A Constraint Satisfaction Model for Recognition of Stop Consonant–Vowel (SCV) Utterances

C. Chandra Sekhar and B. Yegnanarayana

***Abstract*—In this paper, we propose a model for recognition of utterances of consonant–vowel (CV) units. The acoustic–phonetic knowledge of the CV classes is incorporated in the form of constraints of a constraint satisfaction model. The model combines evidence from multiple classifiers. The significant feature of this model is that discrimination of the CV units could be enhanced by a combination of even weak evidence derived from the features. The evidence is obtained from multilayer feedforward neural networks trained for subgroups of CV classes. The evidence is enhanced using a set of feedback subnetworks in the constraint satisfaction model. The weights for the connections in the feedback subnetworks are derived using acoustic-phonetic knowledge and the performance statistics of the trained networks. The performance of the proposed model is demonstrated for recognition of utterances of a large number (80) of stop consonant-vowel units for the Indian language Hindi.**

***Index Terms*—Consonant–vowel units, constraint satisfaction model, neural networks, speech recognition.**

I. Introduction

IN THIS PAPER, we propose a constraint satisfaction neural network model for recognition of consonant–vowel (CV) units of speech. CV units occur frequently in normal speech and recognition of these units is crucial for development of any speech recognition system. Moreover, they are also natural units of speech production in the sense that, typically most syllables are of CV type [1]. Human beings are able to extract the relevant parameters or features from the speech signal and recognize them effortlessly most of the time [2]. The remarkable characteristic is that they are able to do this in a speaker independent manner even in adverse environmental conditions. This ability of humans may be attributed to the knowledge they have acquired about the sound units, besides the sophisticated auditory processing and neural classification mechanism. While developing a speech recognition system, this knowledge factor must be kept in mind.

Among the CV units that occur in a text of the Indian language Hindi, about 45% of the units belong to the category of stop consonant–vowel (SCV) units [3]. In this paper, we propose a new approach for developing a recognition system for SCV units. It takes into account the constraints at various levels:

1) *language level:* legal sound units, relative frequency of occurrence;
2) *acoustic-phonetic level:* speech production features [4];
3) *signal level:* parameters from the speech signal;
4) *dynamic constraints:* discrimination among units within a group.

Some of the knowledge at various levels can be represented through a constraint satisfaction model [5]–[7].

The next section describes the overall approach used for developing the neural network model for SCV recognition. Section III deals with the preparation of speech data of SCV units for this study. It also describes preprocessing and parametric extraction stages. Section IV presents the modular networks for classification of SCV units using different grouping criteria based on speech production. A constraint satisfaction model is developed incorporating the acoustic–phonetic knowledge as weak constraints. The model is described in detail in Section V. Section VI describes the operation of the constraint satisfaction model. The results of recognition of SCV units uttered in isolation are given in Section VII.

II. Proposed Approach for Recognition of SCV Units

Speech data for these studies consists of SCV units of a typical Indian language. The SCV units for the language Hindi [8] are organized along three broad categories, namely, the manner of articulation (MOA) of the consonant, the place of articulation (POA) of the consonant and the vowel. There are four manners of articulation: unvoiced unaspirated (UVUA), unvoiced aspirated (UVA), voiced unaspirated (VUA), and voiced aspirated (VA). There are four places of articulation for Hindi stop consonants: velar, alveolar, dental, and bilabial. The five vowel categories used in our studies are: /a/, /i/, /u/, /e/, and /o/. Combinations of all MOAs, POAs, and vowel categories for forming SCV units result in 80 different SCV classes.

Normally each of the SCV classes has unique production characteristics and when uttered in isolation these production characteristics are well reflected in the resulting speech signal. However, due to closeness of the shapes of the vocal tract for some of these classes, it is difficult to extract the discriminant features by processing the speech signal alone. The acoustic–phonetic knowledge of the SCV units suggests that these units can be grouped into different subgroups, so that a modular network can be developed for the classifier. Each module is designed to classify the units in that subgroup having fewer units. The advantage is that the complexity of the classifier will be reduced. However, for these modular networks to be successful, one needs to know the subgroup identity of a given SCV unit from the input speech. Table I gives the grouping of the 80 SCV units of Hindi into different subgroups based on three different criteria, namely, manner of articulation (MOA), place of articulation (POA), and vowel (V). We refer

Manuscript received February 9, 2000; revised June 3, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rafid A. Sukkar.

The authors are with the Department of Computer Science and Engineering, Indian Institute of Technology, Madras, Chennai-600036, India (e-mail: chandra@cs.iitm.ernet.in; yegna@cs.iitm.ernet.in).

Digital Object Identifier 10.1109/TSA.2002.804298

TABLE I
LIST OF SCV CLASSES AND THE SUBGROUPS BASED ON DIFFERENT GROUPING CRITERIA FOR EACH CLASS

MOA	POA	Vowel subgroup				
Subgroup	Subgroup	/a/	/i/	/u/	/e/	/o/
UVUA	Velar	ka	ki	ku	ke	ko
	Alveolar	ṭa	ṭi	ṭu	ṭe	ṭo
	Dental	ta	ti	tu	te	to
	Bilabial	pa	pi	pu	pe	po
UVA	Velar	kha	khi	khu	khe	kho
	Alveolar	ṭha	ṭhi	ṭhu	ṭhe	ṭho
	Dental	tha	thi	thu	the	tho
	Bilabial	pha	phi	phu	phe	pho
VUA	Velar	ga	gi	gu	ge	go
	Alveolar	ḍa	ḍi	ḍu	ḍe	ḍo
	Dental	da	di	du	de	do
	Bilabial	ba	bi	bu	be	bo
VA	Velar	gha	ghi	ghu	ghe	gho
	Alveolar	ḍha	ḍhi	ḍhu	ḍhe	ḍho
	Dental	dha	dhi	dhu	dhe	dho
	Bilabial	bha	bhi	bhu	bhe	bho

to each type of SCV unit as a class. Hence, the problem is to develop a classifier for these 80 SCV classes.

