AI & Soc (2009) 23:603–612
DOI 10.1007/s00146-008-0179-9

OPEN FORUM

Effect of retroflex sounds on the recognition of Hindi voiced and unvoiced stops

Amita Dev

Received: 3 August 2006 / Accepted: 27 January 2007 / Published online: 8 March 2008

Abstract As development of the speech recognition system entirely depends upon the spoken language used for its development, and the very fact that speech technology is highly language dependent and reverse engineering is not possible, there is an utmost need to develop such systems for Indian languages. In this paper we present the implementation of a time delay neural network system (TDNN) in a modular fashion by exploiting the hidden structure of previously phonetic subcategory network for recognition of Hindi consonants. For the present study we have selected all the Hindi phonemes for recognition. A vocabulary of 207 Hindi words was designed for the task-specific environment and used as a database. For the recognition of phoneme, a three-layered network was constructed and the network was trained using the back propagation learning algorithm. Experiments were conducted to categorize the Hindi voiced, unvoiced stops, semi vowels, vowels, nasals and fricatives. A close observation of confusion matrix of Hindi stops revealed maximum confusion of retroflex stops with their non-retroflex counterparts.

1 Introduction

ANN models attempt to achieve real time response and human-like performance using many simple operating elements operating in parallel as in biological nervous system. These models have the greatest potential in areas such as speech and image recognition where many hypotheses are pursued in parallel. Neural networks offer two potential advantages over existing approaches. First they could provide high computation rates required for speech recognition. The advent of new learning

A. Dev (✉)
Department of Training and Technical Education, Ambedkar Institute of Technology, Shakarpur, Near Madhuban, Delhi 110092, India
e-mail: amita_dev@hotmail.com

procedures and the availability of high-speed parallel super computers have created a renewed interest in connectionist models of intelligence. These models are particularly interesting for cognitive tasks that require massive constraint satisfaction, i.e. parallel evaluation of many clues and facts. Other advantages of neural networks are powerful discrimination-based learning procedure and relatively mild assumptions about statistical distribution. Multilayer perceptrons are the most commonly used neural networks. Multilayer perceptron with two hidden layers is capable of forming complex decision surfaces. One of the main limitations of multilayer perceptron neural network architecture is its inability to deal with dynamic nature of time varying signals like speech. An important aspect of this limitation is that a network has to capture the temporal relationship between acoustic events while at the same time provide invariance under translation in time. One way to overcome this limitation is to provide a neural network with memory. This is accomplished by introducing time delays in the synaptic structure of the network (Lippmann 1987). The time delay neural network architecture (TDNN) is one such network. The TDNN is a multilayer feed forward network whose hidden and output neurons are replicated across time and it was devised to capture explicitly the concept of time.

2 TDNN architecture for phoneme recognition

In a TDNN, a basic unit is extended by introducing delays D1 through Dn for each input pattern component as shown in Fig 1. Each undelayed and delayed input is multiplied by a separate weight.

