Neural Network Models for Preprocessing and Discriminating Utterances of Consonant-Vowel Units

Suryakanth V. Gangashetty, A. Nayeemulla Khan,
S.R. Mahadeva Prasanna, and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
Email: {svg,nayeem,prasanna,yegna}@speech.cs.iitm.ernet.in

Abstract - *In this paper, we demonstrate the significance of nonlinear neural network models for compression of feature vectors and also develop classifiers for syllable-like units. We consider the standard 80 Stop Consonant-Vowel units of most Indian languages. This set consists of dynamic sounds and hence require large size feature vector to represent the acoustic characteristics of these units. To develop classifiers with limited training data, it is necessary to compress the size of the feature vector. We show that nonlinear compression by autoassociative neural network model is useful, and is superior to the compression by linear principal component analysis.*

I. Introduction

Among all the subword units of speech, syllable-like units are more relevant from both speech production and perception point of view. Hence it appears that syllable-like units are also relevant from machine recognition point of view [1] [2]. The number of syllable-like units in a language are large (> 5000), and many of them have similar acoustic features. But the number of units that occur most frequently (over 85% of time) are less than, about 450, and the most basic among them are Vowel (V) and Consonant-Vowel (CV) units, which are about 150 (at least for most Indian languages). Representation of each of these units require large size (50-100 dimension) feature vector. The classes of these units are separated by highly nonlinear hypersurfaces in the feature space. Neural network (NN) models are best suited for capturing the hypersurfaces dividing the different classes. Since the number of classes are large, and the number of patterns available for each class are limited, it is difficult to train the network to capture the discriminating hypersurfaces. Therefore it is preferable to reduce the size of the feature vector representing each pattern. Even for reduction of the size of the vector, nonlinear compression techniques are useful, which can be realized using neural network models.

The objective of this paper is to explore neural network models for data reduction and classification of a subset of syllable-like units. We will also study the discrimination characteristics of different subgroups of these units, to show that all units are not equally distinct.

This paper is organized as follows: Speech data used in the study and its representation in the form of feature vectors is explained in Section II. Neural network models used for nonlinear data compression technique are described in Section III. Classification of Stop Consonant-Vowel (SCV) units using multilayer feedforward neural network models is presented in Section IV. The last section gives conclusions from this study.

II. Speech Data and Representation

In this study we consider the important subset of the basic units, namely the Stop Consonant-Vowel (SCV) subset, which for most Indian languages are 80 in number. Stop consonants are the sounds produced by complete closure at some point along the vocal tract, build up pressure behind the closure, and release the pressure by sudden opening. These units and their acoustic characteristics of production are given in Table I. These units are not only important from speech production and perception point of view, but they also carry significant information about the speech message. But some of them are highly confusable (for example, /pa/ & /ba/, /ta/ & /tha/, etc.,). To capture the acoustic characteristics of these units, it is necessary to represent each of these units as a sequence of frames, and extract the spectral information corresponding to each frame. The units have three distinct regions in the production characteristics: the region just before the onset of the vowel, the region immediately after the release of the stop sound, and the

0-7803-7278-6/02/$10.00 ©2002 IEEE

steady vowel region. It is obvious that all these units have a distinct vowel onset point (VOP) in their production [3] [4]. While it is useful to identify the region before VOP to correspond to Manner of Articulation (MOA), and the transition region after VOP to Place of Articulation (POA), and the remaining part to steady Vowel (V), it is difficult to isolate these regions precisely. Moreover, the acoustic characteristics of each will influence the other. Thus all these regions need to be represented together as a single feature vector [3].

TABLE I

List of SCV 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/
Unvoiced Unaspirated (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
Unvoiced Aspirated (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
Voiced Unaspirated (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
Voiced Aspirated (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

Since the vowel region is prominent in the signal due to its large amplitude characteristics and also due to its distinct periodic excitation property, it seems most appropriate to derive the feature vector anchored around the VOP. One frame of 20 msec is considered to the left of VOP and four overlapping frames, each of 20 msec with a shift of 5 msec are considered to the right of VOP. Thus a 55 msec segment anchored around VOP is used to represent each SCV unit. Each frame is represented by 10 weighted linear prediction (LP) cepstral coefficients (WLPCC) which are obtained from an 8^{th} order LP analysis on speech sampled at 8 kHz [5]. Thus the feature vector for each SCV unit is a 50 dimensional vector [6] [7].

