International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 3 (2011) pp.088-095
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Tandem MLNs based Phonetic Feature Extraction for Phoneme Recognition

Mohammed Rokibul Alam Kotwal[1], Foyzul Hassan[2], Ghulam Muhammad[3] and Mohammad Nurul Huda[4]

[1]United International University, Department of Computer Science and Engineering,
House 80, Road 8/A, Satmasjid Road, Dhanmondi, Dhaka-1209, Bangladesh
rokib_kotwal@yahoo.com

[2]United International University, Department of Computer Science and Engineering,
House 80, Road 8/A, Satmasjid Road, Dhanmondi, Dhaka-1209, Bangladesh
foyzul.hassan@gmail.com

[3]King Saud University, Department of CE, College of CIS
Riyadh 11451, Kingdom of Saudi Arabia
gmd_babu@yahoo.com

[4]United International University, Department of Computer Science and Engineering,
House 80, Road 8/A, Satmasjid Road, Dhanmondi, Dhaka-1209, Bangladesh
mnh@cse.uiu.ac.bd

***Abstract*: This paper presents a method for automatic phoneme recognition for Japanese language using tandem MLNs. Here, an accurate phoneme recognizer or phonetic type-writer, which extracts out-of-vocabulary (OOV) word for resolving OOV problem that occurred when a new vocabulary does not exist in word lexicon, plays an important role in current hidden Markov model (HMM)-based automatic speech recognition (ASR) system. The construction of the proposed method comprises three stages: (i) the multilayer neural network (MLN) that converts acoustic features, mel frequency cepstral coefficients (MFCCs), into distinctive phonetic features (DPFs) is incorporated at first stage, (ii) the second MLN that combines DPFs and acoustic features as input and outputs a 45 dimensional DPF vector with less context effect is added and (iii) the 45 dimensional feature vector generated by the second MLN are inserted into a hidden Markov model (HMM) based classifier to obtain more accurate phoneme strings from the input speech. From the experiments on Japanese Newspaper Article Sentences (JNAS) in clean acoustic environment, it is observed that the proposed method provides a higher phoneme correct rate and improves phoneme accuracy tremendously over the method based on a single MLN. Moreover, it requires fewer mixture components in HMMs. Consequently, less computation time is required for the HMMs.**

Keywords: multilayer neural network, hidden Markov model, automatic speech recognition, mel frequency cepstral coefficients, distinctive phonetic features, out-of-vocabulary**.**

I. Introduction

A new vocabulary word or out-of-vocabulary (OOV) word often causes an “error” or a “rejection” in current hidden Markov model (HMM)-based automatic speech recognition (ASR) systems. To resolve this OOV-word problem, an accurate phonetic typewriter or phoneme recognizer functionality is expected [1]–[3].

Various methods had been proposed to accomplish this phoneme recognition [4], [5] and some of them showed acceptable performances. However, most of them based on HMMs have several limitations. For example, a) they need a large number of speech parameters and a large scale speech corpus to negotiate coarticulation effects using context-sensitive triphone models, and b) they need higher computational cost to get acceptable performances in HMMs.

To resolve the problems of current HMM-based phoneme recognizers, a lower computational cost algorithm with higher recognition accuracy is needed. An articulatory-based or a distinctive phonetic feature (DPF)-based system can model coarticulatory phenomena more easily [6], [7]. In our previous work, a DPF-based feature extraction method was introduced [8], where a multi-layer neural network (MLN) was used to extract DPFs. The DPF-based system i) widens the margin of acoustic likelihood, ii) avoids the necessity of a large number of speech parameters and iii) incorporates context-dependent acoustic vectors to negotiate dynamics. However, because a single MLN is unable to model longer context, it cannot resolve coarticulation effects precisely.

In this paper, we propose a DPF-based phoneme recognition method using tandem MLNs for an ASR system, which consists of three stages, to solve the problems of coarticulation. The first stage extracts a 15 dimensional DPFs vector from acoustic features of an input speech using an MLN. The second stage MLN, which combines DPFs and acoustic features as input, generates a 45 dimensional DPFs vector with less context effect. The third stage incorporates an HMM based classifier to obtain more accurate phoneme strings from the input speech by taking 45 dimensional DPF vectors generated from the second stage MLN. The originality of this paper is to derive hybrid features (output articulatory features of the first MLN and acoustic features extracted from the input speech signal) for constructing input parameters of the second MLN. It is expected that the proposed system generates more precise phoneme strings at low computational cost in HMMs and consequently, gives a functionality of a high performance phonetic typewriter.

