CONSTRAINT SATISFACTION MODEL FOR ENHANCEMENT OF EVIDENCE IN RECOGNITION OF CONSONANT-VOWEL UTTERANCES

Suryakanth V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
email: {svg,chandra,yegna}@cs.iitm.ernet.in

ABSTRACT

In this paper, we address the issues in recognition of a large number of subword units of speech with high confusability among several units. Evidence available from the classification models trained with a limited number of training examples may not be strong to correctly recognize the subword units. We present a constraint satisfaction neural network model that can be used to enhance the evidence for a particular unit with the supporting evidence available for a subset of units confusable with that unit. We demonstrate the enhancement of evidence by the proposed model in recognition of utterances of 145 consonant-vowel units.

1. INTRODUCTION

In this paper, we address the issues in recognition of a large number of consonant-vowel (CV) units of speech. Recognition of subword units of CV type with a high accuracy is important because they occur frequently in speech of many Indian languages. Combinations of 33 consonants and 12 vowels result in a set of about 400 CV units in a language. Because of similarities in their speech production mechanism, the confusability (partial acoustic similarity) among several CV units is high. Therefore, recognition of CV units is a more challenging task than recognition of E-set of English alphabet. Additionally, the number of examples available in a corpus may not be the same for all the units. There may be many units for which only a small number of examples are available. Classification models trained using a small number of examples for a unit may not give a strong evidence in recognition of utterances of that unit. Because of the variability due to context, speaker and background noise in continuous speech, the evidence for a unit may also be weak. Evidence for a CV unit available from the trained classification models may be enhanced by using the evidence for other CV units confusable with that unit. Constraints based on the confusability of CV units have been used to enhance the evidence in a constraint satisfaction neural network (CSNN) model developed for recognition of set of 80 stop consonant-vowel units [1]. In this paper, we extend this model for recognition of a set of 145 CV units belonging to all categories of consonants. We also demonstrate how the evidence for a CV unit is enhanced by the CSNN model.

The paper is organized as follows: In Section 2, we discuss the CSNN model for recognition of CV units. Section 3 demonstrates enhancement of evidence using the CSNN model. In Section 4, we present our study on recognition of isolated utterances of CV units.

2. CONSTRAINT SATISFACTION MODEL

The focus of the studies presented in this paper is on recognition of 145 CV units corresponding to combination of 29 consonants and five vowels that occur in many Indian languages. The list of 145 CV units is given in Table 1. Modular neural networks based approach is used to develop the models for large number of CV units. In this approach, subgroups of CV units are formed using a grouping criterion and a multilayer feedforward neural network (MLFFNN subnetwork) is trained separately for each subgroup. We consider grouping criteria based on the phonetic description of CV units for forming subgroups. A CV unit can be uniquely described using the phonetic features such as manner of articulation (MOA) of consonant, place of articulation (POA) of consonant, and the vowel sound in it. The phonetic description features for each of the 145 CV units are given in Table 1. Subgroups of CV units are formed such that the units in a subgroup have a particular phonetic feature common to them. Thus each phonetic description feature is used as a grouping criterion. For each grouping criterion, a particular CV unit is grouped with a different subset of CV units. Therefore, the output for a CV unit from subnetworks based on different grouping criteria may not be the same. The output from the subnetworks is considered as the evidence for CV units. Multiple grouping criteria lead to multiple sources of evidence

Table 1: List of 145 CV units.

MOA	POA	Vowel				
		/a/	/i/	/u/	/e/	/o/
Unvoiced Unaspirated (UVUA)	Velar	ka	ki	ku	ke	ko
	Palatal	ca	ci	cu	ce	co
	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
	Palatal	cha	chi	chu	che	cho
	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
	Palatal	ja	ji	ju	je	jo
	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
	Palatal	jha	jhi	jhu	jhe	jho
	Alveolar	Dha	Dhi	Dhu	Dhe	Dho
	Dental	dha	dhi	dhu	dhe	dho
	Bilabial	bha	bhi	bhu	bhe	bho
Nasals	Dental	na	ni	nu	ne	no
	Bilabial	ma	mi	mu	me	mo
Semivowels	Palatal	ya	yi	yu	ye	yo
	Alveolar	ra	ri	ru	re	ro
	Dental	la	li	lu	le	lo
	Bilabial	va	vi	vu	ve	vo
Fricatives	Velar	ha	hi	hu	he	ho
	Alveolar	sha	shi	shu	she	sho
	Dental	sa	si	su	se	so

for each unit. The evidence from multiple sources is combined to recognize a CV utterance.