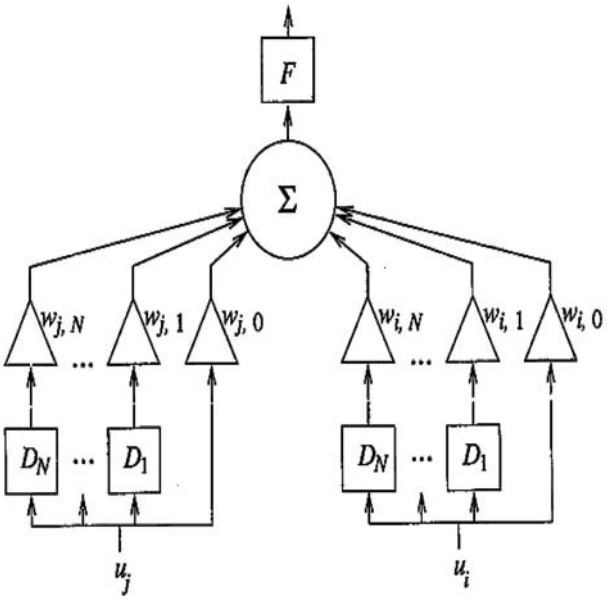

Fig. 1 A time delay neural network unit with only two input channels *j* and *i*

In this way, the TDNN unit gains the ability to represent temporal relations in the stream of events. In addition TDNN is aimed at detection of discrimination of phoneme characteristics and it has a special time shift invariant architecture that is powerful for input interval shifting. For the recognition of phoneme, a three layer network is constructed. Its typical architecture and a set of activities in the units is depicted in Fig. 2. The input to TDNN is a set of parameters from multiple frames of speech signal (Lang et al. 1990). Each frame correspond to signal segment of fixed duration at a given instant of time. Each unit in the hidden layer is connected to the certain number of frames of input (receptive fields). The number of receptive field depends upon the size of the receptive field and the overlap between adjacent receptive fields. First the hidden layer has the role of scanning the input layer in three frame blocks and detecting local phoneme features. In the second hidden layer the local features detected in the first hidden layer are brought together and it plays the role of finding global phonemic features. Here a summing up of features from five frames of the first hidden layer is carried out. In the proposed configuration 15 successive frames each containing 21 spectral coefficients is input to the input layer. Thus the number of units in the input layer is 315 (15 frames × 21 spectral coefficients).

Several learning techniques exist for optimization of neural network. The proposed network model was trained using the standard backpropagation algorithm.

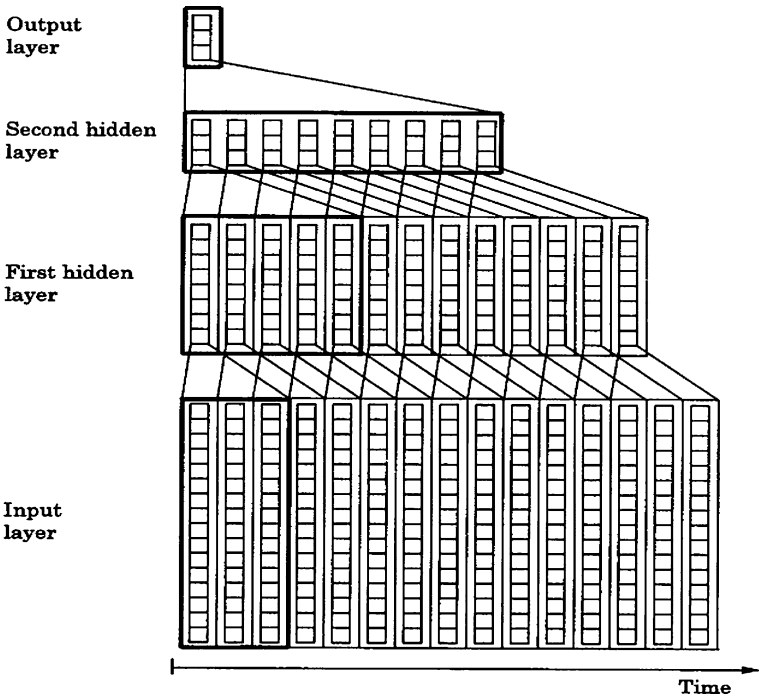

Fig. 2 Model of time delay neural network for phoneme recognition

3 Specific features of Hindi sounds

Each language has a set of abstract linguistic units called phonemes. For example, English can be described by a set of about 42 phonemes whereas Hindi by about 50 phonemes. The sounds of Hindi speech can be conveniently divided into two broad categories of vowels and consonants. Hindi speech contains a set of about 35 consonants, of which about 29 consonants are of frequent usage. These can be conveniently classified according to the manner and place of production as shown in Table 1.