In this study, from the phoneme recognition performance point of view, we investigate and evaluate two types of DPF-based feature extraction methods. These methods are (i) DPF using MLN [8] and (ii) DPF using Tandem MLNs. Another experiment is done for the mel frequency cepstral coefficients (MFCCs), which is directly inserted into the HMM-based classifier for obtaining comparable performance.

The paper is organized as follows: Section II discusses the articulatory features. Section III explains the system configuration of the existing phoneme recognition methods with the proposed. Experimental database and setup are provided in Section IV, while experimental results are analyzed in Section V. Finally, Section VI draws some conclusion with some future remarks

II. Articulatory Features

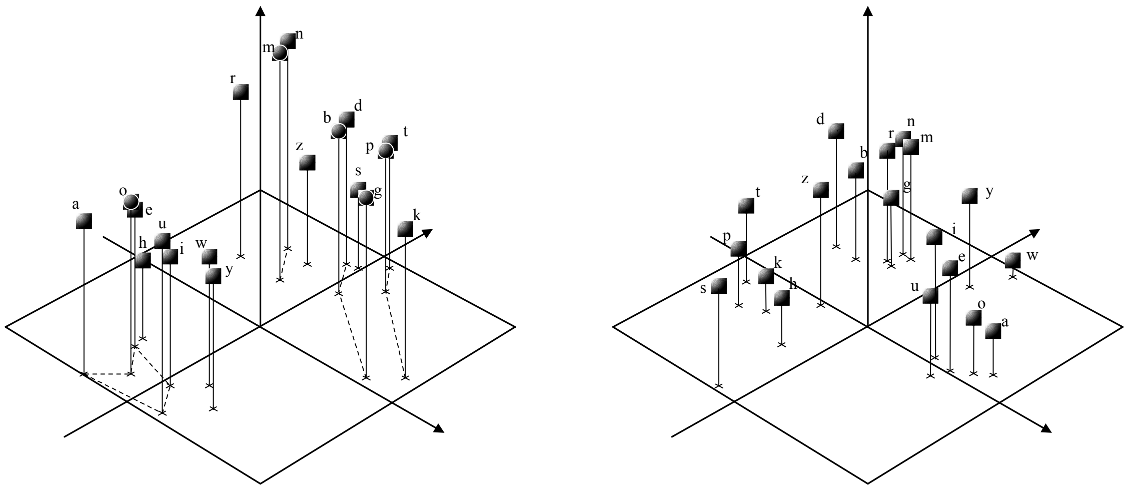
A phone can easily be identified by using its unique articulatory features or distinctive phonetic features (DPFs) set [9]–[11]. Because the traditional-DPF is designed for ASR system with limited domain, the feature vector space composed of the traditional-DPF shows low performance for classifying speech signals. A novel DPF set for classifying Advanced Telecommunications Research Institute International (ATR) with 15 elements, as shown in Table I, which is designed by modifying a Japanese traditional DPF set [12] is used. Windheuser and Bimbot previously proposed a DPF set in which a balance of distances among phonemes is adjusted for classifying English phonemes [13], [14]. The design concept of Japanese balanced-DPF set follows this idea. Each phoneme has five positive elements on average. In Table 1, present and absent elements of the DPF, which are indicated by "+" and "-" signs, are called positive and negative features, respectively. In this DPF set, the balance of distances among phonemes is adjusted by adding new elements, that is, an element "nil" is added as an intermediate expression of "high/low" and "anterior/back" and two elements of "vocalic" and "unvoiced" are also applied. The other change for balancing is the replacement of "fricative" by "affricative". Long vowels (/a:, i:, u:, e:, o:/) have the same positive features as short vowels (/a, i, u, e, o/). On the other hand, silence (/silB, silE/), glottal stop (/q/), and short pause (/sp/) have no positive features in either traditional-DPF or B-DPF. The main difference between the balanced-DPF and the traditional- DPF in Figure 1 is that the consonantal group is separated into two groups of a voiced consonant group and an unvoiced consonant group, that is, the phonemes within the voiced consonant group and the unvoiced consonant group are distributed close to each other. As a result, the balanced-DPF set has three groups consisting of the voiced consonants, the unvoiced consonants, and vowels. Finally, Japanese balanced DPF values are vocalic, high, low, intermediate between high and low <nil>, anterior, back, intermediate between anterior and back <nil>, coronal, plosive, affricate, continuant, voiced, unvoiced, nasal and semi-vowel.

