Neural Network Models for Recognition of Consonant-Vowel (C^nV) Utterances

Suryakanth V. Gangashetty and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
Email: {svg,yegna}@speech.iitm.ernet.in

Abstract

In this paper, we present an approach based on neural network models for recognition of utterances of syllable-like units in Indian languages. The distribution capturing ability of an autoassociative neural network model is exploited to perform nonlinear principal component analysis for compressing the size of the feature vector. A constraint satisfaction model is proposed to incorporate the acoustic-phonetic knowledge and to combine the outputs of subnets to arrive at the overall decision on the class of an input utterance.

1 Introduction

Speech recognition involves transforming input speech into a sequence of units called symbols and converting the symbol sequence into a text corresponding to the message in the speech signal. Speech recognition systems vary from simple isolated word recognition systems to highly complex continuous speech recognition systems.

The basic speech units of a language are called the phonemes. At the level of phoneme, it is difficult to capture and represent the pronunciation variation [1]. Therefore syllable-like units are proposed as the sub-word units for speech recognition. These units in general are of C^nV type, where C represents a consonant and V a vowel. In this study we consider the cases of $n = 0$ and $n = 1$, which correspond to 145 Vs and CVs for the Indian language Hindi. Speech information for each CV units is represented by a sequence of spectral feature vectors, over a duration of about 50 msec, with one vector per analysis frame. The total size of the pattern vector per CV unit is typically in the range of 50-250 components. The representation of CV units is effective, if the duration of the CV utterance is chosen anchored around the vowel onset point (VOP). In fact a fixed number of frames before VOP and a fixed number of frames after VOP can be used to derive the pattern vector for each CV unit.

For effective classification, it is necessary to keep the dimensionality of the pattern representation as small as possible. A compressed feature vector can be derived either by using linear compression technique such as principal component analysis (PCA) or nonlinear compression technique using autoassociative neural network models.

The compressed feature vector is used for developing classification models for the CV units. Since the number of CV units is large, the units are grouped into subgroups using the acoustic-phonetic (AP) knowledge of the production of these units. A modular neural network model can be built with separate classification model for each subgroup.

It is possible to exploit the AP knowledge of the sound units to combine the results of preliminary classification of the subgroups of units to arrive at overall decision on the class of an input utterance. A constraint satisfaction neural network model is developed to use the AP knowledge to combine the results of modular networks [2].

This paper is organized as follows: Feature extraction anchored around the VOP is explained in Section 2. Extraction of reduced dimensional features using nonlinear compression technique is described in Section 3. Preliminary classification of CV units using multilayer feedforward neural network (MLFFNN) models is presented in Section 4. Section 5 describes a constraint satisfaction model to incorporate the AP knowledge and to combine the outputs of subnets to arrive at overall decision on the class of an input utterance. Performance of the recognition of CV units uttered in isola-

0-7803-7044-9/01/$10.00 ©2001 IEEE

tion are given in Section 6.

2 Extraction of Acoustic Features Around the Anchor Point

In order to obtain fixed duration patterns automatically from varying duration subword unit segments, it is necessary to identify the anchor point. One of the proposed anchor point is vowel onset point (VOP). For a CV utterance, the VOP is the instant at which the vowel part begins. Once the anchor point is identified, a portion of the speech signal with a fixed duration around the anchor point can be processed to obtain a fixed length pattern vector. The segments around the VOP contain most of the information relevant for recognition of sound units.

The sound units considered in this study are listed in Table 1. Speech data for isolated utterance of each of these units are collected for training and testing. The speech database consists of 12 utterances per unit per speaker for three speakers. Ten of these utterances per speaker are used for training and the remaining two for testing. A segment of 55 msec anchored around the manually marked VOP [3] [4] is considered from each utterance. The segment is divided into five frames as follows: One frame of 20 msec to the left of the VOP and four frames, each of size 20 msec and frame shift of 5 msec, to the right of the VOP. Each frame is represented by 10 weighted linear prediction cepstral coefficients (acoustic features), which are obtained from an 8^{th} order linear prediction analysis [5]. Thus each utterance is represented by a 50 dimensional feature vector.

