J. Indian Inst. Sci., May–June 1999, **79**, 197–206.

Reordering network as postprocessor in modular approach-based neural network architecture for recognition of consonant–vowel (CV) utterances

C. CHANDRA SEKHAR AND J. Y. SIVA RAMA KRISHNA RAO
Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India.
email: chandra@svalpha2.iitm.ernet.in, Phone: +91-44-4458341, Fax: +91-44-4458352

Abstract

Recognition of consonant–vowel (CV) utterances in Indian languages is a challenging task because of the large number of classes and the high confusability among several classes. Modular approach based on artificial neural network models is considered for recognition of CV utterances. In this approach, the large number of classes is divided into subgroups and a separate network is trained for each subgroup. Three different grouping criteria are considered and the performance of modular networks based on these criteria is studied. An improved performance is obtained by combining evidence from the three modular networks. Because of similarities among several classes, the class of a test utterance may not always have the strongest evidence. However, it may be among a small set of alternative classes with strong evidence. We propose to train another neural network to further discriminate among these classes and reorder the alternatives. A significant increase in the performance is obtained by using the reordering network as a postprocessor for recognition of isolated utterances of 65 CV classes in Indian languages.

Keywords: Speech recognition, consonant–vowel (CV) units, modular networks, reordering network.

1. Introduction

Development of suitable models for recognition of subword units is important for vocabulary-independent speech-recognition systems. In Indian languages, consonant–vowel (CV) segments occur with high frequency and they are also the basic units of speech production. Some of the coarticulation effects are also captured in CVs. Therefore, CVs are useful as subword units for Indian languages.[1] There are 29 consonants that are commonly used in Indian languages. A consonant can be followed by any of the five vowels to form a CV unit. Therefore, the number of CV classes is large (145). Because of similarities in their production, the sounds of CV classes are confusable. Recognition of 145 CV classes is a greater challenging task than the recognition of English alphabet.[2]

We consider a modular approach for handling large number of CV classes. The consonants in Indian languages belong to the following categories: (1) Stop consonants, (2) Nasals, (3) Semivowels, (4) Fricatives, and (5) Affricates. There are 16 stop consonants. The confusability among the 80 stop–consonant–vowel (SCV) classes is high. Therefore, a separate recognition system is developed for these 80 classes.[3] In the present paper, we address the issues in developing a system for recognition of isolated utterances of the 65 CV classes corresponding to the consonants of the other four categories.[4] It will be necessary to combine these two systems to

develop a system for recognition of all the 145 CV classes in Indian languages. This approach can be extended for recognition of CV segments in continuous speech.[5]

In the next section, we present the modular approach for classification of CVs. In Section 3, we describe the studies on classification of CVs and present the performance of subnets, modular networks and the combined evidence method. We also present the effect of using a reordering network on the performance in recognition of CVs.

2. A modular approach for classification of CVs

When the number of classes is large and the similarity amongst the classes is high, it is difficult to train a monolithic neural network classifier based on the all-class-one-network (ACON) architecture to form the necessary decision surfaces in the input pattern vector space. It is possible to develop a classifier based on the one-class-one-network (OCON) architecture in which a separate network is trained for each class. But the discriminatory capability of the OCON classifiers is poor. Modular approaches[6, 7] can be used to overcome the limitations of the ACON and OCON architectures. In modular approaches, the large number of classes is grouped into small subgroups with a separate trained neural network (subnet) for each subgroup. A postprocessor is used to combine the outputs of the subnets (Fig. 1). For a given set

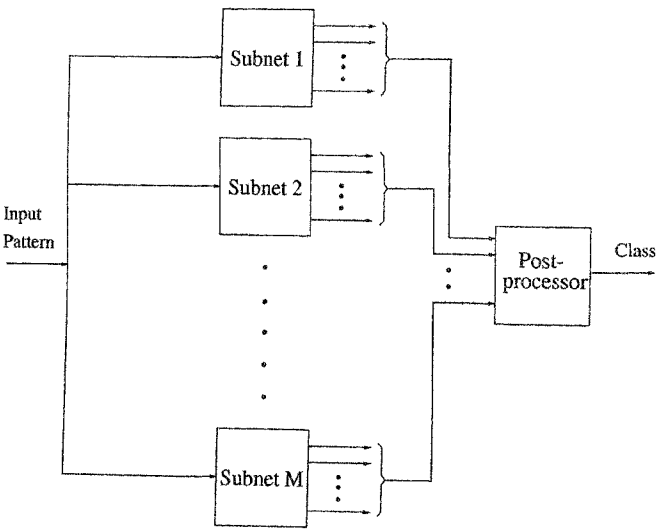

FIG. 1. Block diagram of a modular neural network.