III. Why DPF based method is necessary?

This section describes the necessity of phonetic features in ASR. Figure 2(a) and 2(b) show the phoneme distances of five Japanese vowels in an utterance, /ioi/ that are calculated with a MFCC-based ASR system and a DPF-based system using an MLN, respectively. In both the systems, each distance is measured using the Mahalanobis distance between a given input vector and the corresponding vowel set of mean and covariance in a single-state model. The input sequence in the figures, /i/../i//o/../o//i/../i/, exhibits phoneme for each frame and has total 20 frames in which first three frames, middle 13 frames, and last four frames are phonemes /i/, /o/, and /i/, respectively. The MFCC-based system (Figure 2(a)) shows seven misclassification of phonemes (/u/ output for /o/ and /i/ input) for frames 4, 5, 13, 14, 15, 16, and 17, while two misclassification (/o/ and /u/ output for /i/ input) for frames 17 and 18 are observed by the DPF-based system (Figure 2(b)). Therefore, the DPF-based system outputs few misclassifications. However, because some errors caused by coarticulation still remain, as shown in Figure 2(b), the DPF-based system using a single MLN requires further modifications.

Table 1. Japanese Balanced DPF-Set for classifying ATR Phonemes.

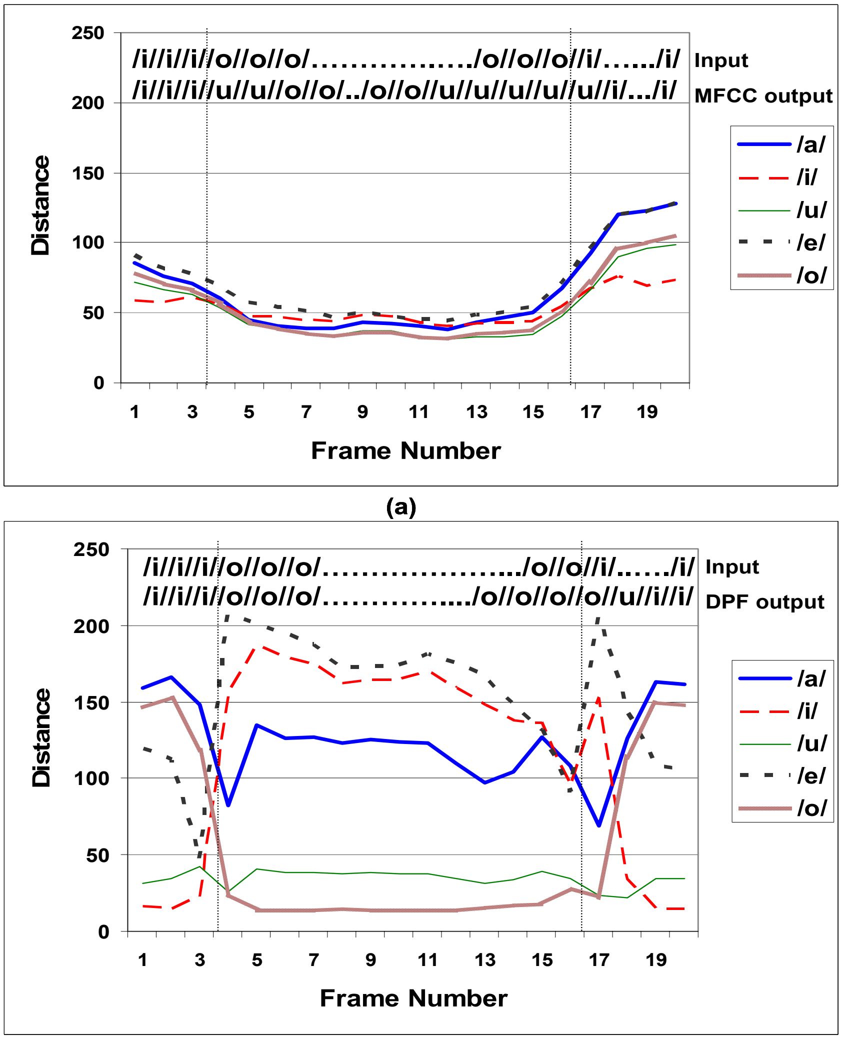
DPFs	a	i	u	e	o	N	w	y	j	my	ky	dy	by	gy	ny	hy	ry	py	p	t	k	ts	ch	b	d	g	z	m	n	s	sh	h	f	r
vocalic	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
high	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	+	-	-	+	-	-	-	-	+	-	+	-
low	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
nil	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	+	+	-	+	+	+	+	-	-	-	+
anterior	-	-	-	-	-	-	-	-	+	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	-	+	+
back	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-
nil	-	+	-	+	-	+	-	+	-	-	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
coronal	-	-	-	-	-	-	-	-	+	-	-	+	-	-	+	-	+	-	-	+	-	+	+	-	+	-	+	-	+	+	+	-	-	+
plosive	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	+	+	+	+	-	-	+	+	+	-	-	-	-	-	-	-	-
affricative	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+	-	-	-	-	-	-	-
continuant	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	+	+	+	-
voiced	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	+
unvoiced	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	+	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	-
nasal	-	-	-	-	-	+	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-
semi-vowel	-	-	-	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+