Table 1: List of consonants and vowels used in the study. Each of the 29 consonants can be followed by any of the five vowels.

Stop Consonant Vowels	Velars	/ka/ /kha/ /ga/ /gha/
	Alveolars	/Ta/ /Tha/ /Da/ /Dha/
	Dentals	/ta/ /tha/ /da/ /dha/
	Bilabials	/pa/ /pha/ /ba/ /bha/
Non-Stop Consonant Vowels	Affricates	/cha/ /chha/ /ja/ /jha/
	Fricatives	/sha/ /sa/ /ha/
	Semivowels	/ya/ /ra/ /la/ /va/
	Nasals	/na/ /ma/
Vowels		/a/ /i/ /u/ /e/ /o/

3 Autoassociative Neural Network Models for Nonlinear Compression of Feature Vectors

The CV units are important information-bearing sound units from production and perception point of view [1]. The number of basic CV units are large (145), and majority of them are confusable. Each unit is represented by a sequence of feature vectors with a dimension of 50. With this large dimension feature vector, a large number of training samples are required to design a neural network (NN) classifier.

Since it is difficult to collect sufficiently large number of training data samples, it is worthwhile exploring methods to compress the size of the feature vector. An AANN model is used for nonlinear compression of the feature vectors. For a given compression ratio, the nonlinear method seems to give better performance over linear methods. The nonlinear compression is accomplished by a five layer AANN model [6] [7] [8], and linear compression can be realized by PCA.

There are two other main reasons to keep the dimensionality of the pattern representation as small as possible: (1) Measurement cost and (2) Classification accuracy.

The five layer AANN model which performs nonlinear principal component analysis is shown in Fig.1 [7]. The second and fourth layers of the network have more units than the input layer. The third layer has fewer units than the first or fifth. The activation functions at the third layer may be linear or nonlinear, but the activation functions at the second and fourth layers are essentially nonlinear. The structure of the AANN model considered in this study is *50L 75N 19N 75N 50L.*

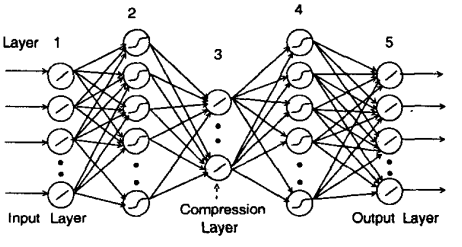

Figure 1: Five layer AANN model used for nonlinear compression of feature vectors.

The trained AANN model is used to compress the 50 dimensional feature vector to 19 dimensional feature vector. The nonlinear compression is performed on each subgroup of the sound units separately.

4 Multilayer Feed Forward Neural Network Models for Preliminary Classification

In this study, the set of 145 CV units are divided into 9 subgroups [3], based on the production characteristics of the sound units [4]. Each subgroup consists of 5 to 20 units, depending on the nature of the subgroup. Table 2 gives the grouping of the 145 CV units into different subgroups based on three criterion, namely place of articulation (POA), manner of articulation (MOA) and vowel (V).

Table 2: List of CV classes and the subgroups based on different grouping criteria for each class.

Stop Consonant Vowels						
MOA	POA	Vowel subgroup				
Subgroup	Subgroup	/a/	/i/	/u/	/e/	/o/
UVUA	Velar	ka	ki	ku	ke	ko
	Alveolar	Ta	Ti	Tu	Te	To
	Dental	ta	ti	tu	te	to
	Bilabial	pa	pi	pu	pe	po
UVA	Velar	kha	khi	khu	khe	kho
	Alveolar	Tha	Thi	Thu	The	Tho
	Dental	tha	thi	thu	the	tho
	Bilabial	pha	phi	phu	phe	pho
VUA	Velar	ga	gi	gu	ge	go
	Alveolar	Da	Di	Du	De	Do
	Dental	da	di	du	de	do
	Bilabial	ba	bi	bu	be	bo
VA	Velar	gha	ghi	ghu	ghe	gho
	Alveolar	Dha	Dhi	Dhu	Dhe	Dho
	Dental	dha	dhi	dhu	dhe	dho
	Bilabial	bha	bhi	bhu	bhe	bho
Non-stop Consonant Vowels						
	Affricates	cha	chi	chu	che	cho
		cha	chi	chu	che	cho
		cha	chi	chu	che	cho
		cha	chi	chu	che	cho
	Fricatives	sha	shi	shu	she	sho
		sa	si	su	se	so
		ha	hi	hu	he	ho
	Semivoels	ya	yi	yu	ye	yo
		ra	ri	ru	re	ro
		la	li	lu	le	lo
		va	vi	vu	ve	vo
	Nasals	na	ni	nu	ne	no
		ma	mi	mu	me	mo
Vowels						
		a	i	u	e	o

