Speech Communication 11 (1992) 273–282
North-Holland

Integrated phoneme and function word architecture of Hidden Control Neural Networks for Continuous Speech Recognition

Bojan Petek[1], Alex H. Waibel and Joseph M. Tebelskis
School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213-3890, USA

Received 26 September 1991
Revised 23 January 1992

Abstract. We present a context-dependent, phoneme and function word based, Hidden Control Neural Network (HCNN-CDF) architecture for continuous speech recognition. The system can be seen as a large vocabulary extension of the word-based HCNN system proposed by Levin in 1990. Initially, we analysed context-*in*dependent HCNN modeling principle in the framework of the Linked Predictive Neural Network (LPNN) speech recognition system and found that it results in a 6% increase of the word recognition accuracy at perplexity 402. Significant savings compared to the LPNN in the resource requirements and computational load for the HCNN implementation can be achieved. In speaker-dependent recognition experiments with perplexity 111, the current versions of the LPNN and HCNN-CDF systems achieve 60% and 75% word recognition accuracies, respectively.

Zusammenfassung. Wir stellen im folgenden eine kontextabhängige auf Phonemen und Funktionswörtern basierende Hidden Control Neural Network Architektur (HCNN-CDF) für die Erkennung von kontinuierlicher Sprache vor. Das System ist eine Erweiterung des wortbasierten HCNN Systems von Levin in 1990 auf ein großes Vokabular. Wir haben zuerst das Prinzip der kontextunabhängigen HCNN-Modellierung im Rahmen des Linked Predictive Neural Network (LPNN) Spracherkennungssystems untersucht und eine Verbesserung der Worterkennungsrate um 6% bei einer Perplexität von 402 festgestellt. Für die HCNN-Implementation konnte eine bedeutende Parameterreduktion und Einsparung von Rechenzeit gegenüber LPNN erreicht werden. Bei sprecherabhängigen Erkennungsexperimenten mit der Perplexität 111 erreichten die aktuellen Versionen des LPNN und des HCNN-CDF Systems Worterkennungsraten von 60% bzw. 75%.

Résumé. Nous présentons une architecture d'Hidden Control Neural Network (HCNN-CDF) dépendant du contexte pour la reconnaissance de la parole continue, basée sur les phonèmes et les mots fonctionnels. Le système peut être considéré comme une large extension du vocabulaire du système HCNN basé sur les mots, proposé par Levin. Initialement, nous avons analysé les principes de modélisation de HCNN sous une forme indépendante du contexte, dans le cadre du système de reconnaissance de la parole Linked Predictive Neural Networks (LPNN) et avons trouvé qu'il aboutit à une augmentation de 6% dans la précision de reconnaissance de la parole à un degré de perplexité 402. Comparé à LPNN, nous avons pu obtenir des réductions significatives dans les exigences de ressources et les charges computationnelles grâce à notre implémentation HCNN. Dans des expériences de reconnaissance dépendant du locuteur, avec un degré de perplexité 111, les versions actuelles des systèmes LPNN et HCNN-CDF obtiennent respectivement une précision de reconnaissance de mots de 60% et 75%.

Keywords. Automatic Speech Recognition; Hidden Control Neural Network; large vocabulary recognition; context-dependent modeling; function-word modeling.

1. Introduction

Early connectionist approaches to Automatic Speech Recognition (ASR) predominantly used connectionist models as classifiers of either full words (e.g. digits), or subword units (e.g. syllables, phonemes) (Waibel and Lee, 1990). In a classification based approach, consecutive frames of speech input vectors are clamped to the network inputs and are mapped into a binary pattern representing a finite set of recognition classes. This approach treats the adjacent speech feature vectors as independent variables and does not explicitly model the causality between them.

[1] The author is now with the Department of Computer Science, Faculty of Electrical Engineering and Computer Science, University of Ljubljana, Slovenia.

The application of artificial neural networks (ANN) in the framework of Hidden Markov Models (HMM) was analysed in (Bourlard and Wellekens, 1990; Bourlard, 1991; Franzini et al., 1991). This work showed that ANNs could be effectively used to estimate emission probabilities for the HMMs. This hybrid approach has the advantages of improved discrimination and the ability to incorporate multiple sources of knowledge without assumptions of distributions or statistical independence (Bourlard, 1991; Morgan et al., 1991; Franzini et al., 1991).