(A) Traditional DPF

(B) Balanced DPF

Figure 1. Three dimensional DPF space for (a) Traditional-DPF and (b) Balanced-DPF



(a)

(b)

Figure 2. Phoneme distances for utterance, /ioi/ using (a) MFCC-based system and (b) DPF-based system

IV. Phoneme Recognition Systems

A. MFCC-based System using MLN

Figure 3 shows the DPF-based phoneme recognition method using MLN. At the acoustic feature extraction stage, input speech is converted into MFCCs of 38 dimensions (12 MFCC, 12ΔMFCC, 12ΔΔMFCC, ΔP, ΔΔP, where P is energy of the raw input signal). MFCCs are input to an MLN with three layers, including two hidden layers, after combining preceding (t-3), (t-2), (t-1) frames and succeeding (t+1) (t+2) (t+3) frames with the current t-th frame. The MLN has 15 DPFs output for current t-th frame. The two hidden layers consist of 500 and 30 units, respectively. The MLN is trained by using the standard back-propagation algorithm. The DPF-based method using a single MLN yields comparable recognition performance. However, Because of lacking of feedback connection, the single MLN suffers from an inability to model dynamic information precisely.

B. Proposed System

In the proposed method shown in Figure 4, Tandem MLNs with large context window are used instead of a single MLN. Acoustic features, MFCCs from input speech are extracted as the same way described in Section III.A. MFCCs are input to the first stage MLN with three layers, including two hidden layers, after combining preceding (t-3), (t-2), (t-1) frames and succeeding (t+1) (t+2) (t+3) frames with the current t-th frame. The MLN has 15 DPFs output for current t-th frame. The architecture of first MLN is same as MLN mentioned in Section IV.A. Then, these output DPFs and input seven continuous frames MFCC, which is 281 (=38×7+15) dimensions, are inserted into second MLN that produces 45 dimensional DPF vector (15 DPF for "t-3" th frame, 15 DPF for "t" th frame, and 15 DPF for "t+3" th frame). Here, for first and second stages MLNs, <input layer, first hidden layer, second hidden layer, output layer> is assigned by the values <266, 500, 30, 15> and <281, 500, 90, 45>, respectively and each of both MLNs is trained by the standard back-propagation algorithm, where momentum coefficient is used not for getting trapped in local optima.

V. Experiments

A. Speech Database

The The following two clean data sets are used in our experiments.

D1. 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 [15].

D2. Test Data Set. This test data set comprises 2379 JNAS [16] sentences uttered by 16 different male speakers (16 kHz, 16 bit).

B. Experimental Setup

Frame length and frame rate are set to be 25 ms and 10 ms, respectively. MFCCs consist of a vector of 38 dimensions (12 MFCC, 12Δ, 12ΔΔ, ΔP and ΔΔP, where P is log energy of raw signal).

In our experiments of the single MLN and tandem MLNs, the non-linear function, (1/(1+exp(-x))) is a sigmoid from 0 to 1 for the hidden and output layers.

Phoneme correct rates (PCRs) and phoneme accuracy (PAs) for D2 data set are evaluated using an HMM-based classifier. The D1 data set is used to design 38 Japanese monophones HMMs with five states, three loops, and left-to-right models. 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. To evaluate PCRs and PAs using D2 data set, the following two experiments are designed, where input features for the HMM-based classifier are DPFs of 15 and 45 dimensions respectively for the existing and proposed methods.

(a) MFCC (dim:38)

(t) DPF (MFCC-MLN,dim:15)

(11) DPF (MFCC-TandemMLNs, dim:45) [Proposed].