An SCV unit occurs in a different combination of other units in a subgroup for each grouping criteria. The classifier developed for each subgroup captures discriminating features of the units in the subgroup. Thus, the output for an SCV unit from the classifiers based on the three grouping criterion can be viewed as evidence from three different classifiers.

The objective of this paper is to propose a constraint satisfaction (CS) model [5]–[7] that takes the outputs of the three classifiers as input and combines it with the production knowledge of the SCV units incorporated in the model as constraints. 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 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 in the parametric form.

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 multilayer feedforward neural networks (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 SCV 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 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 interactive activation and competition (IAC) model [7]. 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 SCV recognition is shown in Fig. 1.

III. PREPROCESSING AND REPRESENTATION OF SPEECH DATA

We consider speech data corresponding to isolated utterances of the SCV units. For each SCV unit, 12 repetitions of isolated utterances of the unit are collected for each of three male speakers. Out of these, four for each unit and speaker are used as the training set 1 and four others for each unit and speaker are used as the training set 2. All the speech data is collected in a laboratory environment using a sampling frequency of 10 kHz and 16 bits per sample.

Speech data for each unit is processed as follows: The point at which the consonant ends and the vowel begins in an SCV unit is defined as the vowel onset point (VOP). The VOP for each SCV unit is identified using the method given in [9]. We consider 60 ms of data before and 140 ms of data after the VOP for analysis. The 200 ms segment of speech data is analyzed frame by frame, with each frame of duration 20 ms and with a frame shift of 5 ms. Each frame of data is represented using 12 weighted cepstral coefficients derived from eight linear prediction coefficients [10]. The number of cepstral coefficients is larger than the order of the linear prediction (LP) analysis. This will help representing the linear prediction spectrum better. The cepstral coefficient vectors of adjacent frames are averaged. Thus, each SCV unit is represented by $20 \times 12 = 240$ parameters. It should be noted that parametric representation is crucial in the sense that we must ensure minimum loss of information in order to obtain good discrimination among SCV classes. LP derived cepstral coefficients were found to be suitable parameters for speech recognition studies [10]. We have also conducted a separate study to evaluate different parametric representations for classification of the place of articulation of the consonant in the SCV units based on the transition region (of 30 ms after the VOP in an SCV unit) [11]. Table II shows the results of this study. The improved performance for the weighted LP cepstral coefficients over the mel-scale cepstral coefficients can be at-

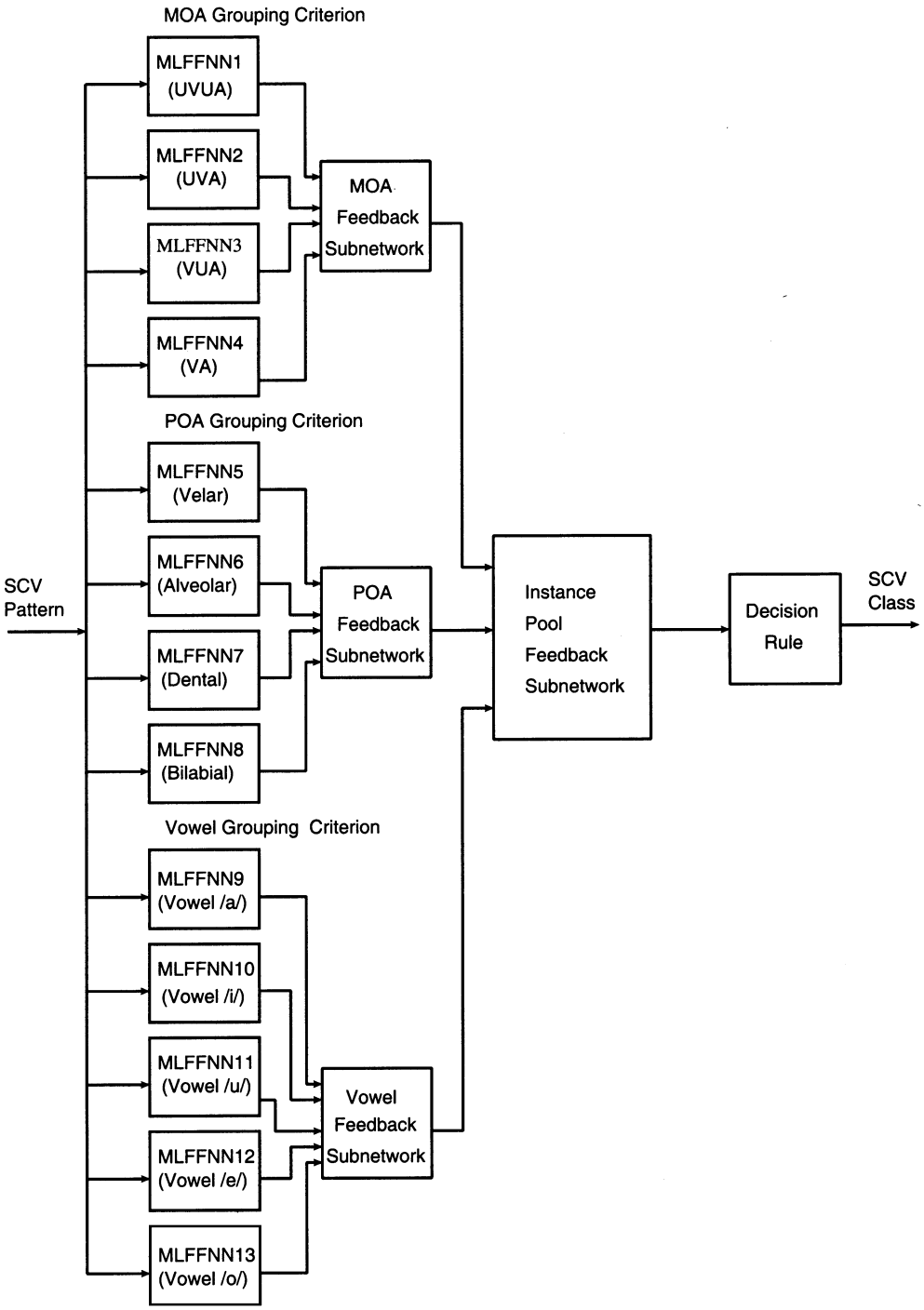


Fig. 1. Block diagram of the proposed system for recognition of SCV units.

tributed to the fact that LP coefficients preserve the formant transitions better than the mel-scale spectral or cepstral parameters.