For manner of articulation, only the four manners of articulation (Unvoiced Unaspirated (UVUA), Unvoiced Aspirated (UVA), Voiced Unaspirated (VUA), and Voiced Aspirated (VA)) of the stop consonant vowels are considered. Likewise for vowel (V) category also, only the 16 stop consonants in each vowel category are considered. Separate neural network (NN) classifiers are developed for each subgroupi, and the performance of each classifier is examined separately. The NN classifier is a MLFFNN with different structure for each subgroup. The structure of each network varies only in the number of units in the input and output layers. The rest of the structure of the NN is identical for all the classifiers of different subgroups. If the size of input vector of a sound unit for a particular group is m (p, in case of compressed feature vector), and the number of classes in that group are n, then the structure of the classifier is given by *(m/p)L 100N 60N nN*, where L refers to linear unit and N refers to nonlinear unit.

The performance of the classifier network for each of the subgroups based on the place of articulation (POA), Manner of Articulation (MOA) and Vowel, is evaluated using uncompressed (Order=50) and compressed (Order=19) feature vectors. Tables 3, 4 and 5 show the results for preliminary classification of the subgroups of the units.

Only for the POA grouping, the performance for the 8 subgroups are shown. It is interesting to note that in most cases the compressed feature vector performs equally well as the uncompressed feature vector.

A CV unit occurs in different combinations of other units in a subgroup for each grouping criteria. The classifier developed for each subgroup captures the discriminating features of the units in the subgroup. Thus the output for a CV unit from the classifier can be viewed as evidence from three different classifiers. It is possible to exploit the AP knowledge of the CV units to combine the results of preliminary classifiers of the subgroups of the units to arrive at overall decision on the class of the input utterance.

5 Constraint Satisfaction Model for Incorporating Acoustic-Phonetic Knowledge

The proposed constraint satisfaction (CS) model takes the outputs of the three classifiers as inputs and combines it with the production knowledge of the CV units incorporated in the model as constraints [3] [9] [10] [11]. A CS model is a feedback neural network in which each node represents a hypothesis, and the weights connecting the nodes represent the constraints. A global

"goodness of fit" function is defined in terms of the activation state of the nodes and the weights of the network. The advantage of such a network representation is that, when the network relaxes to a stable equilibrium state, the resulting state represents a situation when the constraints are satisfied to the maximum extent. Such a result will be obtained even though the constraints are weak due to partial knowledge of the domain specification, and also due to poor representation of the information.

The proposed system thus consists of two stages. The first stage has three modular networks corresponding to the three different grouping criteria. The modular network for each grouping criterion in turn consists of MLFFNNs for different subgroups. The second stage of the proposed system consists of feedback neural network modules. The outputs of the MLFFNNs in the first stage are used as evidence to the corresponding feedback subnetwork in which the feedback connections are provided using the knowledge of speech production for the CV units. It is interesting to note that the values of the weights on the connecting links in the feedback subnetworks are not unique. That is why the knowledge represented by these weights is termed as 'weak'. It is the combined effect of the entire feedback network network that reinforces even the weak evidence both from the external input (outputs of MLFFNNs) as well as the knowledge of the production incorporated in the model.

The evidence from all the three feedback subnetworks is further reinforced by combining them through another feedback subnetwork which uses the concept of instance pool as in the interaction activation and computation (IAC) model [9]. All the four feedback subnetwork modules have feedback connections among the nodes within the module, and they also have bidirectional connections to the nodes in the instance pool feedback subnetwork. The block diagram of the proposed system for CV recognition is shown in Fig.2.