Methods such as addition of evidence and majority voting have been used for combining the evidence from multiple sources [2]. We propose a CSNN model to combine the evidence available from outputs of subnetworks based on different grouping criteria. The model includes a feedback network for each grouping criterion to enhance the evidence. The connections in these networks represent the constraints based on the confusability among CV units. The evidence enhancing networks interact with one another through another feedback network, the instance pool, that combines the enhanced evidence [3].

The block diagram of the proposed system for recognition of CV units is shown in Fig. 1. It is seen that the system consists of two stages. The first stage consists of MLFFNN subnetworks trained for subgroups of CV units formed using different grouping criteria. The input to each subnetwork is a compressed acoustic feature vector obtained from a CV utterance using the method explained in Section 4. The outputs of each subnetwork are used to determine the evidence available for each CV unit in that subgroup. The evidence

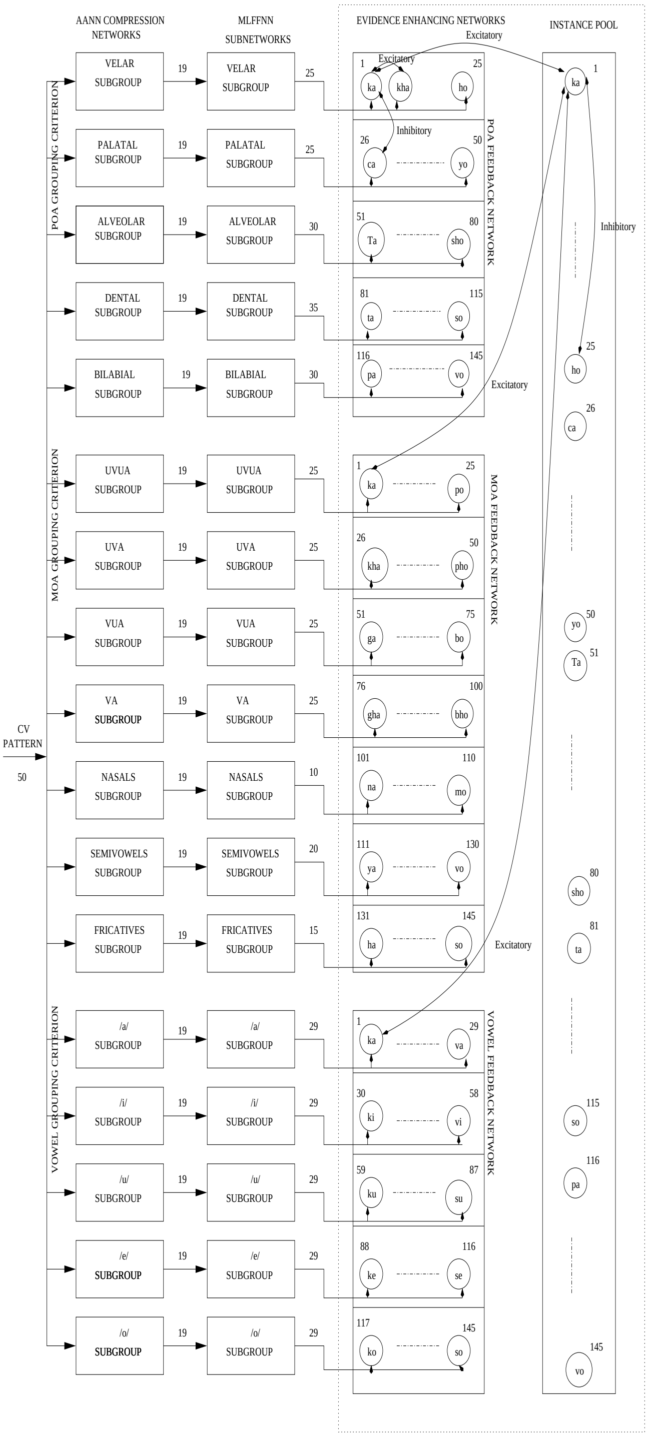

Figure 1: Block diagram of the proposed model for recognition of CV units. Output of a compression network is a 19-dimensional acoustic feature vector. The MLFFNN subnetworks have 100 nodes in the first hidden layer and 60 nodes in the second hidden layer. Number of outputs from an MLFFNN subnetwork correspond to the number of CV units in its subgroup. Numbers on nodes in the feedback networks represent the indices of CV units.