Recently, a non-linear prediction approach has been proposed for ASR, and used in several systems, e.g. "Neural Prediction Model", NPM, by Iso and Watanabe (1990, 1991), "Hidden Control Neural Network", HCNN, by Levin (1990) and "Linked Predictive Neural Networks", LPNN, by Tebelskis and Waibel (Tebelskis and Waibel, 1990; Tebelskis et al., 1991). In these systems, the connectionist networks, used as acoustic models of speech, are trained to learn the temporal correlations between adjacent speech patterns, thus presenting a dynamical systems approach to ASR (Tishby, 1990).

Initial evaluations of these models were carried out on small vocabulary recognition tasks, such as speaker-independent digit recognition (Iso and Watanabe, 1990; Levin, 1990, 1991), yielding high recognition performances, and large vocabulary continuous speech recognition extensions (Iso and Watanabe, 1991; Tebelskis et al., 1991). The work of Tebelskis uncovered a discriminatory problem of the predictive approach on the English database (Tebelskis et al., 1991).

This paper describes a large vocabulary, context-dependent HCNN system. The system can be seen as a large vocabulary extension of the word-level HCNN system of Levin (1990). After reviewing the basic concepts of the predictive connectionist approach to ASR, we present results of a preliminary comparison between the LPNN and HCNN, using context-*in*dependent models. Section 4 presents a context-dependent HCNN modeling principle for large vocabulary ASR and function word integration to the system. The evaluation of the HCNN-CDF system's performance on the CMU Conference Registration Database and its comparison to the baseline LPNN (Tebelskis et al., 1991) are presented in Section 5. We conclude the paper with a discussion of the results and promising future research directions.

2. Predictive connectionist approach to ASR

The basic idea of the NPM, HCNN and LPNN predictive modeling is shown in Figure 1.

An n-frame window, $n = n_M + n_P$ (Figure 1), of speech vectors is input into a multi-layer feed-forward net. The network is trained to approximate as close as possible the speech frame at time t. The prediction error, defined as the distance between the predicted and the actual speech frame, is used as an error criterion for the backpropagation training. During training a pool of canonical subword models (e.g. phonemes, demi-syllables) learns to become specialized to predict corresponding portions of an utterance. This means that they develop the lowest prediction errors in the regions on which they have been trained, and the higher errors outside of those regions. A word is typically represented as a sequence of predictors that best predict the observed speech. The Viterbi algorithm is used for finding the best sequence of predictors over time in order to match the observed speech signal.

2.1. LPNN and HCNN modeling: a comparison

When comparing the LPNN and HCNN modeling principles, the following differences can be observed.

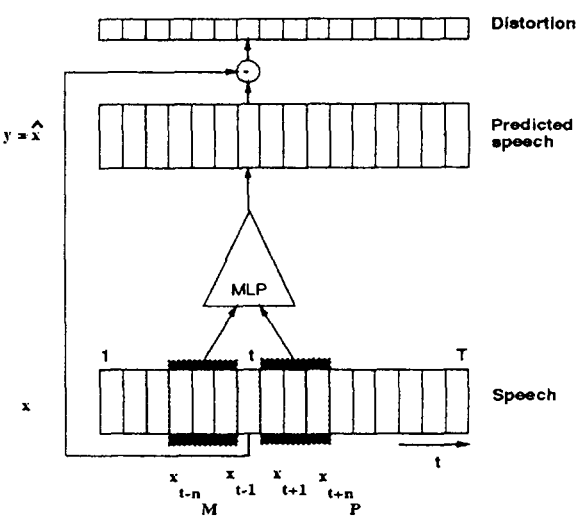

Fig. 1. Predictive connectionist modeling principle for ASR.

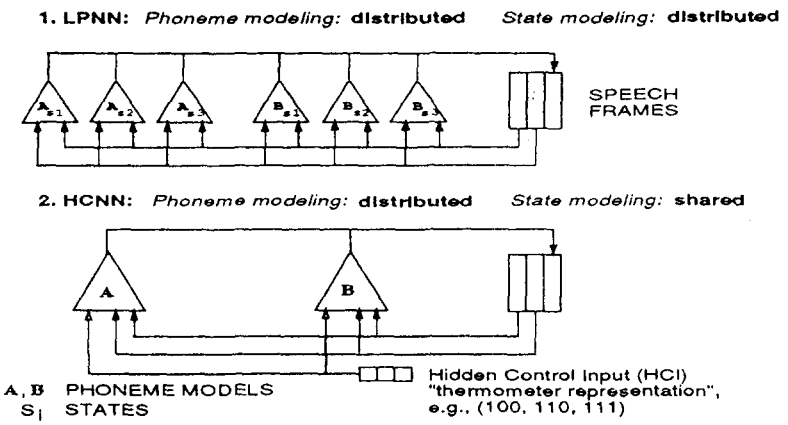

Fig. 2. LPNN and HCNN modeling principles.