In the next section, we discuss the development of feature extraction for SCV units using trained neural networks.

IV. Neural Networks as Nonlinear Feature Extractors

We develop a MLFFNN, which, after training, is interpreted as a nonlinear filter. The filter is designed in such a way that it provides discrimination among the classes within the subgroup used for training the network. One such network is used for each subgroup consisting of 16 or 20 classes depending on the grouping criterion used for organizing the SCV units. The list of 80 SCV units and the subgroups based on different grouping criteria are given in Table I. For example, the class /ka/ belongs to UVUA subgroup based on MOA, "velar" subgroup based on POA and "/a/" subgroup based on vowel.

It may be noted that if we train the neural network using samples for one class only, i.e., a single output unit network, then the resulting network does not capture the discrimination characteristics with respect to other classes. On the other hand, if the number of classes is large and the classes are close, it may

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT PARAMETRIC REPRESENTATIONS FOR CLASSIFICATION OF THE PLACE OF ARTICULATION OF THE CONSONANT IN SCV UNITS USING THE TRANSITION REGION

Parametric representation	Classification accuracy (in %)
Formant frequencies	29.8
Mel-scale spectral coefficients	43.5
Mel-scale cepstral coefficients	50.5
Weighted LP cepstral coefficients	58.6

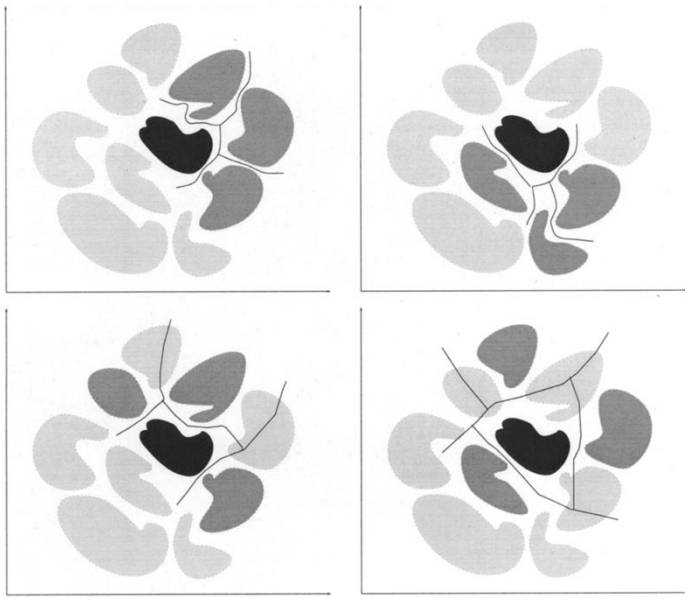


Fig. 2. Illustration of the effect of grouping a class with different subsets of classes on the decision surfaces formed. The three classes with intermediate level of shading only are used for grouping with the class of interest (dark shading).

be difficult to capture the discrimination by a single network. Therefore, a modular network derived by grouping the units into subgroups is a compromise among the conflicting requirements [12].

The shape of the decision surface formed for a class by an MLFFNN depends on the classes in a subgroup. This behavior is illustrated in Fig. 2 for an arbitrary two-dimensional (2-D) pattern space. In this figure the regions for 10 different classes are shown. The region for the class under consideration is shown in dark shade. When this class is grouped with sets of three other classes (shown by intermediate level shading) only, then the decision surfaces formed for the class may vary as illustrated for four cases in Fig. 2. Therefore the performance of MLFFNNs on the test data of a class can vary for different grouping criteria.

The structure of the MLFFNN used in these studies consists of 240 nodes in the input layer, 70 nodes in the first hidden layer, and 50 nodes in the second hidden layer [3]. The results of the trained MLFFNNs as classifiers for each of the SCV classes in their subgroups is given in Table III. Each entry in the table shows the percentage of the total number of test patterns of an SCV class that are correctly classified by the MLFFNN. It can be noted from Table III that the MLFFNNs of the three grouping criteria do not give the same performance for many classes.

TABLE III
CLASSIFICATION PERFORMANCE OF MLFFNNS BASED ON DIFFERENT GROUPING CRITERIA (MOA: MANNER OF ARTICULATION, POA: PLACE OF ARTICULATION, V: VOWEL) FOR TEST DATA OF EACH SCV CLASS

SCV Class	Grouping Criterion MOA	POA	V	SCV Class	Grouping Criterion MOA	POA	V
ka	58	83	42	ṭa	92	92	83
kha	92	67	92	ṭha	92	83	58
ga	67	83	83	ḍa	83	67	67
gha	50	50	25	ḍha	67	75	58
ki	58	50	58	ṭi	75	83	58
khi	50	67	58	ṭhi	67	100	58
gi	42	33	67	ḍi	67	67	42
ghi	33	42	50	ḍhi	67	67	58
ku	67	75	83	ṭu	83	100	92
khu	75	58	67	ṭhu	58	92	67
gu	50	50	25	ḍu	67	92	75
ghu	33	17	33	ḍhu	58	67	58
ke	75	67	50	ṭe	92	92	100
khe	58	42	42	ṭhe	67	92	75
ge	50	42	67	ḍe	50	92	58
ghe	58	42	42	ḍhe	92	92	67
ko	75	67	92	ṭo	75	83	75
kho	67	58	75	ṭho	75	92	50
go	42	42	67	ḍo	83	83	83
gho	67	58	17	ḍho	75	75	67
ta	100	100	100	pa	92	92	75
tha	42	58	42	pha	92	100	58
da	67	67	42	ba	83	83	75
dha	67	92	50	bha	67	75	33
ti	92	75	67	pi	50	42	67
thi	75	100	92	phi	92	83	92
di	75	83	58	bi	75	75	75
dhi	58	83	42	bhi	42	100	92
tu	75	83	67	pu	83	67	83
thu	67	67	50	phu	42	67	75
du	75	42	75	bu	92	83	75
dhu	42	75	92	bhu	75	58	83
te	75	92	83	pe	83	100	67
the	75	100	58	phe	67	92	83
de	42	58	8	be	58	58	33
dhe	50	100	75	bhe	75	83	83
to	67	75	67	po	75	92	75
tho	75	75	83	pho	75	92	92
do	42	67	58	bo	50	75	58
dho	42	67	50	bho	83	83	75