Speech data for these studies consists of isolated utterances of the 80 SCV units from three different speakers. The database consists of 10 utterances of each unit from each speaker. The total number of utterances are $3X80X10 = 2400$. The VOPs for all the units are marked manually. For each unit, a 50 dimensional feature vector is derived anchored around the VOP.

III. Neural Network Models for Data Compression

As mentioned earlier, it is preferable to reduce the size of the feature vector in order to develop a suitable classifier for discriminating among different classes. A standard method of data reduction is to use the linear method of principal component analysis (PCA) [8]. But it is known that for data compression, nonlinear PCA performs better than linear PCA [8]. Nonlinear PCA can be implemented using a five layer Autoassociative Neural Network (AANN) model [9] [10] [11]. A five layer AANN model which performs nonlinear PCA is shown in Fig.1 [11]. The structure of the AANN model used in the present studies is *50L 75N pN 75N 50L*, where p is the number of units in the compression layer (order of compression), L refers to a linear unit and N refers to a nonlinear unit. The activation function of the nonlinear unit is a hyperbolic tangent function. The network is trained using error backpropagation algorithm for 1000 epochs [12]. The number of epochs was chosen using cross-validation for verification, to obtain the best performance for this data.

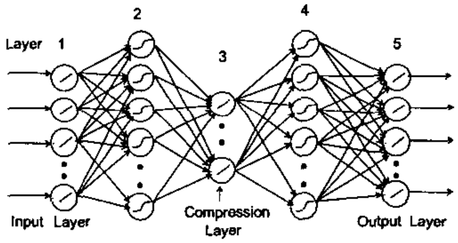

Fig. 1. Five layer AANN model used for nonlinear compression of feature vectors.

The performance of an autoassociation network with compression layer can also be interpreted as a distribution capturing network. The distribution of the feature vectors in the feature space cannot in general be represented by a mixture of Gaussians. The complex non-Gaussian mixture distribution is captured in the AANN model with a compression layer as discussed in [13].

To compare the relative performance of linear and nonlinear PCA methods for data compression, a trained AANN model is used for nonlinear PCA and the standard eigenanalysis for PCA. The recognition performance of SCV units for different compression levels is given in Table II. The table gives the percentage of correct classification of different subsets of the SCV units. The subsets are: 3 Speakers, 5 Vowels, 4 Place of Articulation (POA) and 4 Manner of Articulation (MOA). The structure of the classifier for each of the subsets is a feedforward neural network with *pL 5pN 3nN nN*, where L and N refer to the linear and nonlinear units, p is the number of units in the input layer (equal to the units in the compression layer of the AANN), and n is the number of classes in each subnet. The first hidden layer has *5p* units, and the second hidden layer has *3n* units. Table II also shows the results of classification for uncompressed data, for which $p = m = 50$. The results show that the nonlinear compression of AANN performs better than the compression by linear PCA. The classification performance improves as p is increased. For $p = 18$, the performance is nearly the same as for the uncompressed case. Thus a compression of about one-third (18/50) can be achieved without affecting the classification performance. As discussed earlier, smaller dimension of the feature vector is preferable for developing a classifier with limited amount of training data.

Table II also illustrates that, among different subsets, Speakers and Vowels subsets have better discrimination than POA and MOA subsets. In fact at lower dimension one can visualize the distribution of these features in the data. For example, Fig. 2 shows the distribution of Speakers, Vowels, POA and MOA in the compressed space of $p = 3$. It shows that, when feature vectors are compressed to low dimension, Speaker and Vowel categories are better discriminated than POA and MOA categories. Even between Speakers and Vowels, Vowels have better discrimination than Speakers. This can be seen from the results in Table II for the case $p = 03$. This observation helps to design the compression network for developing a hierarchical classification network.