The Hindi consonants possess certain special features which are not so common to European languages and American English. The most significant differences are in stops and affricates, which use both voicing and aspiration (फ थ ठ छ ख भ ध ढ झ घ), to distinguish them from other languages. The trills /र/ and /ल/ have large allophonic variations in different contexts. Consonant clusters like CVCCCV (केन्द), etc. very frequently occur in Hindi speech. During clusters the properties of consonants are greatly affected by the following and preceding sounds. Retroflexion is another feature, which occupies a prominent place in Hindi alphabet. (ट ड ठ ढ). Similarly geminates like (दिल्ली बिल्ली लल्ली) are very common in Hindi speech. Many intervocalic /r/ and retroflex plosives (in non-geminated context) manifest as taps or flaps.

Table 1 Articulatory classification of Hindi consonants

MoP/PoA	Bilabials	Dentals	Retroflex	Palatal	Velar	Glottal
UvUa	प	त	ट	च	क	
	P	t	t*	tʃ	**k**	
VoUa	ब	द	ड	ज	ग	
	b	d	d*	d_z	g	
UvAs	फ	थ	ठ	छ	ख	
	p^h	t^h	t^h	$tʃ^h$	$\mathbf{k^h}$	
VoAs	भ	ध	ढ	झ	घ	
	b^h	d^h	d^h	d_z^h	g^h	
Fricative		स		श		ह
		s		ʃ		h
Vowel like	व	ल	र	य		
	w	l	r	y		
Nasal	म	न				
	m	n				

Abbreviations/Symbols used:

MoP: Manner of Production

PoA: Place of Articulation

UvUa: Unvoiced Unaspirated

VoUa: Voiced Unaspirated

UvAs: Unvoiced Aspirated

VoAs: Voiced Aspirated

4 Hindi speech database, analysis and labeling

A vocabulary of 207 Hindi words was designed for the task specific environment. The spoken samples were recorded in a studio under high-quality recording conditions (*s*/*n* ratio ranging from 50 to 60 dB). The speech data signal is low pass filtered and digitized at 16 kHz sampling rate. A Hindi speech labeling tool, which displays waveform, spectrogram was used for labeling work. By inspecting the spectrogram, segmentation was done phoneme by phoneme as shown in Fig. 3.

These labeled files were Fourier transformed (window length 16 ms, Hamming window, window advance 4 ms) and converted into 21 channel mel frequency spectral coefficients (MFCCs). These 21 MFCCs were normalized and fed into the network.

For the recognition of Hindi phoneme three layer neural net has been used. The input layer consist of 15 × 21 units so that 15 successive feature vectors each containing 21 spectral coefficients serve as input to the network. This layer is connected to the first hidden layer of 13 × 9 units and layer is further connected to second hidden layer of 9 × 4 units.

With this proposed configuration, network can classify four different phonemes, as the output layer holds four units. Each vector of first hidden layer consisting of nine units is fully connected to one feature vector and two delayed versions of the input layer. Thus the weights from a three frame wide window shifting along the input layer connects the input vectors with the first hidden layer. Similar to this connection the first hidden layer and second hidden layer are connected by a five frame wide window. Finally the activation of the output unit is obtained by integrating the activation of corresponding units in the second hidden layer.