The MLFFNN trained for classes in a subgroup can be viewed as a set of class dependent filters, where the filter characteristics for a class are designed to discriminate that class against the other classes in that subgroup. Thus, there are 16 or 20 filters in each subgroup and 80 filters for each grouping criterion. It may be noted that each SCV class occurs with a different subset of SCV classes for each of the three groupings. We can also interpret the network as a filter set tailored to the classes in a subgroup. This is like Gabor filters used for texture classification where the filters are tailored to the characteristics of the texture classes under consideration [13]. The characteristics to be optimized in the case of Gabor filters are resolution, orientation, and spatial frequency.

Normally, the trained MLFFNNs can be used directly as classifiers for the subgroups of classes. However, the filter interpretation provides greater flexibility and robustness in the development of a classifier for all the SCV classes. Once the MLFFNNs are trained, then they are used as nonlinear filters. The outputs of the filters for each subgroup for a given training pattern are used to form a feature vector. The distribution of the feature vectors is obtained for each class from the training set 2. The distribution is represented in terms of a mean vector and a variance parameter derived from the feature vectors for the class.

The outputs of the sets of filters designed in this section are given as input to the feedback subnetworks of the constraint satisfaction model. The next section will describe the feedback subnetworks and explain the method of determining the weights for the connections in the feedback subnetworks. These weights represent the constraints and they are derived using the acoustic-phonetic knowledge and the performance statistics of the MLFFNNs.

V. Feedback Subnetworks for Different Grouping Criteria

We first build three different feedback subnetworks, one for each of the three grouping criteria. Since the SCV classes within a subgroup have been designed to compete among themselves during training of the MLFFNN for that subgroup, we provide excitatory connections between the nodes corresponding to the classes within a subgroup. All the connections across the subgroups are made inhibitory. The weights for the excitatory and inhibitory connections have been derived from the confusion matrices obtained from the classification performance of the MLFFNNs.

The confusion matrices for different manners of articulation, places of articulation and vowels are given in Table IV. The rounded values in the parentheses are interpreted as (symmetric) similarity measures. For example, the similarity between UVUA and UVA is indicated as 0.03 which is the rounded value of the average of the two entries for UVUA and UVA in the confusion matrix, i.e., $((3.1 + 2.5)/2)/100$.

The similarity measures are used to determine the weights for the excitatory and inhibitory connections in the feedback subnetworks. An excitatory connection is provided between nodes of two SCV classes within a subgroup if they differ in only MOA or POA or vowel characteristic. The weight of an excitatory connection is equal to the similarity measure between the differing production features of the two classes. For example, in grouping based on MOA, the class /ka/ belongs to the UVUA subgroup. Of the 20 classes present in this subgroup (/ka/, /ṭa/, /ta/, /pa/, /ki/, /ṭi/, /ti/, /pi/, /ku/, /ṭu/, /tu/, /pu/, /ke/, /ṭe/, /te/, /pe/, /ko/, /ṭo/, /to/, and /po/), an excitatory connection is provided between /ka/ and each of the following seven classes only: /ṭa, /ta, /pa/, /ki/, /ku/, /ke/, and /ko/. The remaining 12 classes in this subgroup differ with /ka/ in both POA and vowel and hence no connection is provided between the nodes of /ka/ and these 12 classes. The weight for the excitatory connection between /ka/ and /ku/ is 0.02, which is the similarity measure between the vowels /a/ and /u/ as given in Table IV(c).

An inhibitory connection is provided between nodes of the classes in different subgroups only if the two classes differ either in MOA or POA or vowel only. For the earlier example of class /ka/ in the grouping based on MOA, an inhibitory connection is provided between /ka/ in the UVUA subgroup and each of the following classes: /kha/ in UVA, /ga/ in VUA and /gha/ in the VA subgroup. All the other classes in the UVA, VUA, and VA subgroups differ with /ka/ not only in MOA but also in POA or/and vowel. The weight for an inhibitory connection is inversely proportional to the similarity measure between the differing production features of the two classes. If the similarity measure is C (in the range 0.0 to 1.0), then the inhibitory weight W is assigned as follows:

$$W = -\frac{1}{100 * C}. \tag{1}$$

If C is less than 0.01, then the corresponding inhibitory weight is assigned as -1.0. The weights of the connections for the class /ka/ in the feedback subnetworks for different grouping criteria are given in Table V.

The connections in the feedback subnetwork for the grouping criterion of POA are illustrated in Fig. 3. The excitatory connections for the class /ka/ in the 'Velar' subgroup are shown in Fig. 3(a) and the inhibitory connections for the class are shown in Fig. 3(b).

TABLE IV
Confusion Matrices for Different Manners of Articulation, Places of Articulation and Vowels. The Values in the Parentheses are Interpreted as Similarity Measures Derived From the Confusion Matrix

(a) Confusion matrix for different manners of articulation.

MOA	UVUA	UVA	VUA	VA
UVUA	86.9 (0.87)	2.5 (0.03)	7.7 (0.06)	2.9 (0.02)
UVA	3.1 (0.03)	84.2 (0.84)	4.4 (0.04)	8.3 (0.08)
VUA	4.2 (0.06)	3.3 (0.04)	78.3 (0.78)	14.2 (0.12)
VA	0.4 (0.02)	7.9 (0.08)	9.6 (0.12)	82.1 (0.82)

(b) Confusion matrix for different places of articulation.