enhancing network for a grouping criterion consists of a node for each CV unit. The evidence for a CV unit determined from the outputs of subnetworks is given as input to its node in the evidence enhancing network. The node of a CV unit has bidirectional connections to the nodes of a subset of CV units that are highly confusable with that unit. For example, consider the CV unit /ka/ in the UVUA subgroup of MOA grouping criterion. The CV units /ki/, /ku/, /ke/, /ko/, /ca/, /Ta/, /ta/, /pa/ in the same subgroup are highly confusable with /ka/, because each of them differs with /ka/ only in POA or in vowel. Because the subnetwork for UVUA subgroup has been trained to discriminate among the units in that subgroup, evidence for these confusable units is used to increase the evidence for /ka/. The CV units /kha/, /ga/, /gha/, /ha/ are also highly confusable with /ka/ because each of them differ with /ka/ only in MOA. Because these units are in the subgroups different from that of /ka/, the evidence for these units is used to decrease the evidence for /ka/. Excitatory connections and inhibitory connections between /ka/ and each of its confusable units are used in the MOA feedback network to increase or decrease the evidence for /ka/ respectively. Similar connections are provided to the node of each unit in the evidence enhancing networks for each grouping criterion.

The instance pool feedback network consists of a node for each CV unit. The node for a CV unit has a bidirectional connection to the nodes of the same unit in the evidence enhancing networks so that the enhanced evidence from the three grouping criteria can be combined. The nodes of all the CV units in the instance pool have inhibitory connections from one to another so that the units compete with one another.

For a given test utterance, the CSNN model is initialized using the outputs of subnetworks. Deterministic synchronous update is performed until a stable state is reached [4]. When the model is in a stable state, the node in the instance pool with maximum evidence determines the class of the test utterance.

3. ENHANCEMENT OF EVIDENCE

Enhancement of evidence by the CSNN model is demonstrated for a test utterance of the CV unit /khi/. The outputs of the MLFFNN subnetworks are shown in Figs. 2(a), 2(b) and 2(c) for the POA, MOA, and vowel as grouping criterion respectively. The index of the unit /khi/ is 7. It is seen that the output for /khi/ is not the largest for any grouping criterion. Therefore, the conventional modular networks would have misclassified the test utterance. The evidence determined using the outputs of subnetworks is plotted in Figs. 2(d), 2(e) and 2(f) for each grouping criterion. It may be noted that evidence for /khi/ is significant, though it is not the highest. The outputs of the nodes of CV units in the instance pool when the CSNN model has reached to the stable state are shown in Fig. 2(h). It is seen that the node of the unit /khi/ has the largest output indicating that the enhanced and combined evidence is the highest for that CV unit. It is also seen that combined evidence is also high for confusable CV units such as /ghi/ (with the index of 9) and /thi/ (index 89). It is interesting to note that initially high evidence for the CV units such as /ki/ (index 6) and /ti/ (index 88) has been significantly reduced because each of these units do not have supporting evidence from its confusable CV units. This behavior of the CSNN model to enhance the evidence for a CV unit based on the supporting evidence from its confusable units is helpful in improving the CV recognition performance as presented in the next section.

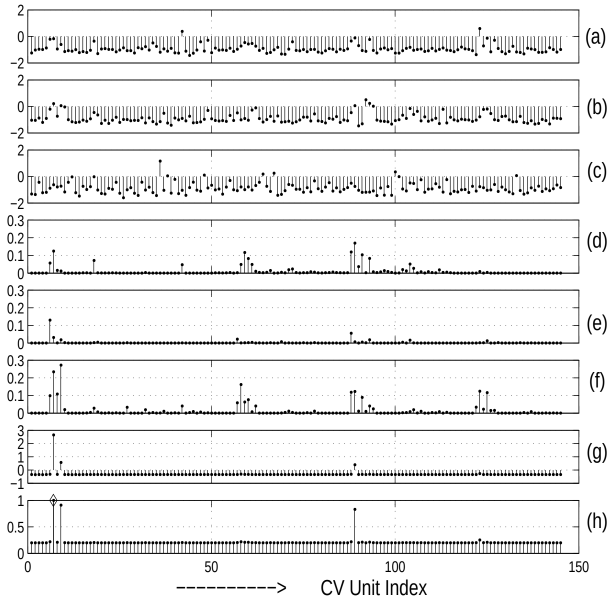

Figure 2: Evidence for CV units at different stages of the CV recognition system for an utterance of /khi/. (a)-(c) Outputs of MLFFNN subnetworks corresponding to POA, MOA and vowel grouping criterion, respectively. (d)-(f) Evidence for CV units based on the outputs of subnetworks. (g) Net input to the instance pool nodes when the model is in a stable state. (h) Outputs of the instance pool nodes when the model is in a stable state.

4. RECOGNITION OF CV UNITS

In this section, we present our study on recognition of isolated utterances of 145 CV units that occur in an Indian language. Speech data is collected for 12 utterances of each CV unit from each of three male speakers. For each utterance a 50-dimensional acoustic feature vector is extracted using the method given in [5]. Nonlinear principal component analysis using an autoassociative neural network (AANN) model for each subgroup of CV units is carried out to obtain a 19-dimensional feature vector for each utterance [5].