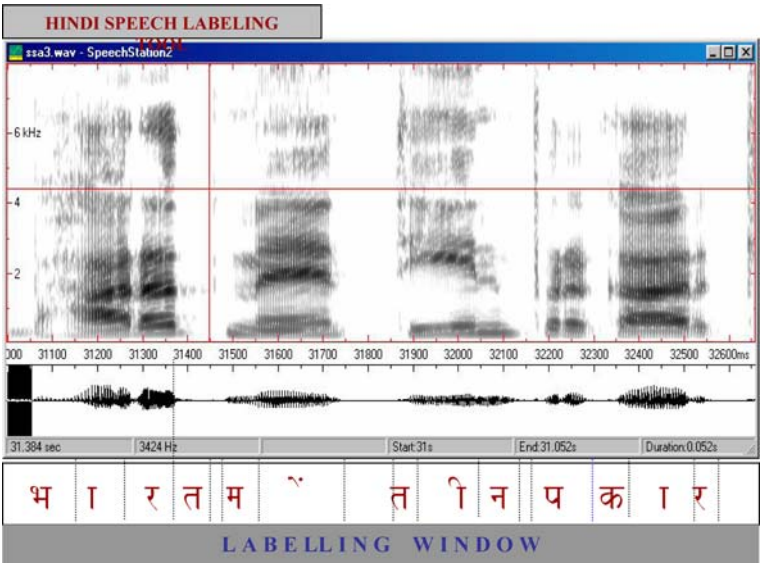

Fig. 3 Hindi speech labelling tool

The training of the TDNN requires that each shift of the window which connects two layers be treated separately first. But after each iteration the averages of the resulting weight changes are used, since they actually belong to the same set of connections. In this way, a correct classification unaffected by temporal shifts is achieved. The usual sigmoid function was chosen for the nonlinear function $F(a) = 1/(1 + e - a)$ because of its convenient properties (Waibel et al. 1989).

5 Learning in a TDNN

Although several learning techniques exist for optimizing neural networks, simple back propagation learning procedure was used. Mathematically back propagation minimizes the mean squared error as a function of the weights by gradient descent. Each iteration requires two passes through the network. During the forward pass, an input pattern is applied to the network initially with small random weights and then the output of the units at each level are computed starting at the input layer (Sarma et al. 1990). The output is compared with the desired output and its error is calculated. During the backward pass, this error is propagated back through the network, and all the weights are adjusted so as to decrease the error. This is repeated many times for all the training tokens until the network converges to produce the desired output (Traunmuller and Lacerda 1987).

6 Recognition results

A total of six nets aimed at the major coarse phonetic classes in Hindi consonants were trained which included

1. unvoiced unaspirated stops (प, त, ट, क)
2. voiced unaspirated stops (ब, द, ड, ग)
3. nasals (म, न)
4. fricatives (स, श, ह)
5. semivowels of both liquids and glides (य, र, ल, व)
6. vowels (आ, ई, ऊ, ओ, ऐ)

Each of the net was given four to five phonemes to distinguish. Evaluation of each net on the test data revealed that TDNN achieved an average recognition score of 99% for all vowel classes, and a recognition rate of 87% for unvoiced stops, 82% for voiced stops, 94.7% for semi vowels, 98.1% for nasals and 96.4% for fricatives as shown in Fig. 4.

These results were compared with the result of German and Japanese language as depicted in Table 2.

It was found that except for voiced and unvoiced stop consonants, results were almost same for rest of the phoneme classes. A close observation of confusion matrix for Hindi voiced and unvoiced stops (Tables 3, 4) revealed that the retroflex stop /ट/ and /ड/ got confused with their nonretroflex counterparts i.e. /त/ and /ट/.

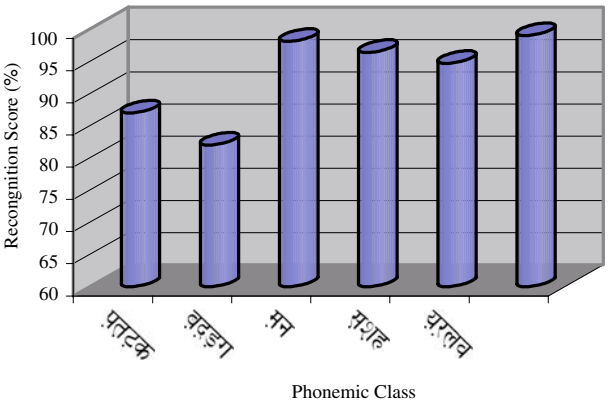

Fig. 4 Recognition scores of phonemic classes of Hindi

Table 2 Comparison of recognition score of different phoneme classes of Hindi, German and Japanese language