Table 2 shows phonemes and their frequencies in the test data set. From the table it is shown that some phonemes (for example: dy, by and py) are less frequent with respect to some other phonemes (for example: a, i, u, e, o). It can be mentioned from the table that beginning and end silences (silB, silE) and short pause (sp) are more frequent in the test data set.

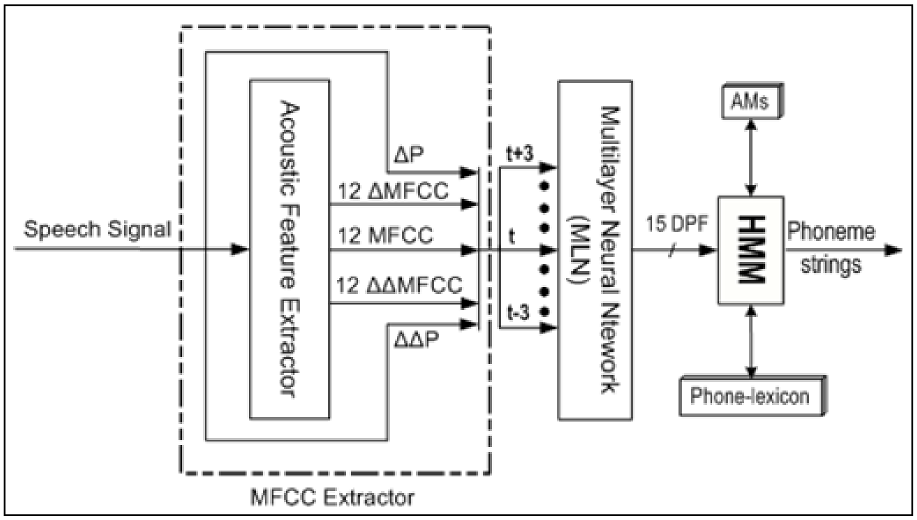


Figure 3. Phoneme recognition method using a single MLN

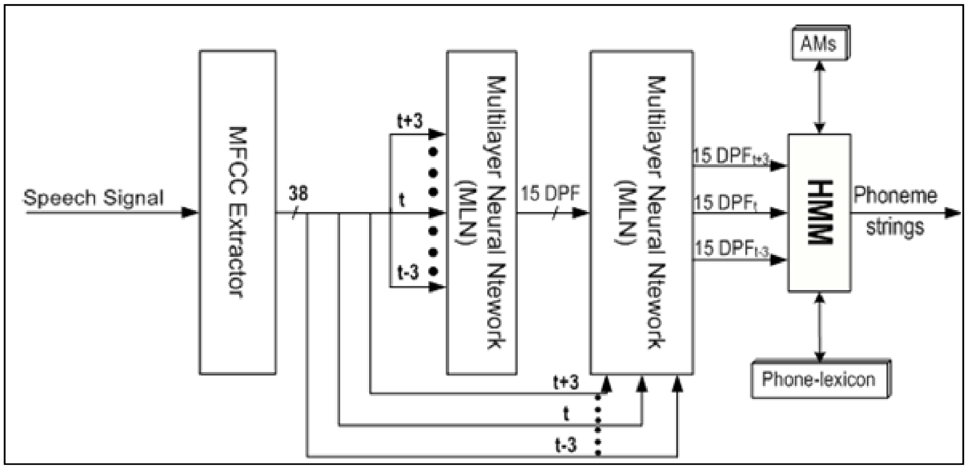
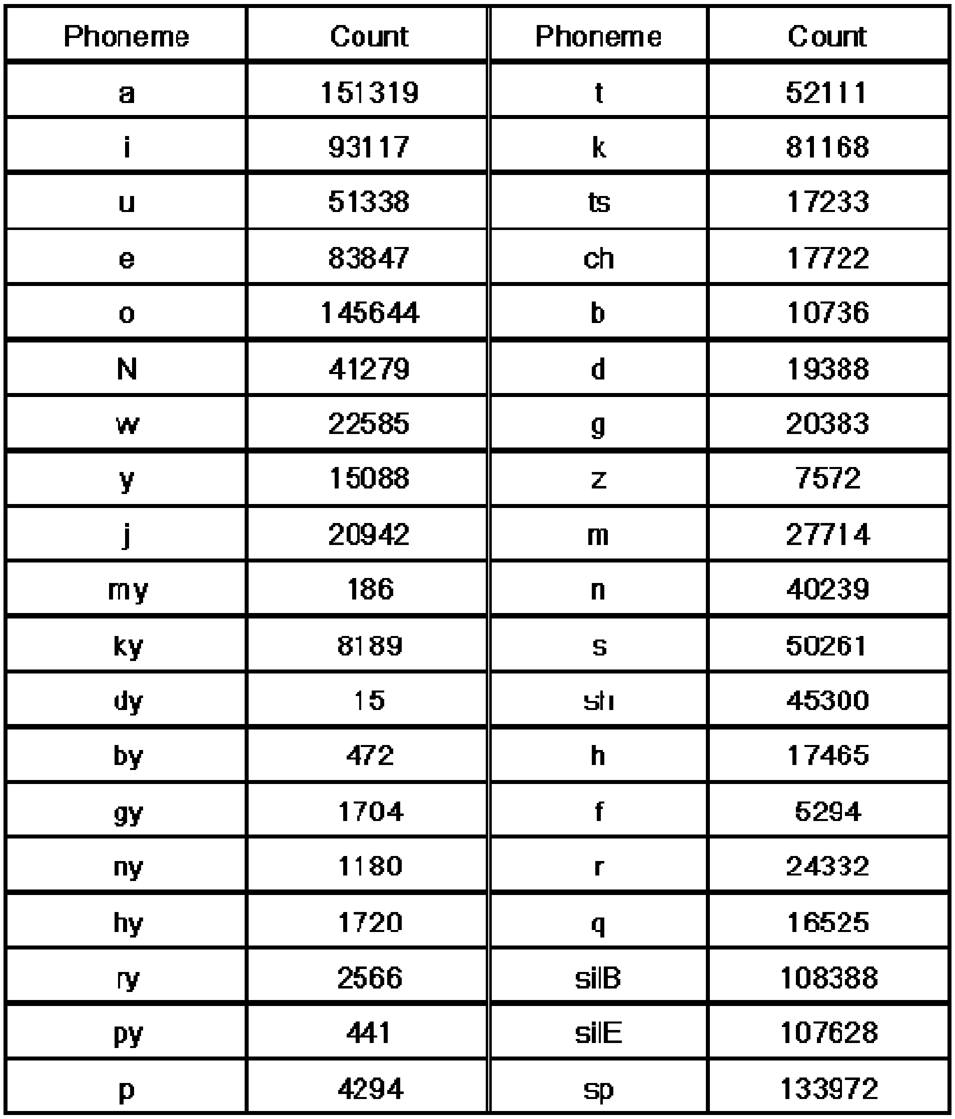


Figure 4. Proposed Phoneme recognition method using Tandem MLNs

Table 2. Phonemes and their frequencies in test data set.



Phoneme	Count	Phoneme	Count
a	151319	t	52111
i	93117	k	81168
u	51338	ts	17233
e	83847	ch	17722
o	145644	b	10736
N	41279	d	19388
w	22585	g	20383
y	15088	z	7572
j	20942	m	27714
my	186	n	40239
ky	8189	s	50261
dy	15	sh	45300
by	472	h	17465
gy	1704	f	5294
ny	1180	r	24332
hy	1720	q	16525
ry	2566	silB	108388
py	441	silE	107628
p	4294	sp	133972

Table 3. Comparison of PCRs for the methods (a), (t) and (11).

Methods	Phoneme Correct Rate (%)				
	1 Mix	2 Mix	4 Mix	8 Mix	16 Mix
MFCC (dim: 38)	62.44	67.12	69.78	71.92	73.24
(t) DPF(MFCC-MLN,dim:15)	76.19	76.57	76.91	77.05	77.35
(11) DPF(MFCC-TandemMLNs,dim:45)	73.03	75.09	77.23	77.61	78.44

VI. Experimental Results and Analysis

Figures 5 and 6 shows the PCRs and PAs comparison between a single MLN and tandem-MLNs based methods, respectively, for MFCC input. It is observed from the Fig. 5 that the tandem-MLNs provide higher PCR than a single MLN for all mixture components except 1 and 2. In the case of PA of Figure 6, tandem-MLNs used in the proposed method shows its superiority for all mixture components except 1. For an example, at mixture component 16, a tandem-MLNs provide 78.44% PCR and 56.80% PA, while a single MLN exhibit 77.35% PCR and 47.89% PA. The method (t) needs higher mixture components in the HMMs to obtain higher PCR and PA. On the other hand, the proposed requires fewer mixture components for obtaining a higher phoneme recognition performance. It may be mentioned from the Fig. 6 that the proposed method using tandem MLNs provides tremendous improvement of PAs over the method (t), while the PCRs improvements in the proposed method (11) are less significant (see Figure 5).