Table I
CV classes in subgroups formed using consonant as the grouping criterion

Subgroup	CV Classes				
/m/	ma	me	mɪ	mp	mu
/n/	na	nr	ni	no	nu
/c/	ca	ce	cɪ	co	cu
/ch/	cha	che	chɪ	cho	chu
/j/	ja	je	jɪ	jo	ju
/jh/	jha	jhe	jhɪ	jho	jhu
/s/	sa	se	sɪ	so	su
/ṣ/	ṣa	ṣe	ṣɪ	so	ṣu
/h/	ha	he	hi	ho	hu
/y/	ya	ye	yi	yo	yu
/r/	ra	re	rɪ	ro	ru
/l/	la	le	lɪ	lo	lu
/v/	va	ve	vɪ	vo	vu

of classes, there can be different modular networks. The modular network based on a particular grouping criterion consists of the subnets for all the subgroups formed using that criterion.

2.1. *Grouping criteria*

The criterion used for grouping the CV classes into subgroups decides the constitution of each subgroup. We consider criteria based on the common phonetic features (based on speech production) among the subsets of CV classes. The first grouping criterion is based on the consonant. In this criterion, all the CV classes in a subgroup have the same consonant. There are 13 subgroups with five classes in each subgroup (Table I). The modular network based on this criterion has 13 subnets.

The second grouping criterion is based on the category of the consonant. All the CV classes in a subgroup have the consonants of the same category. The four subgroups formed using this criterion and the number of CV classes in each subgroup are as follows: Nasals(10), Affricates(20), Fricatives(15) and Semivowels(20).

The third grouping criterion is based on the vowel in CVs. There are five subgroups with one subgroup for each of the five vowels: /a/, /i/, /u/, /e/ and /o/. Each subgroup consists of 13 CV classes, all of which have the same vowel.

For any particular CV class, the subgroup in which it is present will have different classes for different grouping criteria. Therefore, the outputs of the subnets for a pattern of the CV class will also be different depending on the criterion.

2.2. *Multiple modular networks*

The classification performance of a modular network based on a particular grouping criterion depends on the number of subnets in it, their performance and the method used to process the outputs of subnets. Even though the average performance of different modular networks is the same, the subsets of the test data that are correctly classified by them may not be the same. The evidence in the outputs of the subnets based on different grouping criteria for a test pattern can

be combined to decide its class. We study the performance of different modular networks and also the performance of the method in which evidence from the subnets based on two or three grouping criteria is combined.

Because of similarity among several classes, the class of the test pattern may not have the strongest combined evidence. It may be present among a small set of alternatives with strong evidence. We study the performance of different number of alternatives considered.

2.3. *Reordering of alternatives*

The small set of alternatives given by the combined evidence method for a test pattern is obtained by considering the outputs of the subnets for all the 65 CV classes. By focusing on discrimination among the small set of alternative classes, it is possible to reorder the alternatives. A neural network is trained to discriminate the alternative classes and then the test pattern is given as input to this network. The outputs of this network are used to reorder the alternatives and then decide the class of the test pattern.

2.4. *Neural network architecture for classification of CVs*

The CV recognition system, whose architecture is shown in Fig. 2, consists of subnets of the subgroups based on three different grouping criteria. Each subnet is trained with the data of

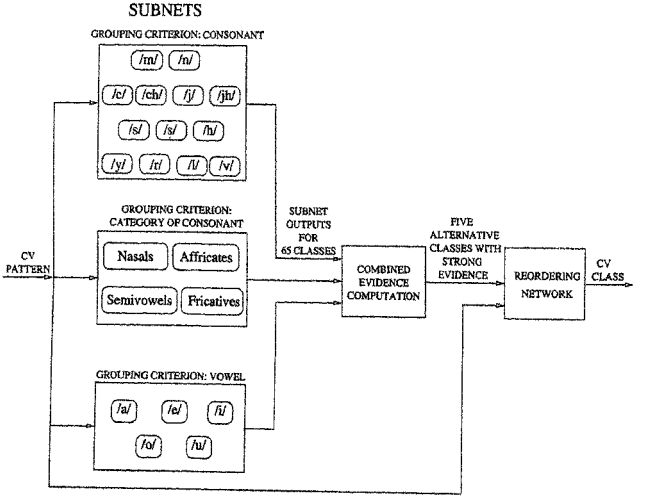

FIG. 2. Block diagram of CV-recognition system.

the classes in its subgroup. During recognition, the test pattern is given as input to all the subnets. The combined evidence for a class is computed by adding the output values of that class from three different subnets corresponding to the subgroups that have the class. Then the reordering network is trained with the data of the alternative classes with strong evidence. The test pattern is then presented to the reordering network and the outputs of the network are used to decide its CV class.