Phoneme class	Hindi	German	Japanese
प, त, ट, क	87	96	98.7
ब, द, ड, ग	82	98	98.6
म, न	98.1	97	96.6
स, श, ह	96.4	96	99.3
य, र, ल, व	94.7	98	99.9
आ, ई, ऊ, ओ, ऐ	99	100	99

Table 3 Confusion matrix for voiced Hindi stops training set (350 tokens)

	ब	द	ड	ग	Sum
ब	90	5	5	0	100
द	10	**85**	**28**	0	123
ड	0	**10**	**12**	0	22
ग	0	0	5	100	105
Sum	100	100	50	100	350

Table 4 Confusion matrix for unvoiced Hindi stops training set (350 tokens)

	प	त	ट	क	Sum
प	90	0	5	0	95
त	10	**95**	**20**	0	123
ट	0	**05**	**20**	0	25
क	0	0	5	100	105
Sum	100	100	50	100	350

Table 5 Formant frequencies of the bursts of Hindi stop consonants in CV syllables

POA	F1	F2	F3	F4
Bilabial	500	1,000	2,200	3,500
Dental	450	1,650	2,550	3,700
Retroflex	450	1,700	2,700	3,700
Palatals	400	2,100	2,800	4,000
Velar	550	1,500	2,400	3,600

POA place of ariculation

A close observation of confusion matrix for Hindi revealed that the retroflex stop got confused with their non-retroflex counterpart. A detailed and systematic acoustic phonetic study of Hindi stops was conducted. Study of formant frequencies of bursts of stop consonants in consonant vowel (CV) syllable context revealed that the burst spectrum of dental stops and retroflex stops was almost same in F1, F2 and F4 region and there was slight change in F3 region as shown in Table 5.

Effect of co-articulation

With the view to further investigate the reason for confusion between dental and retroflex sounds, experiment was further extended to study the variations in the spectral characteristics of phonemes particularly dental and retroflex due to change in vowel context. It was noticed that vowel context not only changed the spectral characteristics, but the duration of the phoneme also. The first four formant frequencies of the bursts of Hindi stop consonants in CV context are shown in Table 6.

As evident from Table 6, formant frequencies of burst are different for different place of articulation in each vowel context. It is also observed that second and third formant frequency increases as the place of articulation of consonants moves from the lips to the palate. For velar F2 and sometimes F3 is relatively lower than dentals and retroflex. F1 is found to be higher in case of bilabials and velars as compared to its value in dental, retroflex and palatals. Yet another important observation is that the burst spectrum of dental stops and retroflex stops is almost same in F1, F2, F3 and F4 region.

In order to test the correctness of recognition scores, yet another study was specifically conducted in order to study the durational characteristics of Hindi consonants as per place of articulation. It was revealed that the acoustic features of Hindi speech such as aspiration, burst, VOT and voice bar have considerable differences in duration as well as spectral characteristics as shown in Table 7.

The duration of burst and VOT has been found to be highest in Velars and least in bilabials whereas retroflex and dentals exhibits almost same duration of burst and VOT. Similarly duration of voice bar has been found to be highest in bilabials whereas dentals and retroflex sounds observe almost same duration for voice bar.

Earlier in one of the study conducted to study the significant features in the perception of Hindi consonants in CVC syllables, it was also observed that dentals

Table 6 Formant frequencies of the bursts of Hindi stop consonants in different vowel context

Formant	Vowel	Bilabial	Dental	Retroflex	Palatal	Velar
F1	/ई/	300	200	200	300	300
	/ए/	400	450	450	500	400
	/आ/	500	450	450	400	550
	/ओ/	400	400	450	500	500
	/ऊ/	400	400	300	300	500
F2	/ई/	1700	2000	2000	2350	1800
	/ए/	1800	1800	1850	1850	1900
	/आ/	1000	1600	1800	2100	1500
	/ओ/	900	1200	1500	1600	900
	/ऊ/	1200	1500	1700	2200	1000
F3	/ई/	1800	2800	3000	3000	2600
	/ए/	2400	2400	2400	2600	2300
	/आ/	2200	2500	2700	2800	2400
	/ओ/	2400	2600	2600	2750	2500
	/ऊ/	1800	2600	2800	2900	2600
F4	/ई/	4000	3300	3850	3900	4200
	/ए/	3600	3400	3500	3250	4200
	/आ/	3500	3700	3700	4000	3600
	/ओ/	3450	3500	3600	3250	3450
	/ऊ/	3400	3400	3200	3400	3700