Table 3 exhibits the comparison of phoneme correct rates for the methods (a), (t) and (11) for investigated mixture components. It is observed from the experiments that the MFCC-based method that does not incorporate artificial neural network provides poor recognition performance. For example, the proposed method (11) shows 73.03%, 75.09%, 77.23%, 77.61% and 78.44% PCR for the mixture components one, two, four, eight and 16, while the corresponding values for the MFCC-based method are 62.44%, 67.12%, 69.78%, 71.92% and 73.24% for the respective mixture components.

It is claimed that the proposed method reduces mixture components in the HMMs and hence computation time. The required time for the HMM-based classifier is O(ms2T), where m, S and T represent the mixture components used in the HMM, the number of HMM states and the number of observation sequences. For an example from the Figure 6, approximately 47.50% phoneme recognition accuracy is obtained by the methods (t) and (11) at mixture components 16 and two, respectively. For (t), the required time in the HMMs is 16x52x200 (=80K), while the corresponding time for the proposed method (11) is 2x5^{2}x200 (=10K) assuming number of observation sequence is 200 frames. Therefore, the proposed method requires fewer mixture components as well as less computational cost in the HMMs.

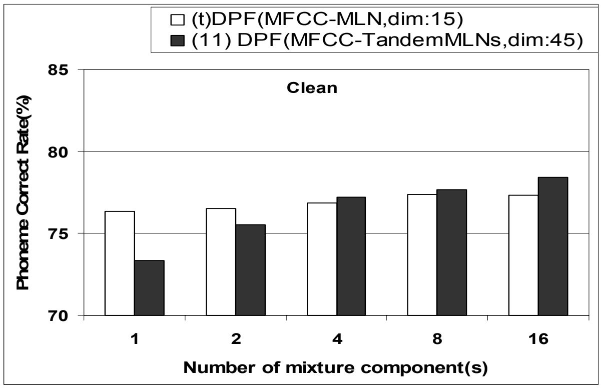


Figure 5. Comparison of PCR between MLN and Tandem-MLNs based methods for input MFCC

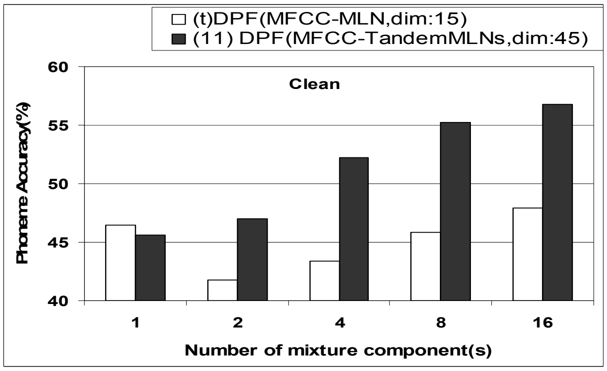


Figure 6. Comparison of PA between MLN and Tandem-MLNs based methods for input MFCC

VII. Conclusion

This paper has presented a DPF-based automatic phoneme recognition method using Tandem MLNs. The following conclusions are drawn from the study.

i) The proposed system outperforms the method using a single MLN.

ii) It is obvious that tremendously higher phoneme recognition accuracy is obtained by the proposed method.

iii) The proposed method requires fewer mixture components in the HMM-based classifier. Consequently, less computation time is required for the proposed method.

iv) The neural network based method with single MLN and tandem MLNs output higher phoneme correct rate over the method based on MFCC.

In near future, the authors would like to do some experiments for evaluating Bangla (can also be termed as Bengali) phonemes spoken by Bangladeshi People. Moreover, we have intension to evaluate word recognition performance using the proposed method.

References

[1] I. Bazzi and J. R. Glass, "Modeling OOV words for ASR," *Proceedings of ICSLP*, Beijing, China, p. 401-404, 2000.

[2] S. Seneff, et. al, "A two-pass for strategy handling OOVs in a large vocabulary recognition task," *Proc. Interspeech*, 2005.

[3] K. Kirchhoff "OOV Detection by Joint Word/Phone Lattice Alignment, "ASRU, Kyoto, Japan, Dec 2007.