POA	Velar	Alveolar	Dental	Bilabial
Velar	72.1 (0.72)	9.0 (0.08)	7.1 (0.08)	11.8 (0.08)
Alveolar	6.0 (0.08)	75.4 (0.75)	7.3 (0.10)	11.3 (0.09)
Dental	8.1 (0.08)	12.9 (0.10)	67.7 (0.68)	11.3 (0.10)
Bilabial	4.2 (0.08)	7.5 (0.09)	8.5 (0.10)	79.8 (0.80)

(c) Confusion matrix for different vowels.

Vowel	/a/	/i/	/u/	/e/	/o/
/a/	89.3 (0.90)	1.0 (0.01)	2.9 (0.02)	0.8 (0.01)	6.0 (0.05)
/i/	0.5 (0.01)	85.7 (0.86)	3.1 (0.02)	9.9 (0.08)	0.8 (0.00)
/u/	1.3 (0.02)	1.6 (0.02)	82.3 (0.82)	1.3 (0.01)	13.5 (0.16)
/e/	1.3 (0.01)	6.8 (0.08)	1.3 (0.01)	90.3 (0.90)	0.3 (0.00)
/o/	4.4 (0.05)	0.0 (0.00)	18.8 (0.16)	0.0 (0.00)	76.8 (0.77)

TABLE V
ILLUSTRATION OF WEIGHTS OF CONNECTIONS FOR CLASS /KA/ IN THE FEEDBACK SUBNETWORKS FOR DIFFERENT GROUPING CRITERIA

Grouping Criterion	Excitatory Connections		Inhibitory Connections	
	Class	Weight	Class	Weight
MOA	/ṭa/	0.08	/kha/	-0.33
	/ta/	0.08	/ga/	-0.16
	/pa/	0.08	/gha/	-0.50
	/ki/	0.01		
	/ku/	0.02		
	/ke/	0.01		
	/ko/	0.05		
POA	/kha/	0.03	/ṭa/	-0.125
	/ga/	0.06	/ta/	-0.125
	/gha/	0.02	/pa/	-0.125
	/ki/	0.01		
	/ku/	0.02		
	/ke/	0.01		
	/ko/	0.05		
Vowel	/ṭa/	0.08	/ki/	-1.0
	/ta/	0.08	/ku/	-0.5
	/pa/	0.08	/ke/	-1.0
	/kha/	0.03	/ko/	-0.2
	/ga/	0.06		
	/gha/	0.02		

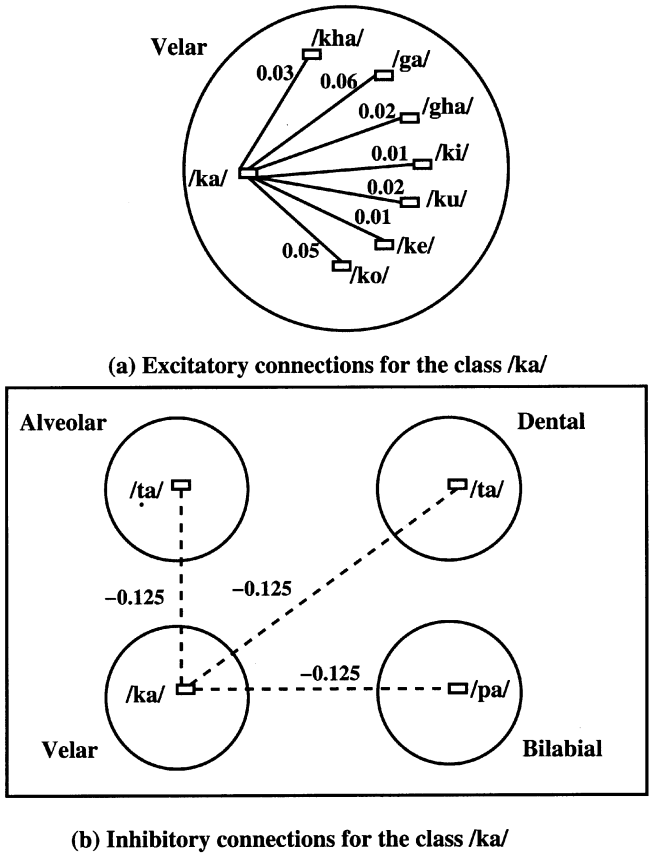


Fig. 3. Connections for the class /ka/ in the POA feedback network. (a) Excitatory connections for the class /ka/ in the "velar" subgroup. (b) The inhibitory connections for the class /ka/.

The main function of each feedback subnetwork is to enhance the evidence available from the filters for the class of the input utterance by giving positive contributions from the evidence for the classes close to it in a subgroup and to reduce the evidence for the classes which are in the other subgroups but are close to it. The weights of the connections based on similarities among classes help the feedback subnetwork to perform its function as a constraint satisfaction network.

Each node in a feedback subnetwork is associated with a mean vector μ and a variance parameter σ^2 representing the distribution of the feature vectors for the class of the node. We assume a symmetric Gaussian distribution, which can be described by the mean vector and the diagonal variance matrix. The mean vector and the variance parameter are obtained from the training set 2. A training pattern belonging to the class of the unit is given as input to the MLFFNN for the subgroup containing the class. The output of the MLFFNN is used to form a feature vector. The dimension of the feature vector is the same as the number of classes in the subgroup. If $\mathbf{y}_i$ is the feature vector obtained for the ith training pattern and N is the number of training patterns for each class, then the kth element of the mean vector, $\boldsymbol{\mu}_k$, is computed as follows:

$$\boldsymbol{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} y_{ik} \tag{2}$$

where y_{ik} is the kth element of $\mathbf{y}_i$. The variance parameter σ^2 is computed from the mean vector and the feature vectors for the N training patterns as follows:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}(y_{ik} - \boldsymbol{\mu}_k)^2 \tag{3}$$

where M is the dimension of the feature vectors.