6 Results And Discussion

The recognition performance of CV units using CS model is significantly higher than the recognition obtained when the outputs of the individual subnet are directly combined (modular network approach). For example, for the result of 80 SCV classess we have obtained a performance improvement from 35% to 70%.

The superior performance of the CS model is that, the output from the MLFFNNs of each grouping criteri-

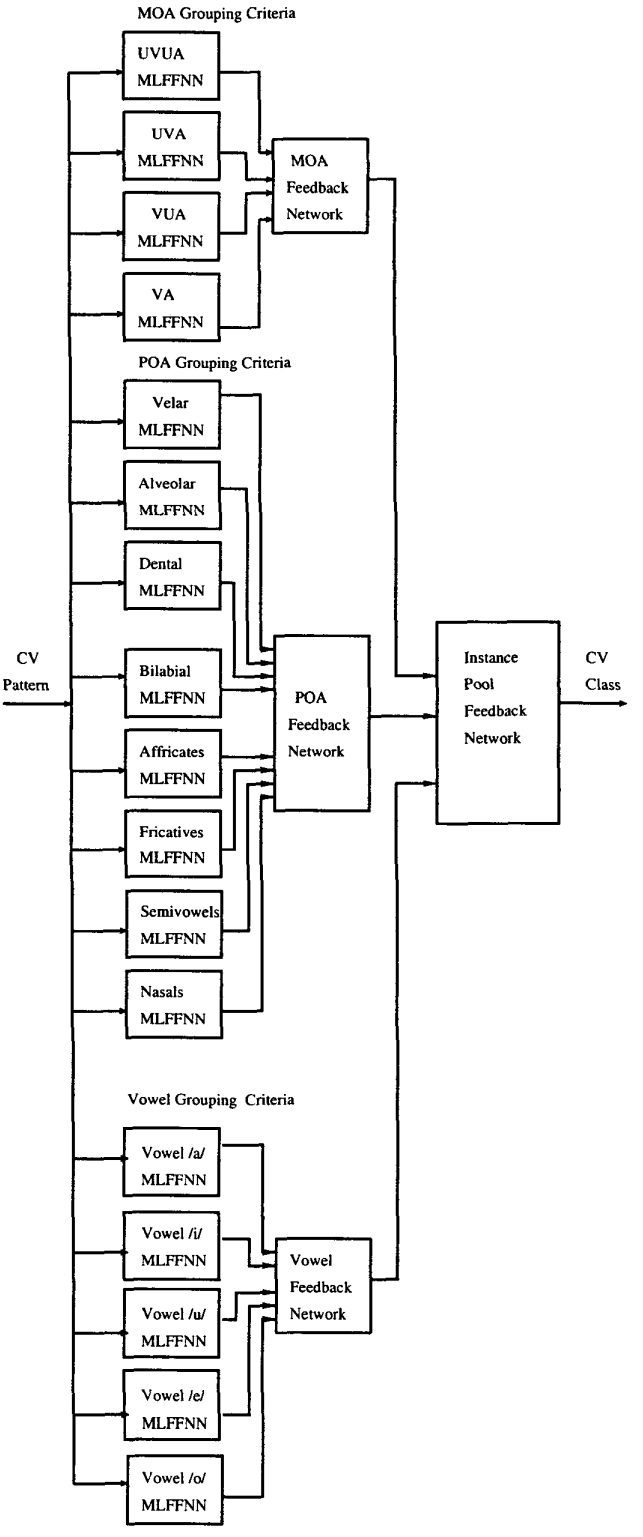

Figure 2: Block diagram of proposed cs model for recognition of CV units.

on are processed by the feedback subnetwork for that grouping. Similarities among the classes are represented by the weights of the connections in the feedback subnetwork. Evidence available from different groupings is combined by letting the feedback subnetworks interact with one another through the instance pool. Therefore the CS model not only uses the knowledge about the similarities among the classes but also combines the evidence from multiple classifiers in performing the classification. On the other hand, the postprocessor in a modular network processes the outputs of the MLFFNNs in that network to decide the class. The postprocessor simply assigns the class of the largest output value without using the similarity information available in other outputs. The modular networks for different groupings operate independent of each other. Hence the performance of the CS model is superior to directly combining the outputs of modular networks.