The state sequence of a phoneme is modeled by a single Hidden Control Neural Network while the LPNN system uses a sequence of distinct neural networks (Figure 2).

This gives the HCNN system the advantage of having fewer parameters to train for a given amount of training data while having more training data per neural network. This should be advantageous for obtaining a better model.

Since models for different states within a phoneme or function word are *shared* (Figure 2), the HCNN system is less computationally demanding than the LPNN. The hidden unit input activations (e.g. the solid lines in Figure 3) for a given speech frame from the input layer to the first hidden layer need not be recomputed for more than one state; i.e. only the hidden control inputs change (e.g. the dotted lines in Figure 3).

3. Baseline experiments

This section briefly summarizes the results obtained from our preliminary experiments. First, the context-*in*dependent HCNN and LPNN modeling principles were compared. Following this, Section 3.1 describes the HCNN function word modeling experiments.

In the preliminary set of experiments, unrelated to the issue of function word modeling, we compared the performance of the context-*independent* LPNN and HCNN modeling *without* function words under the same testing conditions. The word recognition accuracy was tested every 20 epochs of training and was analysed twice, starting from two different initial conditions.

Both systems consisted of 40 phoneme models. While keeping the size of the prototypical neural

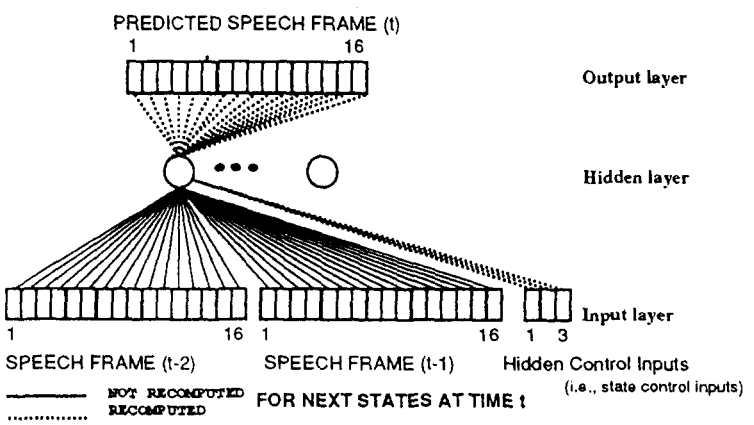

Fig. 3. Reduction of computational load by the HCNN modeling principle.

network about the same, the total number of free parameters in the system was reduced from 80 960 (LPNN modeling) to 42 080 when HCNN modeling was used. In our comparison, each phoneme model had 32 input units, 10 hidden and 16 output units. In the HCNN case, two more input units were used as the hidden control inputs. All networks were fully connected.

In these experiments, the "mblw" speaker of the CMU's Conference Registration Database and a 2-state phoneme topology were used for performance evaluation.

A comparable performance (96% word recognition accuracy) of both modeling principles was obtained on the same task at perplexity 7. When testing at perplexity 402, the HCNN modeling showed 6% better word recognition accuracy. This absolute difference in word recognition accuracy was measured as the smallest difference at any particular number of epochs. At that point on the learning curve, the LPNN and HCN systems achieved 14% and 20% word accuracies, respectively.

3.1. *HCNN function word modeling*

Function word modeling for continuous speech recognition was shown to be an important issue in ASR (Lee, 1988). Function words have strong coarticulation effects, are very frequent in continuous speech and are often poorly articulated by the speaker. Poor modeling of these words can considerably degrade the overall word accuracy of an ASR system.

Initially, our system consisted of 40 canonical phoneme models. Preliminary analysis of the recognition results showed that most misrecognitions occur on short words and function words. We thus decided to add additional resources to the system to achieve better modeling and improved recognition accuracy for these words.