The mean vector and the variance parameter describe a symmetric Gaussian distribution and they are computed for each of the 80 SCV classes and for each of the three grouping criteria during the second level of training. For the classification of an SCV utterance, the pattern belonging to the utterance is given as input to all the MLFFNNs. The outputs of the MLFFNNs are given as input to the feedback subnetwork corresponding to that grouping.

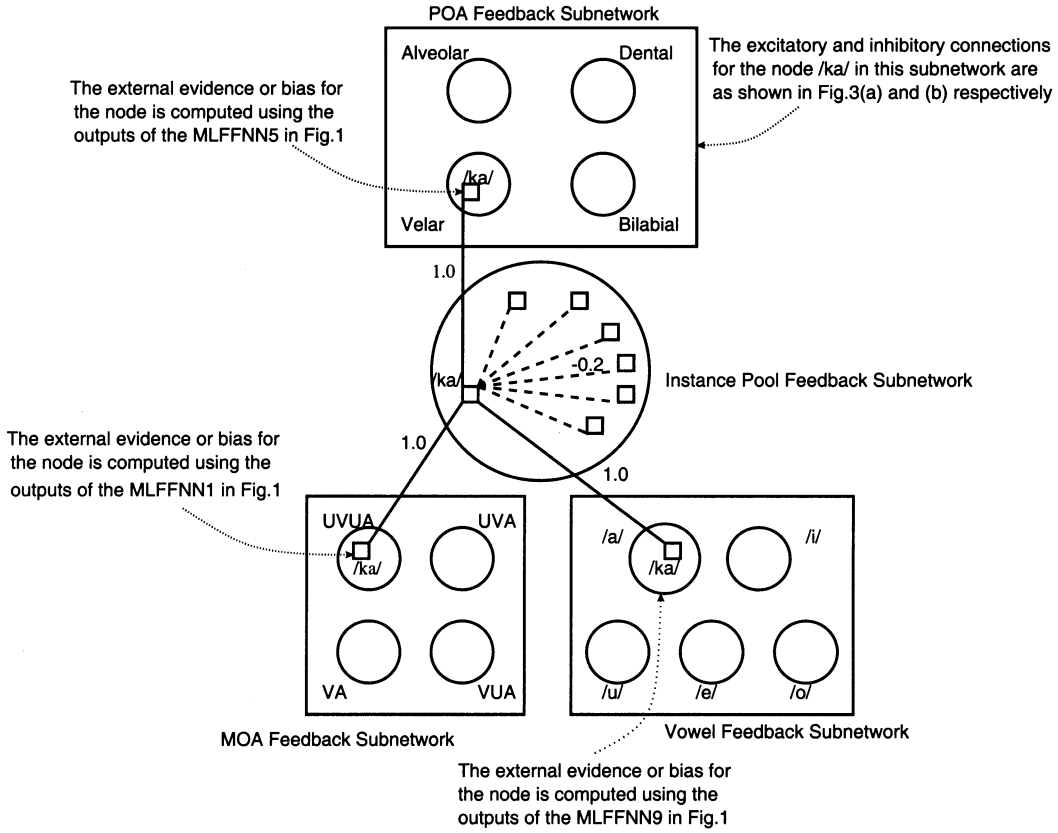


Fig. 4. Constraint satisfaction model for classification of SCV utterances. The model consists of three feedback subnetworks for the three grouping criteria and an instance pool through which the three feedback subnetworks interact. Only connections for the class /ka/ are shown for illustration.

VI. Constraint Satisfaction Model for Classification of SCVs

The three feedback subnetworks corresponding to the three different grouping criteria interact with each other through a pool of nodes in another feedback subnetwork, called instance pool [14]. There are as many (80) nodes in the instance pool as the total number of SCV classes. Each node in the instance pool has a bidirectional excitatory connection with the corresponding nodes in each of the feedback subnetworks. For example, the node corresponding to the class /ka/ in the instance pool has a bidirectional connection to the nodes corresponding to /ka/ in MOA, POA, and vowel feedback subnetworks, as shown in Fig. 4. The nodes within the instance pool compete with each other and hence are connected by a negative weight. Although the choice of value of this weight is not critical, a value of -0.2 was found suitable from experimental studies. The weight typically depends on how many of the other nodes in the instance pool contribute to the activation of a given node and the extent of their contribution, which depends on the output values of those nodes. Normally, the sum of weights (both excitatory and inhibitory) from all active nodes should be nearly zero.

The three feedback subnetworks and the instance pool subnetwork constitute the constraint satisfaction (CS) model reflecting the known speech production knowledge of the SCVs, as well as the knowledge derived from the trained MLFFNNs for different grouping criteria. The CS model developed for classification of SCVs is shown in Fig. 4. There are four feedback subnetworks and all the connections in the network are bidirectional. Note that only the excitatory connections (solid lines) linking the nodes corresponding to /ka/ across all the four feedback subnetworks are shown in the figure. A few of the inhibitory connections (dashed lines) are shown in the instance pool subnetwork. The connections within each subnetwork are as shown in Fig. 3(a) and (b), where the connections in the feedback subnetwork for the POA grouping criterion are given.