3. Studies on classification of CVs

3.1. *Implementation details*

Isolated utterance data of 65 CV classes was collected from three male speakers. For each class, 12 tokens were collected from each speaker. A fixed-duration portion of the signal around the vowel onset point (VOP) of a CV utterance is processed to derive a pattern vector.[1] A pattern vector is derived from a 200-ms portion of the signal with 60 ms before and 140 ms after the VOP. This signal is processed to extract 40 frames with 12 weighted cepstral coefficients in each frame. In order to reduce the size of the pattern vector, the average coefficients of every two adjacent frames are used. Thus, a 20-frame pattern vector is used to represent a CV utterance. Multilayer perceptron (MLP) model is used to build subnets and reordering network. The MLP model has 240 nodes in the input layer, and 70 and 50 nodes in the first and second hidden layers. The number of nodes in the output layer of a subnet is equal to the number of classes in its subgroup. The training data set for a subnet includes the pattern vectors of utterances of CV classes belonging to its subgroup only. Four tokens from each speaker are used as the training data for a CV class. Training of each subnet is carried out until the error is small. The reordering network has five output nodes and is trained using the training data of the alternative classes only. This training data includes four tokens from each speaker for a class. Training of the reordering network can be stopped after a fixed number of epochs, or when its performance on generalization test data is adequate.[6]

3.2. *Performance of subnets*

The test data set of a subnet consists of the pattern vectors of eight test utterances from each of the three speakers for every CV class in the subgroup. The classification performance is given as a percentage of the total number of pattern vectors in the test data set that are correctly classified by a subnet. The performance of a subnet depends on the number of classes in its subgroup and the confusability among these classes. The performance of the subnets based on different grouping criteria is given in Table II.

The better average performance of the subnets based on the consonant as grouping criterion is mainly due to the small number (5) of classes in a subgroup. It is observed that the performance of the subnets for the subgroups 'Affricates' and 'Semivowels' is poor. It is mainly due to the larger number (20) of classes in these subgroups.

3.3. *Performance of modular networks*

The pattern vector derived from the test utterance of any one of the 65 CV classes is given as input to all the subnets in the modular network. The outputs of the subnets are considered as

Table II
Classification performance of the subnets for subgroups of CV classes formed using different grouping criteria: (a) Consonant, (b) Category of consonant, and (c) Vowel

(a) Consonant		(b) Category of consonant	
Subgroup	Performance	Subgroup	Performance
/m/	83.4	Nasals	71.7
/n/	84.2	Affricates	49.6
/c/	79.2	Fricatives	61.2
/ch/	67.5	Semivowels	56.2
/j/	63.4	Average	57.7
/jh/	59.2		
/s/	68.4	(c) Vowel	
/s/	61.7		
/h/	69.2	/a/	70.2
/y/	77.5	/e/	57.7
/r/	75.0	/i/	60.0
/l/	80.9	/o/	69.9
/v/	81.7	/u/	66.1
Average	73.2	Average	64.7

Table III
Classification performance of modular networks based on different grouping criteria: Consonant (C), Category of consonant (CC), and Vowel (V)

Grouping criteria	*N* used in decision criterion				
	1	2	3	4	5
C	4.5	11.7	19.3	25.4	31.7
CC	22.5	37.9	50.0	57.9	62.6
V	22.5	40.2	52.3	60.6	65.4

the evidence for different classes. A decision criterion that can be used to assign a CV class to the input pattern is to choose the class with the largest output value. Because of the large number of CV classes and similarity among several classes, the output value for the class of the input pattern may not be the largest value always. If the output value for the class of the input pattern is amongst the N largest output values, then that class will be one of the N alternative classes hypothesized by the modular network. The performance of a modular network is given for different cases of the decision criteria used for the classification of CV utterances. The performance for the Case_N is given as a percentage of the total number of pattern vectors in the test data set for which the correct class is among the N alternatives hypothesized. The classification performance of the modular networks based on different grouping criteria is given for different cases (for $N = 1, 2, 3, 4$ and 5) of the decision criterion in Table III.