We selected the function word "the" for the experiments. In order to decide on which modeling principle models function words better, two sets of experiments were conducted. In the first set of experiments we modeled a function word with a sequence of phoneme models (e.g. DH AX), and in the second set with the word-level function word model.

A training algorithm that clusters function words was developed, as described in (Petek et al., 1991).

In order to concentrate on issues of function word modeling, we initially tested on excerpted words, where word boundaries are given.

When tested on 105 function words, the experiments showed a decrease of 23% in the number of substitution errors, i.e. from 40 to 16 errors, when the word-level model was used.

4. HCNN-CDF system for large vocabulary CSR

This section describes a context-dependent Hidden Control Neural Network Architecture (HCNN-CDF) for large vocabulary continuous speech recognition. In order to design a large vocabulary ASR system, subword units, e.g. phonemes or syllables, have to be used. Subword modeling is then improved by applying explicit context-*dependent* modeling to the Hidden Control Neural Network model (Figure 4) (Petek et al., 1991).

The HCNN-CDF system explicitly models function words on the word level by using a *single* Hidden Control Neural Network. For example, a function word "the" is modeled with a single "THE" function word model, rather than a sequence of phoneme models. As shown by our preliminary experimental results, word-level modeling of function words improves their recognition rate.

Another significant extension compared to the HCNN system of Levin (1990) is the use of *explicit* context-dependent modeling, as described in Section 4.1.

The HCNN-CDF system consists of N context-dependent phone models and M context-dependent function word models. The $N=40$ phone models used in this work are given in Table 1.

Word models are defined by concatenating the phone models, as defined in the system's dictionary. Given that any word can be described by a concatenation of canonical phone models, large vocabulary speech recognition can be achieved. Each phone or function-word model has a transition diagram as shown in Figure 5.

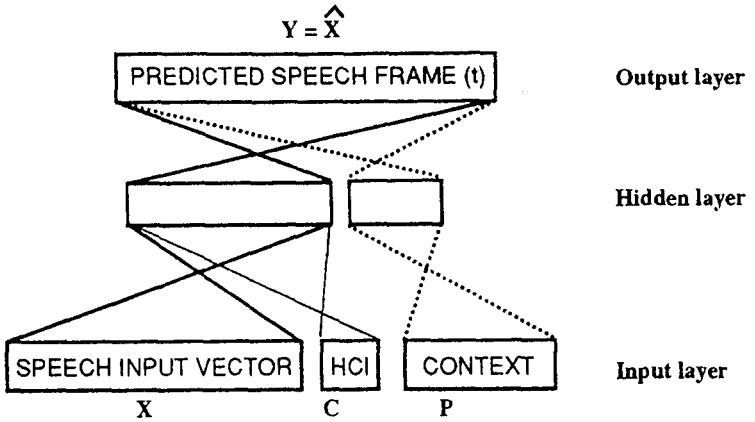

Fig. 4. Context-dependent HCNN model.

Table 1
Canonical phone models of the HCNN system

Phoneme	Example	Phoneme	Example	Phoneme	Example	Phoneme	Example
SIL	(silence)	IH	bit	F	fluff	Q	(garbage)
AA	father	IY	beet	G	gig	R	roar
AE	bat	OW	coat	HH	how	S	sass
AH	but	UH	book	JH	fudge	SH	ship
AO	hot	UW	boot	K	key	T	time
AW	cow	B	bib	L	lull	TH	thin
AY	bite	CH	chip	M	maim	V	valve
EH	bet	D	dime	N	anna	W	one
ER	bird	DH	the	NG	bang	Y	you
EY	bait	DX	at	P	ship	Z	zoos

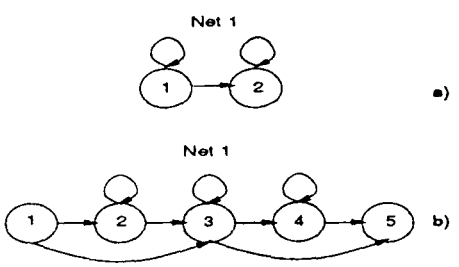

Fig. 5. State transition diagrams of the HCNN model.

4.1. Context-dependent HCNN modeling

One of the strengths of connectionist models is their ability to successfully combine information from several input domains. This capability has also been used in our context-dependent HCNN modeling principle for large vocabulary ASR.