Table 7 Durational characteristics of Hindi stops

S.No	Spectral characteristics	Comparison as per place of articulation
1.	Duration of Voice bar	Bilabial > Velar > Dental > = Retroflex
2.	Duration of Burst	Velar > Retroflex > = Dental > bilabial
3.	Duration of VOT	Velar > Retroflex > = Dental > bilabial

are more frequently confused with retroflex sounds as compared to sounds of any other class as shown in Table 8.

Hence acoustic–phonetic analysis, durational characteristics and perception test also verified the reason for low recognition score of voiced and unvoiced stops due to confusion between dental and retroflex sounds. Experiments were conducted to find out the effect of retroflexion on recognition score. For this purpose the network was trained only with three stop phonemes (प, त, क) (ब, द, ग). There was indeed an improvement in the recognition score as as shown in Table 9.

7 Conclusions

We have presented here a time-delay neural network architecture for the categorization of Hindi phonemes. TDNN was evaluated for six phoneme classes.

Table 8 confusion matrix for the perception of Hindi consonants

	प	त	ट	ख	ब	द	ड	ग	फ	थ	ठ	ख	भ	ध	ढ	घ
प	**523**	**15**			**1**				**1**							
त	**3**	**509**	**26**			**2**										
ट		**41**	**494**				**4**				**1**					
क		**19**	**2**	**518**						**1**						
ब	2				**516**	**4**	**7**	**1**								
द		**4**			**2**	**518**	**13**							**3**		
ड			**3**		**3**	**22**	**509**			**1**				**1**	**1**	
ग				**1**	**1**	**2**	**1**	**533**								**2**
फ									**490**	**24**	**1**	**3**	**13**			
थ									**3**	**525**	**7**			**5**		
ठ							**1**		**10**	**18**	**398**	**5**		**4**	**4**	
ख									**4**	**8**		**524**				**3**
भ					**4**				**13**				**503**	**8**	**3**	**6**
ध						**4**	**1**			**3**			**3**	**524**	**7**	**1**
ढ							**11**			**4**	**9**		**2**	**75**	**435**	**4**
घ								**2**	**1**	**1**		**12**	**2**	**7**	**4**	**511**

Table 9 Improvement in the recognition score of voiced and unvoiced stops

Phoneme class	Recognition score (%)
(प, त, क)	87.5
(ब, द, ग)	90.0

The scores obtained for Hindi consonants and vowels were compared with German and Japanese scores. The recognition scores were nearly same for all the classes except for voiced and unvoiced stop consonants, which were lower in case of Hindi. This was primarily due to the confusion occurring between the retroflex and dental sounds. Acoustic phonetic analysis, durational study and perception study conducted for Hindi stop consonants also indicated considerable similarity between dental and retroflex sounds.

References

Lang KJ, Waibel AH, Hinton GE (1990) A time delay neural network architecture for isolated word recognition. Neural Netw 3:23–43

Lippmann RP (1987) An introduction to computing with neural nets. IEEE ASSP Mag April:8–20

Sarma ASS, Ganesan M, Agrawal SS (1990) Development of speech data base of 200 spoken Hindi words, International workshop on speech technology for man–machine interaction, Bombay, Dec 10–12–1990

Traunmuller H, Lacerda F (1987) Perceptual relativity in identification of two formant-vowels. Speech Commun 6:143–157

Waibel AH, Sawai H, Shikano K(1989) Consonant recognition by modular construction of large phonemic time delay neural networks. ICASSP-89