It is observed that the performance of a modular network based on a particular grouping criterion is poor compared to the average performance of the subnets in it. The difference in the performance is large when the subgroups have small number of classes as in grouping criterion based on the consonant of the CV classes. The performance of modular networks can be improved by using the subnets trained to give low output values for the patterns of the classes that do not belong to their subgroups.[8]

3.4. *Performance of multiple modular networks*

An analysis of the performance of the modular networks has shown that though the subnets do not give the largest output values for the class of an input pattern vector, the output is significantly large. It is also observed that though the average performance of the modular networks based on the second and third groupings is approximately the same, the subsets of the test data that are correctly classified by them are not the same. The evidence in the outputs of the subnets based on different grouping criteria for a particular test pattern vector can be combined to decide its class. The classification performance of the CV-recognition systems that

Table IV
Classification performance of CV-recognition systems that combine the evidence from the subnets in multiple modular networks, Consonant (C), Category of consonant (CC) and Vowel (V)

Multiple modular networks	N used in decision criterion				
	1	2	3	4	5
C, CC	24.3	42.1	52.4	57.5	59.7
C, V	40.5	53.3	57.7	59.5	61.2
V, CC	43.1	54.5	60.2	65.0	70.0
C, CC, V	44.3	55.4	62.4	68.4	72.1

use the method of combining the evidence from the subnets of two or three different modular networks is given in Table IV. The combined evidence for a class is obtained by adding the output values for the class from the two or three subnets corresponding to the subgroups having that class. The performance of a CV-recognition system is given as a percentage of the total number of pattern vectors in the test data set for which the correct class is amongst the classes with the N largest values of the combined evidence.

It is noted that the performance of the CV-recognition system that combines the evidence from the subnets of all the three grouping criteria gives the best performance (for $N = 1$) of 44.3%. The correct class is present among the five alternatives for 72.1% of the test patterns (i.e. for 1124 out of 1560 test patterns of 65 classes). It is also observed that the CV-recognition system that combines the evidence from the subnets of two grouping criteria, the category of consonant and the vowel, gives a similar performance.

3.5. *Performance of reordering network*

Here we study the effects of using a reordering network as postprocessor. For recognition of a CV pattern, an MLP is trained with the data of the alternative classes given by the combined evidence method. Here we consider the method in which the outputs of the subnets in all the three modular networks are combined. The set of alternatives can be different for different test patterns of a CV class. Therefore, training of an MLP has to be done for recognition of every pattern. The training is stopped after 160 epochs. The performance of the CV-recognition system that uses the reordering network as postprocessor is given in Table V.

An analysis of the performance has been carried out to study the effects of using the reordering network. The correct class of a test pattern is present in one of the five positions in the list of alternatives given by the combined evidence method. The purpose of this network is to reorder the alternatives such that the class of the test pattern moves to the first position in the reordered list. The number of test patterns for which the correct class is present in a particular position in the list of alternatives is given in Table VI.

We illustrate the effect of the reordering network with a few examples. For a pattern of CV class /vi/, the list of alternatives (in the order of evidence) given by the combined evidence method is as follows: {/hu/, /mi/, /ni/, /yo/, /vi/}. The outputs of the MLP trained for these five classes give the following ordered list: {/vi/, /mi/, /ni/, /hu/, /yo/}. In this case, the class

Table V
Classification performance of the CV-recognition system that uses a reordering network as postprocessor

Recognition system	N used in decision criterion				
	1	2	3	4	5
Without reordering network	44.3	55.4	62.4	68.4	72.1
	53.8	65.9	69.6	71.2	72.1

Table VI
Number of test patterns for which the correct class is present in a particular position in the list of alternatives

Recognition system	Position in the list				
	1	2	3	4	5
Without reordering network	691	174	109	93	57
	839	189	57	26	13

of the test pattern has moved from the fifth to the first position. For 248 test patterns, the reordering network has moved the class of pattern to the first position.

It is also observed that for 100 test patterns, the reordering network has moved the class of the input pattern from the first position. However, the movement is mostly to the second or third position. This behavior may be because the reordering network has been trained for a fixed number of epochs, but not until the error becomes significantly small. As an example of downward movement, we consider a pattern of the CV class /ha/. The list given by the combined evidence method for this pattern is as follows: {/ha/, /ya/, /ma/, /ja/, /na/}. The reordered list is {/ma/, /ha/, /ya/, /na/, /ja/}.

The number of patterns for which there is an upward movement to the first position from different positions, and the number of patterns for which there is a downward movement from the first position to different positions are given in Table VII.

In another study, training of the reordering network is continued until its performance on the generalization test data is adequate. The generalization test data set includes four tokens per class from each speaker. These tokens are different from the tokens in the training data set. Training of the reordering network is stopped when its performance on the generalization test data set reaches 80%. For many test patterns, this performance is reached in about 70 to 80 epochs. Otherwise, training is continued for 160 epochs. The effect of reordering network trained using this stopping criterion on the movement of the correct class of the test patterns is given in Table VII(b).