The total data set is divided into three parts as follows. Training data set 1 (TDS-1) consists of five utterances of each CV unit from each speaker. Training data set 2 (TDS-2) consists of five utterances of each unit from each speaker. Test data set consists of two utterances of each CV unit from each speaker. Utterances in the TDS-1 are used to train the MLFFNN subnetworks for different subgroups of CV units. Then the utterances of a CV unit in TDS-2 are given to subnetworks of subgroups to which that CV unit belongs to. A mean output vector is obtained from the outputs of a subnetwork for each of these utterances. The mean vector is associated with the node of the CV unit in the feedback network of the corresponding grouping criterion. During the recognition, outputs of the subnetworks for a test utterances are obtained. The Euclidean distance of the output vector of a subnetwork with the mean vector of each CV unit in the subgroup is used to compute the initial evidence for each of the CV units and for that grouping criterion [1]. The initial evidence thus obtained for each CV unit is enhanced by the CSNN model as explained in Section 2.

Table 2 gives the average classification performance in recognition of the utterances in the test data set. The outputs of the nodes in the instance pool are used to determine the class of a test utterance. The $k-$best performance is obtained based on the outputs of instance pool nodes after every iteration of the relaxation process of the CSNN model. It is observed that the performance does not vary significantly after five iterations. It is important to note that the CSNN model gives a $1-$best performance of about 60% in recognition of a large number of CV units with high confusability among several units. For comparison, the performance of a Hidden Markov Model (HMM) based system is obtained. A 5-state HMM with a single Gaussian per state is trained for each CV unit using the utterances in TDS-1. We also consider a CV recognition system in which the outputs for a CV unit from subnetworks based on different grouping criteria are added. The test pattern is then assigned to the CV class with the largest total output. The performance of this system, called the multiple modular network (MMN) system, is given along with the performance of the HMM based system and the CSNN model based system in Table 3. It is seen that the CSNN model gives the best performance. The improved performance is mainly due to the enhancement of evidence by the feedback networks in the CSNN model.

Table 2: CV recognition performance of the system using CSNN model. The $k-$best performance based on the outputs of instance pool after every iterations is given.

Iteration	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
1	52.29	56.78	60.34	63.10	64.36
2	59.77	67.47	71.83	76.43	78.73
3	59.88	68.96	75.86	79.65	83.10
4	60.11	69.65	75.86	80.80	83.79
5	60.45	69.08	76.09	80.00	83.44

Table 3: Comparison of the $k-$best classification performance for different CV recognition systems.

System	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
HMM	57.82	62.64	66.44	68.51	70.57
MMN	44.71	56.66	63.44	67.81	71.14
CSNN	60.45	69.08	76.09	80.00	83.44

5. SUMMARY AND CONCLUSIONS

In this paper, we presented a constraint satisfaction model for recognition of a large number of subword units of CV type. It is demonstrated that weak evidence due to the small number of training examples and due to the confusability among several units can be enhanced significantly by the CSNN model. The enhancement of evidence may be useful in recognition of subword units in continuous speech with an improved accuracy. Other methods, such as based on support vector machines, can improve the performance at the subnetworks level significantly, which in turn may improve the performance of the overall CV recognition system [6].

6. REFERENCES

[1] C. Chandra Sekhar and B. Yegnanarayana, "A constraint satisfaction model for recognition of stop consonant-vowel (SCV) utterances," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 7, pp. 472–480, Oct. 2002.

[2] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall International, New Jersey, 1999.

[3] D. E Rumelhart, P. Smolensky, J. L. McClelland, and G. E. Hinton, *Schemata and sequential thought processes in PDP models*, in Parallel and Distributed Processing. vol. 2, J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, Eds., MIT Press, 1986.

[4] P. P. Raghu and B. Yegnanarayana, "Segmentation of Gabor-filtered textures using deterministic relaxation," *IEEE Trans. Image Processing*, vol. 5, no. 12, pp. 1625–1636, Dec. 1996.

[5] S. V. Gangashetty, A. Nayeemulla Khan, S. R. Mahadeva Prasanna, and B. Yegnanarayana, "Neural network models for preprocessing and discriminating utterances of consonant-vowel units," in *Proc. Int. Joint Conf. Neural Networks*, 2002, vol. 1(3), pp. 613–618.

[6] C. Chandra Sekhar, K. Takeda, and F. Itakura, "Close-class-set discrimination method for large-class-set pattern recognition using support vector machines," in *Proc. Int. Joint Conf. Neural Networks*, 2002, vol. 1(3), pp. 577–582.