The outputs of the MLFFNNs corresponding to different subgroups in Fig. 1 are used to compute the external evidence or bias, which is used as external input to the constraint satisfaction model. The external input for each of the nodes in the feedback subnetworks is derived from the 16- or 20-dimensional feature vector of the MLFFNN to which the unit belongs. For example, the external input or bias to the node /ka/ in the POA feedback subnetwork is computed using the 20-dimensional output feature vector ($\mathbf{x}$) of the MLFFNN5 in Fig. 1. A mean vector $\boldsymbol{\mu}$ and a variance parameter σ^2 are associated with each node in the three feedback subnetworks. The mean vector is derived from the training set 2 as discussed in Section V. The bias or external input for a node in the feedback subnetworks is given by

$$b = \frac{1}{\sqrt{(2\pi)^M \sigma^2}} e^{-d/2} \tag{4}$$

where

$$d = \frac{|\mathbf{x} - \boldsymbol{\mu}|^2}{\sigma^2} \tag{5}$$

and

$$|\mathbf{x} - \boldsymbol{\mu}|^2 = \frac{1}{M} \sum_{i=1}^{M} (x_i - \boldsymbol{\mu}_i)^2 . \tag{6}$$

Here, M is the dimension of the feature vector and x_i and $\boldsymbol{\mu}_i$ are the ith elements of the feature vector $\mathbf{x}$ and the mean vector $\boldsymbol{\mu}$, respectively.

Each node in the CS model also computes the weighted sum (a) of the inputs from the other nodes in the model. The net input (z) to a node in the feedback subnetworks is given by

$$z = \alpha * b + \beta * a. \tag{7}$$

The constants α and β determine the relative importance given to the external evidence (b) and to the *a priori* knowledge in the form of constraints reflected in the weighted sum (a) from other nodes in the model. If the external evidence is strong, such as for the case of features extracted from clean speech, then the value of α may be made large, closer to 1.0. If the speech production knowledge is captured well in the weights of feedback subnetworks, then the value β may be made large, closer to 1.0. While the choice of values for α and β is not very critical, some experimentation is useful to determine suitable values for them. We have chosen $\alpha = 0.5$ and $\beta = 0.5$ in our studies.

A sigmoid activation function was chosen for all the nodes in the CS model. The output of a node using the sigmoid function is given by

$$s = f(z) = \frac{1}{1 + e^{-k(z-\theta)}} \tag{8}$$

where $0 < k < \infty$ is the slope of the sigmoid function and θ is the threshold on the activation value z of the node. Larger value of k results in lesser ambiguity in the final output, but the network may get stuck at some local minimum state of the energy landscape of the feedback network. Here, the terms state and energy refer to those of a Hopfield-type feedback network [5], [15]. From our studies, we have found that $k = 1.0$ was adequate. The value of the threshold θ is to be chosen in such a way that the average value of $(z - \theta)$ is close to zero for all the training data. The value of $\theta = 0.3$ was found to be adequate for our studies.

The constraint satisfaction model is initialized as follows: When a new pattern is presented to the MLFFNNs, the feature vectors ($\mathbf{x}$) for all the MLFFNNs are obtained. Note that for each component of the feature vector there is a corresponding node in the three feedback subnetworks. The values of the components of the feature vectors are examined for the utterances in the training set 2 to determine the average value of the feature vector component corresponding to the class of the input data. The average value for all the classes gives an idea of the value of the threshold (δ), which is used to initialize the outputs of all the nodes in the three feedback subnetworks. The outputs of the nodes for which the corresponding feature vector component exceeds the threshold are initialized to $+1.0$ and the outputs of all other nodes in the three feedback subnetworks are initialized to 0.0. A threshold value of $\delta = 0.3$ was chosen in our studies based on observation of feature vector components for the training set 2. The bias for a node in the instance pool subnetwork is computed from the net input to the node, using the initialized values for the outputs of the nodes in the three feedback subnetworks. The output of a node in the instance pool is initialized to $+1.0$, if the net input to the node is greater than 0.0.

After initialization, the constraint satisfaction model is allowed to relax until a stable state is reached for a given input pattern. Deterministic relaxation method is used in this study [13]. In this method, a node in the model is chosen at random and its output is computed as shown in the (8). The state of the model, represented by the outputs of all the nodes in the model, is changed due to the update of output of any one node. All the nodes in the CS model are considered, one at a time at random, to complete one cycle of iteration. The state update is continued for several cycles until there is no significant change in the state of the model, i.e., in the outputs of all the nodes in the model. Usually a stable state is reached within 10 to 15 cycles. At a stable state of the model, the outputs of the nodes in the instance pool are interpreted to determine the class of the input pattern.

If the feature vectors for an input pattern from the MLFFNNs are considered as the evidence for the classes obtained using different grouping criteria, then the outputs of the instance pool nodes in the final stable state of the model can be considered as the combined evidence for each class, after satisfying as many constraints as possible. The class label of the node in the instance pool with the largest output value is assigned to the input pattern. Because of similarity among several SCV classes, we consider the cases in which the correct class can be among the classes corresponding to the K largest output values. In the next section, we present the classification results of the CS model for Case_1, Case_2, Case_3, and Case_4, corresponding to $K = 1$, 2, 3, and 4, respectively.

VII. Results and Discussion

The classification performance of different recognition systems on the test data of 80 SCV classes is obtained for comparison. The hidden Markov model (HMM) based system uses a 5-state, left-to-right, discrete HMM trained for each class. The size of the codebook used is 256. The structure of the 80-class multilayer feedforward network consists of 240 nodes in the input layer, 120 nodes in the first hidden layer and 60 nodes in the second hidden layer. Table VI gives the performance of the HMM based system, the 80-class MLFFNN and the modular networks based on different grouping criteria. The performance of a modular network is obtained by using the decision rule on the outputs of the MLFFNNs in the network. The recognition performance is also obtained by using the decision rule on the combined evidence computed by adding the output for each class from the MLFFNNs in the three modular networks. The performance of all these systems is compared with that of the constraint satisfaction model (CSM). The performance of the CSM for Case_1 is as high as 65% indicating that the instance pool node with the largest output value gives the class of the input utterance correctly for 65% of the total number of test utterances. The performance of the CSM increases to about 82% for the Case_4 of the decision criterion.