In the framework of classification networks it has already been demonstrated that the phonetic context has to be taken into account for very high accuracy in phoneme recognition (Watrous, 1989).

The input layer of the context-dependent HCNN model consists of three groups of units, i.e. speech inputs with N_x units, hidden control inputs with N_c units and N_p contextual units.

$$N_{\text{input}} = N_x + N_c + N_p. \tag{1}$$

The mapping function $\mathcal{F}$ between the observable speech input $\boldsymbol{x}$ and the output $\boldsymbol{y}$ is defined by

$$\boldsymbol{y} = \hat{\boldsymbol{x}}_t = \mathcal{F}_{\omega}(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{p}) \triangleq \mathcal{F}_{\omega,c,p}(\boldsymbol{x}), \tag{2}$$

where ω represents a set of fixed parameters (i.e. weights and biases of the network), and $(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{p})$ the concatenation of the three inputs.

Since the coverage of left *and* right contexts of our database was rather small, we decided to make

the predictive models only *right*-context dependent. The binary pattern of contextual information represents a coding of the linguistic features of the right-hand phoneme, as described in (McClelland, 1986).

We used 10 bits for contextual input representation, coding each phoneme along four dimensions. The first dimension (three bits) was used to divide the phonemes into interrupted consonants (stops and nasals), continuous consonants (fricatives, liquids and semivowels) and vowels. The second dimension (two bits) was used to subdivide these classes. The third dimension (three bits) classified the phonemes by places of articulation (front, middle, back). Finally, the fourth dimension (two bits) divided the consonants into voiced and unvoiced, and vowels into long and short.

Separate first hidden layers for speech/state and context information were used. After the training phase, the activations in the contextual first hidden layer can be cached, thus yielding additional computational savings. Given the context, known hidden unit activations are clamped to the *first hidden layer* instead of the contextual input vector being clamped to the input layer.

As can be seen in Figure 4, the mapping function of the HCNN model is now state as well as context dependent. The hidden control input enables the MLP model to handle large-scale temporal variability (i.e. state transitions in a phoneme/word). The contextual inputs permit more context-specific predictions, thus potentially enhancing the discrimination among acoustic models.

4.2. Context-dependent HCNN as a statistical model

From a statistical point of view, the HCNN model can be described by an equivalent vector source with assumed white Gaussian noise $\boldsymbol{n}_t$ having zero mean and covariance matrix Σ (Levin, 1990).

Following this assumption, a context-dependent HCNN model can be described as

$$\hat{\boldsymbol{x}}_t = \mathscr{F}_{\omega, c_t, p}(\boldsymbol{x}_{t-1}) + \boldsymbol{n}_t, \quad \boldsymbol{n}_t \sim \mathrm{N}(0, \Sigma). \tag{3}$$

Assuming this interpretation, the conditional likelihood of observation $\boldsymbol{x}_t$ is given by

$$\mathrm{P}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \omega, c_t, p) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \times \exp(-\tfrac{1}{2}(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p))' \Sigma^{-1} (\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p))). \tag{4}$$

By assuming the equiprobability of control codes c_i, the joint likelihood of the data and the control can be written as

$$\mathrm{P}(\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{c}_1^{\mathrm{T}} \mid \omega, p) = \frac{1}{(2\pi)^{dT/2}|\Sigma|^{T/2}} \times \exp\left(-\frac{1}{2} \sum_{t=1}^{T} (\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p))' \times \Sigma^{-1} (\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p)) \right), \tag{5}$$

where $\boldsymbol{x}_1^{\mathrm{T}}$ denotes the observed sequence $\{x_1, \ldots, x_T\}$. This assumption considers a special case of the Markov process (Levin, 1990) and can be easily extended for the general case.

The final goal is to maximize $\mathrm{P}(\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{c}_1^{\mathrm{T}} \mid \omega, p)$, given by (5). In order to see the relationship between the minimization of the prediction error[1] and the desired maximization of the joint likelihood, the logarithm of (5) and simplification yields

$$\ln \mathrm{P}(\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{c}_1^{\mathrm{T}} \mid \omega, p) = \sum_{t=1}^{T} (\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p))' \Sigma^{-1} (\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p)) + \ln|\Sigma|. \tag{6}$$

If the non-diagonal components of the covariance matrix are negligible, (6) reduces to

$$\ln \mathrm{P}(\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{c}_1^{\mathrm{T}} \mid \omega, p) = \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p))^2}{\sigma_i^2} + \sum_{i=1}^{d} \ln \sigma_i^2. \tag{7}$$

[1] Performed by the Back-Propagation (BP) training algorithm.