IV. Classification of SCV Units

For classification of all the 80 SCV units, a single network with either uncompressed or compressed feature vectors performs poorly, as it gives about 45% for uncompressed (size $m = 50$) and about 35% for compressed vectors (size $p = 18$). On the other hand, one can use several classification networks for different groups of SCV units, and then combine the results of the outputs from these networks. For this, we use three more classification networks (in addition to the last three classes in Table II), using the combinations Vowel & POA (20), POA & MOA (16) and MOA & Vowel (20). The values within the paranthesis correspond to the total number of classes within that combination group. The results of all the six networks (for $p = 18$), three for Vowel (5), POA (4) and MOA (4), and three for the pairs, are combined to obtain the overall classification performance for all the 80 SCV units. The evidence for each of the Vowel, POA and MOA is derived from the networks, and then the final classified SCV unit is obtained. Table III gives the performance of all the six networks. The combined evidence gives a percentage classification of about 62.50% as shown in Table IV. Due to confusable nature of these 80 SCV units, the realized performance of 62.50% is still significant. Moreover, in most cases, the next best class is similar to the correct one, with only one of the three components (Vowel, POA or MOA) wrong. In all these cases 480 ($3X80X2$) test patterns are used for testing the classification performance. Table IV also gives the performance when at least two or at least one of the three components in each SCV unit is correct.

It is interesting to note that the SCV classification performance is about 89%, if we consider at least any two components of a sound unit are correct. The performance is over 98%, if we consider at least any one component among the three components (Vowel, POA or MOA) is correct. The block diagram of the proposed system for SCV recognition is shown in Fig.3. The performance at the component level is about 83% for all the 480 test patterns of SCV units. Due to confusable nature of these units, even this performance is significant.

V. Summary and Conclusions

In this paper we have shown that nonlinear compression of feature vectors using AANN models gives performance better than linear PCA for the case of 80 SCV units of typical Indian languages. This study confirms similar results obtained for other classification problems involving compression of feature vectors [8]. Dynamic sound units such as SCV units need large size feature vector to represent the acoustic characteristics of the sound units. The advantage of compressed feature vectors is that a classifier can be developed with limited amount of training data. We have also noticed that for a confusable set like SCV units, all units will not have similar discrimination among the classes. Speakers and Vowels subsets have better discrimination than POA and MOA subsets. One can obtain improved classification accuracy by combining the results from classifiers with different combinations. The overall performance of about 62% for these SCV units is significant, since in continuous speech some of the errors can be corrected using contextual knowledge. In this paper we have discussed the results for isolated utterances of SCV units. It is necessary to develop the compression network and the classifiers for the SCV units occuring in continous speech, in order to determine the significance of the results of this paper for practical application.

TABLE II

Comparison of linear and nonlinear PCA for compression of feature vectors for classification of subsets of SCV units. The entries in the table from columns 3 to 11 represents the percentage of correct classification.

Subset	# Classes(n)	$p = 03$		$p = 09$		$p = 18$		$p = 25$		$m = 50$
		Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	
Speaker	03	62.91%	81.25%	91.66%	91.87%	90.83%	95.62%	91.88%	94.58%	93.30%
Vowel	05	75.83%	83.95%	82.91%	85.41%	86.87%	88.75%	90.63%	90.83%	90.20%
POA	04	46.88%	54.38%	66.04%	70.00%	65.63%	74.17%	67.38%	72.70%	72.29%
MOA	04	45.83%	50.20%	64.37%	68.75%	65.00%	69.38%	67.90%	68.75%	70.20%

References

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

[2] P. Eswar, S.K. Gupta, C. Chandra Sekhar, B. Yegnanarayana and K. Nagamma Reddy, "An acoustic-phonetic expert for analysis and processing of continuous speech in Hindi," in *Proc. European Conf. Speech Technology, Edinburgh*, pp. 369–372, Sept. 1987.

[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, Apr. 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 *Proc. Int. Conf. Advances in Pattern Recognition and Digital Techniques (ISI calcutta, India)*, pp. 316–320, Dec. 1999.