[4] D.J Pepper, et. al, "Phonemic recognition using a large hidden Markov model," IEEE Transactions, Volume 40, Issue 6, June 1992.

[5] B. Merialdo (1988), "Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information Training,"*Proc. IEEE ICASSP-88*, pp. 111-114.

[6] K. Kirchhoff, et. al, "Combining acoustic and articulatory feature information for robust speech recognition," Speech Commun.,vol.37, pp.303-319, 2002.

[7] K. Kirchhoffs, “ Robust Speech Recognition Using Articulatory information,” Ph.D thesis, University of Bielefeld, Germany, July 1999.

[8] T. Fukuda, et al, “Orthogonalized DPF extractor for Noise-robust ASR,” IEICE Trans., vol.E87-D, no.5, pp.1110-1118, 2004.

[9] S. King and P. Taylor, “Detection of Phonological Features in Continuous Speech using Neural Networks,” Computer Speech and Language 14(4), pp. 333-345, 2000.

[10] S. King, et. al, “Speech recognition via phonetically features syllables,” Proc ICSLP’98, Sydney, Australia, 1998.

[11] E. Eide, “Distinctive Features for Use in an Automatic Speech Recognition System,” Proc. Eurospeech 2001, vol.III, pp.1613-1616, 2001.

[12] S. Hiki, et al., “Speech Information Processing,” University of Tokyo Press, 1973, *in Japanese.*

[13] C. Windheuser and F. Bimbot, “Phonetic Features for Spelled Letter Recognition with a Time Delay Neural Network,” Proc. Eurospeech’93, pp.1489-1492, Sep. 1993.

[14] S. Okawa, C. Windheuser, F. Bimbot and K. Shirai, “Phonetic Feature Recognition with Time Delay Neural Network and the Evaluation by Mutual Information,” IEICE Technical Report, SP93-131, pp.25-32, Jan. 1994, *in Japanese.*

[15] T. Kobayashi, et al. “ASJ Continuous Speech Corpus for Research,” Acoustic Society of Japan Trans. Vol. 48 No. 12, pp.888-893, 1992.

[16] JNAS: Japanese Newspaper Article Sentences. http://www.milab.is.tsukuba.ac.jp/jnas/instruct.htm

Author Biographies

Mohammed Rokibul Alam Kotwal was born in Dhaka, Bangladesh in 1983. He completed his B. Sc. in Computer Science and Engineering (CSE) Degree from Ahsanullah University of Science and Technology, Dhaka, Bangladesh and M. Sc. in CSE Degree from United International University, Dhaka, Bangladesh. His research interests include Neural Networks, Phonetics, Automatic Speech Recognition, Robotics, Fuzzy Logic Systems, Pattern Classification, Signal Processing, Data Mining and Software Engineering. He is a member of IEEE, IEEE Communication Society and Institution of Engineers, Bangladesh (IEB).

Foyzul Hassan was born in Khulna, Bangladesh in 1985. He completed his B. Sc. in Computer Science and Engineering (CSE) Degree from Military Institute of Science and Technology (MIST), Dhaka, Bangladesh in 2006. He has participated several national and ACM Regional Programming Contest. He is currently doing M. Sc. in CSE in United International University, Dhaka, Bangladesh. His research interests include Speech Recognition, Robotics and Software Engineering.

Ghulam Muhammad was born in Rajshahi, Bangladesh in 1973. He received his B. Sc. in Computer Science and Engineering degrees from Bangladesh University of Engineering & Technology (BUET), Dhaka in 1997. He also completed his M.E and Ph. D from the Department of Electronics and Information Engineering, Toyohashi University of Technology, Aichi, Japan in 2003 and 2006, respectively. Now he is working as an Assistant Professor in King Saud University, Riyadh, Saudi Arabia. His research interest includes Automatic Speech Recognition and human-computer interface. He is a member of IEEE.

Mohammad Nurul Huda was born in Lakshmipur, Bangladesh in 1973. He received his B. Sc. and M. Sc. in Computer Science and Engineering degrees from Bangladesh University of Engineering & Technology (BUET), Dhaka in 1997 and 2004, respectively. He also completed his Ph. D from the Department of Electronics and Information Engineering, Toyohashi University of Technology, Aichi, Japan. Now he is working as an Associate Professor in United International University, Dhaka, Bangladesh. His research fields include Phonetics, Automatic Speech Recognition, Neural Networks, Artificial Intelligence and Algorithms. He is a member of International Speech Communication Association (ISCA).