7 Summary And Conclusions

In this paper, we have proposed NN models for recognition of isolated utterances of syllable-like units. An AANN model is used for dimension compression of feature vectors. A CS model is developed to incorporate acoustic-phonetic knowledge and combine the outputs of subnets to arrive at the overall decision on the class of an input utterance. The models developed for the classification of the isolated utterances of CV units may be useful for spotting CV segments in continuous speech.

References

[1] Steven Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, March 1999.

[2] C. Chandra Sekhar and B. Yegnanarayana, "Modular networks and constraint satisfaction model for recognition of stop_consonant–*vowel* (SCV) utterances," in *Proceedings of International Conforance on Neural Networks, Alaska*, pp. 405–408, May 1998.

[3] C. Chandra Sekhar, *Neural Network Models for Recognition of Stop Consonant-Vowel (SCV) segments in Continuous Speech.* PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, April 1996.

[4] J. Y. Siva Rama Krishna Rao, C. Chandra Sekhar and B. Yegnanarayana, "Neural networks based approach for detection of vowel onset points," in *Proceedings of International Conference on Advances in Pattern Recognition and Digital Techniques (ISI calcutta, India)*, pp. 316–320, December 1999.

[5] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, New Jersey: PTR Prentice Hall, 1993.

[6] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics.*, vol. 59, pp. 291–294, 1988.

[7] Mark A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE*, vol. 37, pp. 233–243, February 1991.

[8] S. P. Kishore and B. Yegnanarayana, "Speaker verification: Minimizing the channel effects using autoassociative neural network models," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (Istanbul)*, pp. 1101–1104, June 2000.

[9] B.Yegnanarayana, *Artificial Neural Networks.* New Delhi: Prentice-Hall of India, 1999.

[10] P.P. Raghu and B. Yegnanarayana, "Supervised texture classification using a probabilistic neural network and constraint satisfaction model," *IEEE Transactions on neural networks*, vol. 9, pp. 516–522, May 1998.

[11] C. Chandra Sekhar and B. Yegnanarayana, "A Constraint Satisfaction Model for Recognition of Stop Consonant-Vowel (SCV) Utterances," *Submitted to IEEE Transactions on Speech and Audio Processing.*

Table 3: Preliminary classification performance of MLFFNN models based on POA. The order is the size of the feature vector used for classification. The entries in columns 3 and 4 represent the percentage of correct classification.

Category (#Training samples/ #Testing samples)	# Classes	Order=50 Uncompressed	Order=19 AANN
Velar(600/120)	20	74.16%	77.50%
Alveolar(600/120)	20	79.16%	80.00%
Dental(600/120)	20	73.33%	73.33%
Bilabial(600/120)	20	71.66%	70.00%
Affricates(600/120)	20	83.33%	82.50%
fricatives(450/90)	15	94.44%	94.44%
Semivowels(600/120)	20	91.66%	92.50%
Nasals(300/60)	10	90.00%	95.00%

Table 4: Preliminary classification performance of MLFFNN models based on MOA. The order is the size of the feature vector used for classification. The entries in columns 3 and 4 represent the percentage of correct classification.

Category (#Training samples/ #Testing samples)	# Classes	Order=50 Uncompressed	Order=19 AANN
UVUA(600/120)	20	92.50%	93.33%
UVA (600/120)	20	69.16%	67.50%
VUA (600/120)	20	85.83%	86.60%
VA (600/120)	20	60.83%	63.33%

Table 5: Preliminary classification performance of MLFFNN models based on Vowel. The order is the size of the feature vector used for classification. The entries in columns 3 and 4 represent the percentage of correct classification.

Category (#Training samples/ #Testing samples)	# Classes	Order=50 Uncompressed	Order=19 AANN
Vowel /a/ (480/96)	16	83.33%	77.08%
Vowel /i/ (480/96)	16	61.45%	61.45%
Vowel /u/ (480/96)	16	76.04%	65.62%
Vowel /e/ (480/96)	16	57.29%	54.16%
Vowel /o/ (480/96)	16	62.50%	65.62%