+GS, dim:33) shows a higher recognition performance. For example, at SNR 20 dB and clean, the DPF(MLN+KLT+GS, dim:33) gives 79.08% and 79.48% respectively, while 78.06% and 77.80% are shown by the methods DPF(MLN+GS, dim:45).

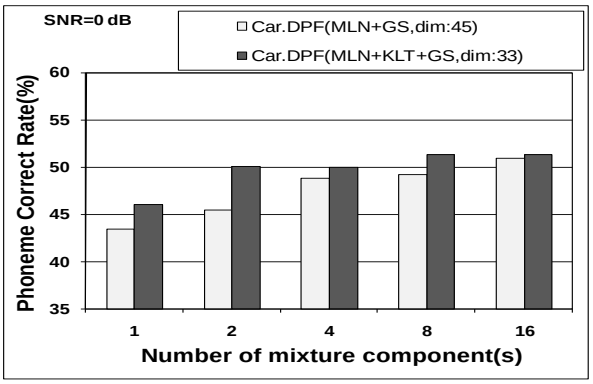

Fig-2: Phoneme recognition performance at 0 dB (car noise)

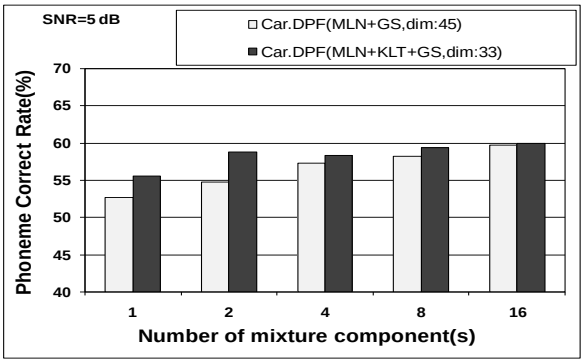

Fig-3: Phoneme recognition performance at 5 dB (car noise)

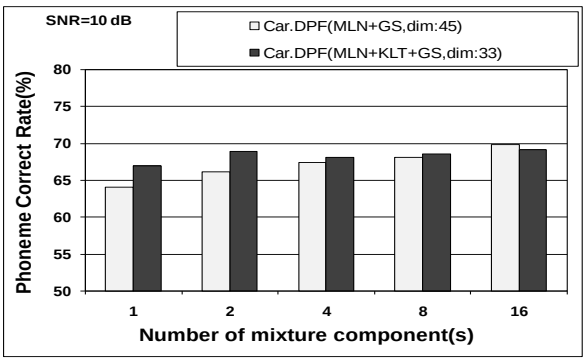

Fig-4: Phoneme recognition performance at 10 dB (car noise)

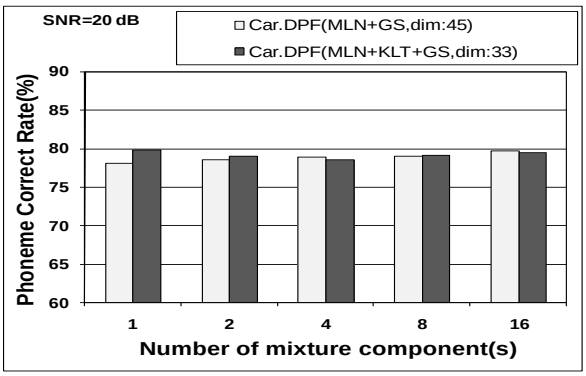

Fig-5: Phoneme recognition performance at 20 dB (car noise)

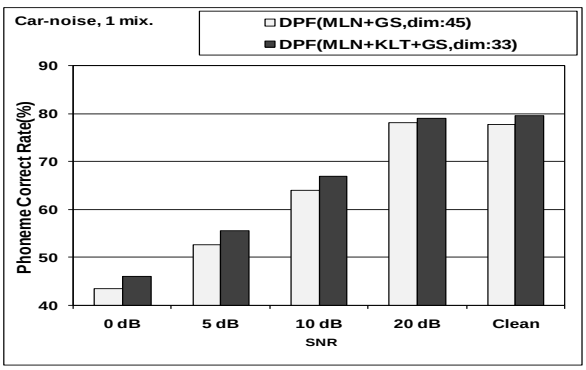

Fig-6: Phoneme recognition performance for different SNRs (car noise, 1 mix.)

CONCLUSIONS

This paper has designed a phoneme recognizer in noisy acoustic environments incorporating Karhunen-Loeve Transformation. The following information concludes the paper.

1) For the KLT-based method, the mixture component two generates the highest level improvement for all SNRs.

2) The proposed KLT-based method has showed the significant improvement of phoneme correct rate in comparison with the method that did not incorporate KLT procedure.

In future, the authors would like to incorporate Recurrent Neural Network (RNN) in KLT-based system for evaluating the performance.

REFERENCES

1. Bazzi, Glass JR; Modeling OOV words for ASR. Proceedings of ICSLP, Beijing, China, 2000; 401-404.
2. Scharenborg O, Seneff S; A two-pass strategy for handling OOVs in a large vocabulary recognition task. InInterspeech'2005-Eurospeech, 9th European Conference on Speech Communication and Technology, 2005; 1669-1672.
3. Kirchhoff K; OOV Detection by Joint Word/Phone Lattice Alignment. ASRU , Kyoto, Japan, Dec 2007.
4. Pepper DJ, Clements MA; Phonemic recognition using a large hidden Markov model. IEEE Transactions on signal processing, 1992; 40(6):1590-5.
5. Merialdo; Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information Training,"Proc. IEEE ICASSP-88, 1988; 111-114.
6. Kirchhoff K, Fink GA, Sagerer G; Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication, 2002; 37(3):303-19.
7. Fukuda T, Nitta T; Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition. IEICE TRANSACTIONS on Information and Systems, 2004; 87(5):1110-8.
8. Zorila TC, Kandia V, Stylianou Y; Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. InThirteenth Annual Conference of the International Speech Communication Association, 2012.
9. Aggarwal R, Singh JK, Gupta VK, Rathore S, Tiwari M, Khare A; Noise reduction of speech signal using wavelet transform with modified universal threshold. International Journal of Computer Applications, 2011; 20(5):14-9.
10. King S, TaylorP; Detection of Phonological Features in Continuous Speech using Neural Networks," Computer Speech and Language, 2000; 14(4): 333-345.
11. Eide E; Distinctive Features for Use in an Automatic Speech Recognition System. Proc. Eurospeech 2001; 3:1613-1616.
12. Nitta T; Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA. Proc. ICASSP'99, 1999; 421-424.
13. Kobayashi T; ASJ continuous speech corpus for research. Acoustic Society of Japan (ASJ) Trans., 1992; 48(12):888-93.
14. JNAS; Japanese Newspaper Article Sentences:http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html.
15. Itahashi; A noise database and Japanese common speech data corpus," J. Acoust. Soc. Jpn., 1991; 47(12): 951–953.