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

[6] Donal G. Sinex and C. Daniel Geisler, "Responses of auditory-nerve fibres to consonant-vowel syllables," *J. Acoust. Soc. Am.*, vol. 73(2), pp. 602–615, Feb. 1983.

[7] Sadaoki Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, vol. 80(4), pp. 1016–1025, Oct. 1986.

[8] K.I. Diamantaras and S.Y. Kung, *Principal Component Neural Networks, Theory and Applications.* 605 Third Avenue, New York, NY: John Wiley & Sons, Inc., 1996.

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

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

[11] S. P. Kishore and B. Yegnanarayana, "Speaker verification: Minimizing the channel effects using autoassociative neural network models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1101–1104, Jun. 2000.

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

[13] S. P. Kishore, B. Yegnanarayana, and Suryakanth V. Gangashetty, "Online text-independent speaker verification system using autoassociative neural network models," in *Proc. Int. Joint Conf. Neural Networks (Washington DC, USA)*, vol. 2(4), pp. 1548–1553, Jul. 2001.

0-7803-7278-6/02/$10.00 ©2002 IEEE

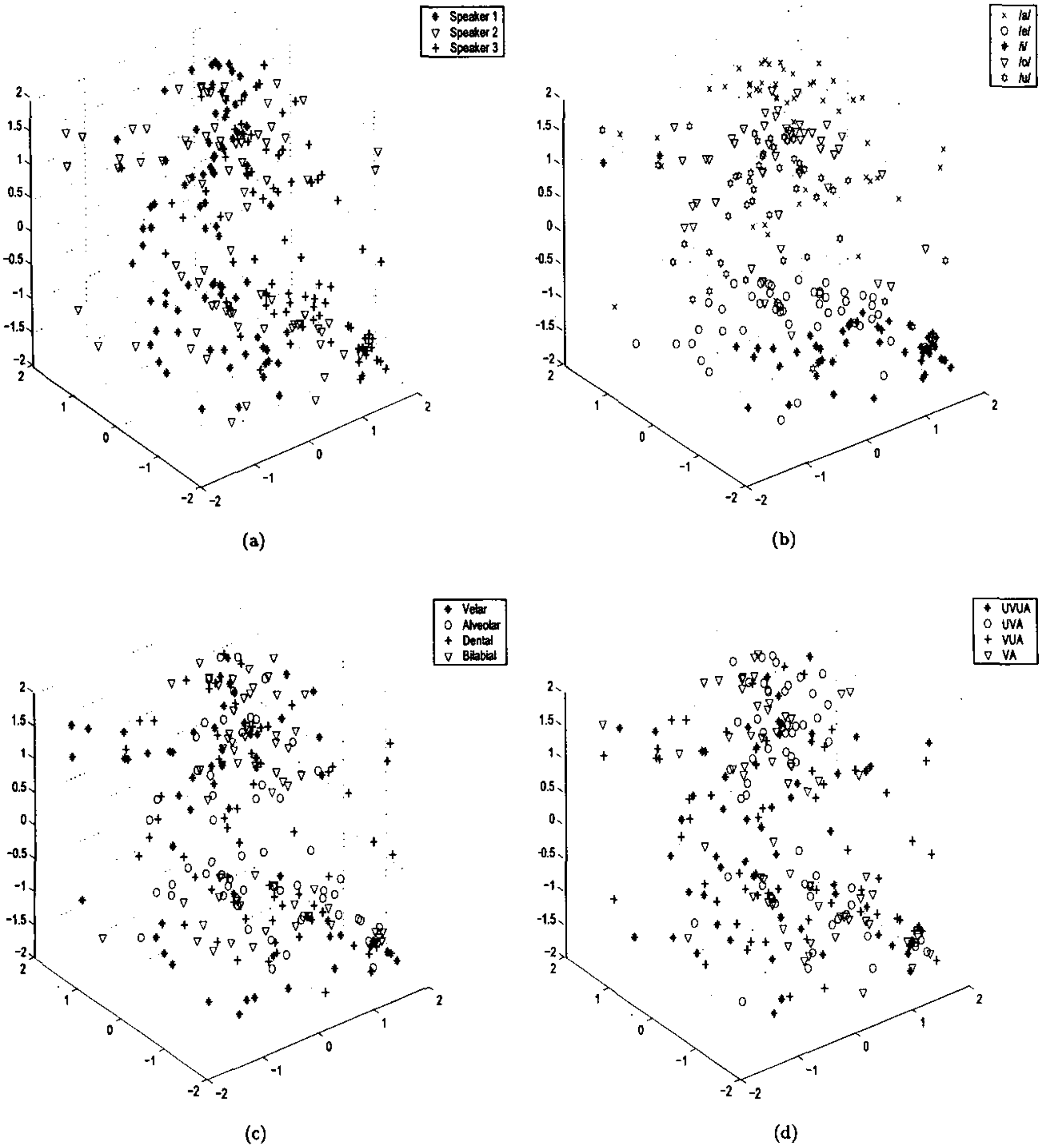

Fig. 2. Distributions of 3-D compressed vectors for different subsets of SCV units.

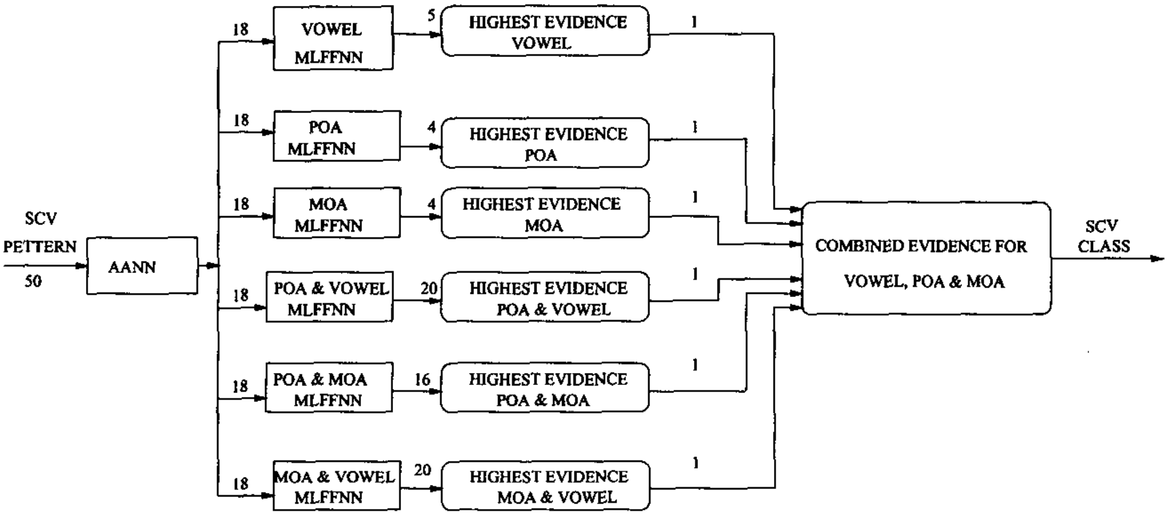

Fig. 3. Block diagram of the proposed model for recognition of SCV units. The numbers represent the number of features or classes.

TABLE III

Classification performance of 6 classifiers.

Category	# Classes	Number of test patterns identified correctly	Percentage of correct classification
Vowel	05	426	88.75%
POA	04	356	74.17%
MOA	04	333	69.38%
POA & Vowel	20	357	74.37%
POA & MOA	16	313	65.20%
MOA & Vowel	20	357	74.37%

TABLE IV

Classification performance of the SCV recognition subsystem that uses the 3 Vowel, 3 POA and 3 MOA evidences obtained from the six classifiers given in Table III.

Percentage of classification of SCV units when		
(a) All the three components correct	:	62.50%
(b) At least two components correct	:	89.58%
(c) At least one component correct	:	98.33%

0-7803-7278-6/02/$10.00 ©2002 IEEE