The assumption of the unity variance, i.e. $\sigma_i^2 = 1$, and further simplification yields[2]

$$\ln P(\boldsymbol{x}_1^T, \boldsymbol{c}_1^T \mid \omega, p) = \sum_{t=1}^{T} \|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t(c_t, p)\|^2. \quad (8)$$

The comparison of (8) with the equation of the BP training algorithm

$$d_k(n) = \min_{\{a_k(t)\}} \sum_{t=1}^{T_n} \|\boldsymbol{x}_t(n) - \hat{\boldsymbol{x}}_t(n, k)\|^2 \quad (9)$$

yields

$$\ln P(\boldsymbol{x}_1^T, \boldsymbol{c}_1^T \mid \omega, p) \sim d_k(n). \quad (10)$$

Therefore, the minimization of the prediction error $d_k(n)$ performed by the BP algorithm is equivalent to the maximization of the joint likelihood $P(\boldsymbol{x}_1^T, \boldsymbol{c}_1^T \mid \omega, p)$,

$$\min_{\Omega, C^T} d(\omega, \boldsymbol{c}_1^T, p) \quad \Leftrightarrow \quad \max_{\Omega, C^T} P(\boldsymbol{x}_1^T, \boldsymbol{c}_1^T \mid \omega, p), \quad (11)$$

which is the desired goal of the training procedure.

4.3. *Integrating function word models into the system*

Function word models have been added to the system by introducing additional context-dependent HCNN *word* models (previously, every word was modeled by a sequence of one or more general purpose phoneme models). The hidden control input now codes for the states within a word, as in the word-level HCNN system of Levin (1990).

The context input code of the word model was chosen to be the same as the first phoneme code in the standard phonetic spelling of the function word.

An important issue when adding additional resources to the predictive system like this is the fact that the *existing resources* (i.e. phoneme models) *need not be retrained.* If the system was classification-based, however, all the models would have to be retrained to insure the proper learning of the discrimination hyperplanes for the increased number of decision classes.

[2] This assumption need not be made if we implement variance modeling.

5. Experimental results

The described integrated phoneme and function word HCNN system was trained on the Conference Registration Database recorded at CMU which consists of 204 English sentences (data from speaker "mjmt" was used in the experiments). For testing, we used the first three dialogs of a separate recording as a test set (the same as reported in (Tebelskis et al., 1991)).

The prototypical network consisted of 64 speech inputs (2 frames of speech with corresponding delta frames), 5 hidden control inputs, 10 context inputs, 30 speech/state units in the first hidden layer, 5 context units in the first hidden layer, and 16 predicted speech output units (1 frame). The networks were fully connected. Speech input vector consisted of 16-dimensional mel-scale filterbank coefficients, corresponding to time frames $t-1$ and $t-2$, and first order difference vectors for these two frames, computed using a 40 msec delta.

In contrast to the LPNN system (Tebelskis et al., 1991), the HCNN phoneme model consisted of 5 states, implemented by a *single* network, as shown in Figure 4. All canonical phoneme (and function word) models had the same topology, except the silence model, which had 3 states.

Each of the phoneme or function word models was implemented by two separate HCNN networks, called *alternates*. This concept improved the modeling accuracy by allocating more than one predictive model per phoneme or word (Tebelskis and Waibel, 1990). During training, only the winning alternate model was reinforced by backpropagating error while competing alternate remained unchanged.