The explanation for the superior performance of the CSM is the following: In the CSM, the outputs from the MLFFNNs of each grouping criterion are processed by the feedback subnetwork for that grouping. Similarities among classes are represented in 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 CSM not only uses the

TABLE VI
CLASSIFICATION PERFORMANCE OF THE CSM AND THE OTHER SCV RECOGNITION SYSTEMS ON TEST DATA OF 80 SCV CLASSES

SCV Recognition System	Decision Criterion			
	Case_1	Case_2	Case_3	Case_4
HMM based system	45.5	59.2	65.9	71.4
80-class MLFFNN	45.3	59.7	66.9	72.2
MOA modular network	29.2	50.2	59.0	65.3
POA modular network	35.1	56.9	69.5	76.6
Vowel modular network	30.1	47.5	58.8	63.6
Combined evidence based system	51.6	63.5	70.7	74.5
Constraint satisfaction model	65.6	75.0	80.2	82.6

knowledge about the similarities among 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 three modular networks is inferior to the CSM.

VIII. SUMMARY AND CONCLUSIONS

In this paper, we have proposed a new approach for developing a model for classification of utterances of 80 SCV classes. In this approach we proposed a constraint satisfaction model to represent the known constraints of the problem. Trained multilayer feedforward neural networks are used as nonlinear filters to extract features. A second level of training is used to derive the distribution of the feature vectors for each class. Since the constraint satisfaction model satisfies a set of even weak constraints in the best possible manner, the results are good in most of the cases.

Parametric representation that captures the crucial vocal tract transition information may help in improving the classification performance further. The parametric representation may also be a limiting factor for realizing speaker-independent classification of SCV utterances. Our studies demonstrate that the constraint satisfaction model can be used to enhance even the weak evidence available in the parametric representation of the input data. The models developed for the classification of the isolated utterances of SCVs may be useful for spotting SCV segments in continuous speech [16].

REFERENCES

[1] S. Greenberg, "Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation," *Speech Commun.*, vol. 29, no. 2–4, pp. 159–176, Nov. 1999.

[2] F. S. Cooper, "Acoustics in human communication: Evolving ideas about the nature of speech," *J. Acoust. Soc. Amer.*, vol. 68, pp. 18–21, 1980.

[3] C. C. Sekhar, "Neural network models for recognition of stop consonant-vowel (SCV) segments in continuous speech," Ph.D. dissertation, Indian Inst. Technol., Madras, 1996.

[4] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1999.

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

[6] P. P. Raghu and B. Yegnanarayana, "Supervised texture classification using a probabilistic neural network and constraint satisfaction model," *IEEE Trans. Neural Networks*, vol. 9, pp. 516–522, May 1998.

[7] D. E. Rumelhart, P. Smolensky, J. L. McClelland, and G. E. Hinton, "Schemata and sequential thought processes in PDP models," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, J. McClell and D. Rumelhart, Eds. Cambridge, MA: MIT Press, 1986, vol. 2, Psychological and Biological Models.

[8] R. P. Dixit, "Glottal gestures in Hindi plosives," *J. Phonetics*, vol. 17, pp. 213–237, 1989.

[9] J. Y. S. R. K. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based approach for detection of vowel onset points," in *Proc. Int. Conf. Advances in Pattern Recognition and Digital Techniques*, Dec. 1999, pp. 316–320.

[10] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[11] C. C. Sekhar and B. Yegnanarayana, "Classification of CV transitions in continuous speech using neural network models," in *Proc. Int. Symp. Speech, Image Processing, and Neural Networks*, 1994, pp. 97–100.

[12] A. Waibel, H. Sawai, and K. Shikano, "Modularity and scaling in large phonemic neural networks," *IEEE Trans, Acoust., Speech, Signal Processing*, vol. 37, pp. 1888–1898, Dec. 1989.

[13] P. P. Raghu and B. Yegnanarayana, "Segmentation of Gabor filtered textures using deterministic relaxation," *IEEE Trans. Image Processing*, vol. 5, pp. 1625–1636, Dec. 1996.

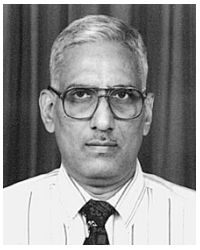
[14] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986, vol. 1, Foundations.

[15] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[16] C. C. Sekhar and B. Yegnanarayana, "Neural network models for spotting stop consonant-vowel (SCV) segments in continuous speech," in *Proc. Int. Conf. Neural Networks*, 1996, pp. 2003–2008.

C. Chandra Sekhar was born in India in 1961. He received the B.E. degree in electronics and communication engineering from Sri Venkateswara University, Tirupati, India, in 1984. He received the M.Tech. degree in electrical engineering and the Ph.D. degree in computer science and engineering from Indian Institute of Technology, Madras, Chennai, India, in 1986 and 1997, respectively.

He was a Lecturer from 1989 to 1997 and an Assistant Professor since 1997 in the Department of Computer Science and Engineering, Indian Institute of Technology. From June 2000 to May 2002, he was a JSPS Postdoctoral Fellow at Itakura Laboratory, Department of Information Electronics, Nagoya University, Nagoya, Japan. His current research interests include speech recognition, neural networks and support vector machines.



B. Yegnanarayana was born in India in 1944. He received the B.E., M.E., and the Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978, in the Department of Electrical Communication Engineering at the Indian Institute of Science. From 1966 to 1971, he was engaged in the development of environmental test facilities for the Acoustics Laboratory at the Indian Institute of Science. From 1978 to 1980, he was a Visiting Associate Professor of computer science at Carnegie Mellon University, Pittsburgh, PA. He was a Visiting Scientist at ISRO Satellite Center, Bangalore, from July to December 1980. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology. He was also the Chairman of the department from 1985 to 1989. He was a Visiting Professor at the Institute for Perception Studies, Eindhoven Technical University, Eindhoven, The Netherlands, from July 1994 to January 1995. Since 1972, he has been working on problems in the areas of speech and image signal processing. His current research interests are in signal processing, speech, vision, neural networks, and man–machine interfaces. He has published papers in reviewed journals in these areas and he is also the author of the book *Artificial Neural Networks* (New Delhi, India: Prentice-Hall of India, 1999).

Dr. Yegnanarayana is a Fellow of the Indian National Science Academy and the Indian National Academy of Engineering.