Table VII
Number of test patterns for which the class of the pattern has been moved to or from the first position in the list of alternatives by the reordering network for different stopping criteria

(a) Fixed number of epochs

Direction of movement	Position in the list			
	2	3	4	5
Upward	121	58	43	26
Downward	73	17	9	1

(b) Generalization performance

Direction of movement	Position in the list			
	2	3	4	5
Upward	145	66	61	39
Downward	60	13	4	1

Table VII
Classification performance of CV-recognition systems using the reordering networks trained using different stopping criteria

Stopping criterion	N used in decision criterion				
	1	2	3	4	5
Fixed number of epochs	53.8	65.9	69.6	71.2	72.1
Generalization perform-	59.2	67.2	70.2	71.7	72.1

It is noted that for 311 test patterns, the reordering network has moved the class of the patterns to the first position. For 78 patterns, the class of the pattern has been moved from the first position. The classification performance of the CV-recognition system using the reordering networks trained with different stopping criteria is given in Table VIII. It is observed that about 59% of the test patterns have been correctly classified. Recognition accuracy has increased by about 15% compared to the performance of the system without reordering network.

The time taken for training ACON architecture-based reordering network in five alternative classes is about one minute (for 160 epochs on a 266-MHz Pentium processor-based workstation). This time is significantly large for online CV recognition. OCON architecture-based reordering network can be used to overcome this limitation. In this architecture, a separate MLP is trained for each of the 65 CV classes a priori. Then the MLPs of the five alternative classes are used for discrimination and reordering of the classes. This architecture does not require training during the recognition phase. The main limitation of the OCON architecture is its poor discrimination capability compared to the ACON architecture.[1]

4. Summary and conclusions

In this paper, we have proposed different grouping criteria based on the common features among the subsets of classes to develop a system for recognition of isolated utterances of a large number of CV classes. A method that combines the evidence from the subnets based on different groupings is shown to give an improved performance over modular networks. For a small set of alternative classes with strong evidence, a postprocessor network is used to reorder the alternatives. Different criteria have been considered to stop online training of reordering network. It has been demonstrated that the performance in recognition of CVs has increased by about 15% due to reordering.

A recognition performance of about 59% is significant considering that the classification of a CV utterance is carried out by discriminating large number (65) of classes. A constraint satisfaction model based on the acoustic–phonetic knowledge of the classes has been developed to combine evidence from the subnets based on different grouping criteria and improve the performance in recognition of 80 SCV classes.[9] A similar model may be explored to obtain an improved performance for recognition of 65 CV classes belonging to other categories of consonants.

References

1. CHANDRA SEKHAR, C. *Neural network models for recognition of stop–consonant–vowel (SCV) segments in continuous speech*, Ph. D thesis, Indian Institute of Technology, Madras, 1996.

2 LOIZOU, P. C. AND SPANIAS, A. S. High-performance alphabet recognition, *IEEE Trans.*, 1996, **SAP-4**, 430–445

3 CHANDRA SEKHAR, C. AND YEGNANARAYANA, B. Modular networks and constraint satisfaction model for recognition of stop–consonant–vowel (SCV) utterances, *Proc. Int. Jt Conf. on Neural Networks*, Alaska, pp. 1206–1211, May 1998.

4. AVANEENDRA, A. *Development of neural network models for recognition of isolated utterances of consonant–vowel (CV) units*, M. Tech. Project Report, Indian Institute of Technology, Madras, 1998.

5. CHANDRA SEKHAR, C. AND YEGNANARAYANA, B. Neural network models for spotting stop–consonant–vowel (SCV) segments in continuous speech, *Proc. Int. Conf. on Neural Networks*, Washington, DC, pp. 2003–2008, June 1996.

6. HAYKIN, S. *Neural networks: A comprehensive foundation*. Prentice-Hall International, 1999.

7. WAIBEL, A., SAWAI, H. AND SHIKANO, K. Modularity and scaling in large phonemic neural networks, *IEEE Trans.*, 1989, **ASSP-37**, 1888–1898.

8. CHANDRA SEKHAR, C. AND SIVA RAMA KRISHNA RAO, J. Y. Multiple modular networks for recognition of utterances of consonant–vowel units, *Natn. Conf. on Communication*, Kharagpur, pp. 157–162, January 1999.

9. CHANDRA SEKHAR, C. YEGNANARAYANA, B. AND SUNDAR, R. A constraint satisfaction model for recognition of stop–consonant–vowel (SCV) utterances in Indian languages, *Proc. Conf. on Communication Technologies* (CT-96), Bangalore, pp. 134–139, December 1996.