5.1. *HCNN-CDF system evaluation*

We tested the context-dependent HCNN system at task perplexity 111. The results are summarized in Table 2. The table includes the Linked Predictive Neural Network (LPNN) system's performance, as given in (Tebelskis et al., 1991). The HCNN-CD denotes the system with 40 canonical phoneme

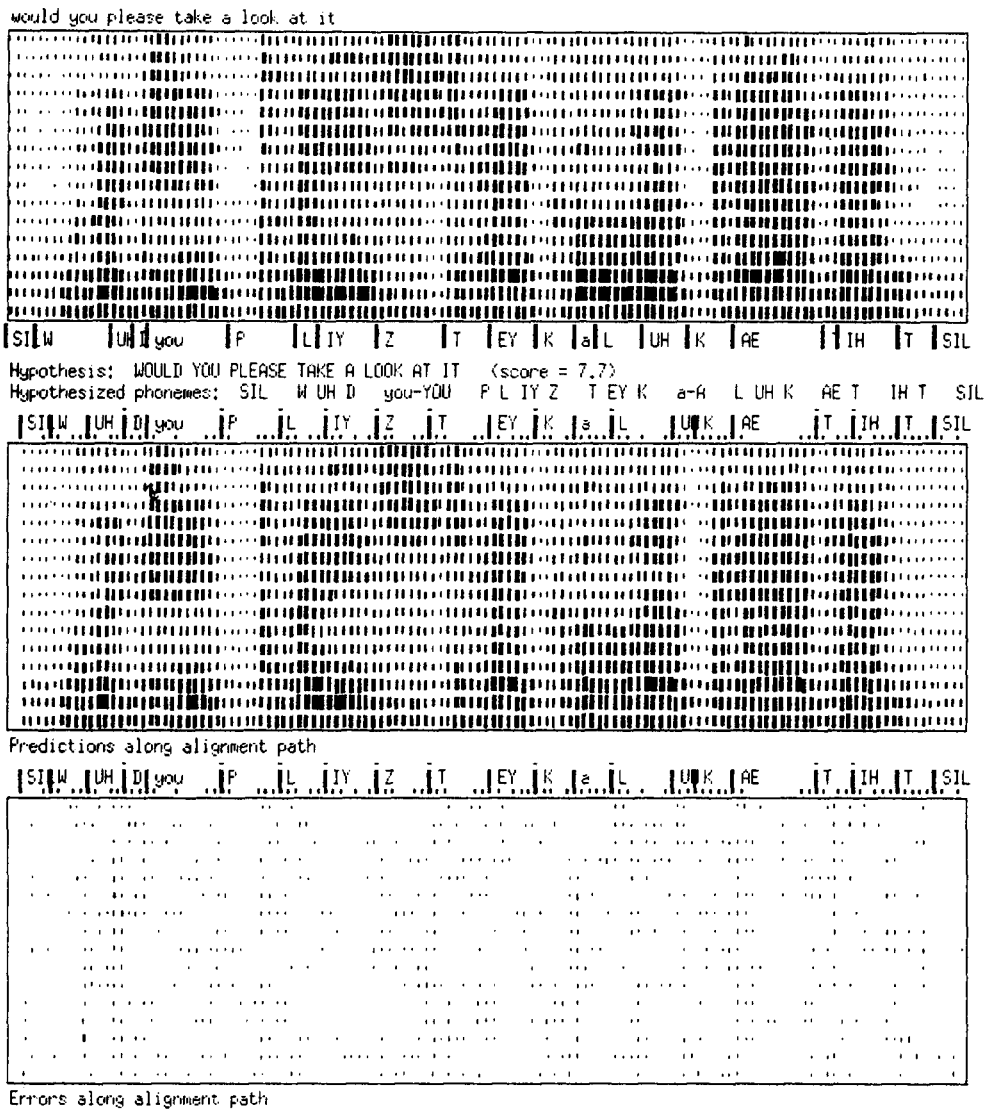

Fig. 6. Typical example of the HCNN-CDF continuous speech recognition. Upper trace: Actual sentence. x-axis: time (frame sampling rate is 10 ms); y-axis: frequency (normalized melscale filterbank coefficients). Middle trace: Predictions and segmentation generated by the HCNN system. Major hash marks: phoneme and function word boundaries; Minor hash marks: HCNN-input "thermometer" input change; Bars above major segmentation marks specify the alternate model index. Lower trace: Prediction errors made along the alignment path.

models (i.e. no function word models), and the HCNN-CDF system additionally models three function words (a, the, you).

We did an additional analysis of the phoneme confusion matrix of the HCNN-CD system on the training data which showed that there is still 20% phoneme-level confusions after the training has been completed. The most confusions occurred on AH, AE, EH, UH, D, K and N phonemes.

A typical example of the recognized sentence by the HCNN-CDF system is given in Figure 6.

6. Discussion and conclusions

We presented a context-dependent, phoneme and function word based, Hidden Control Neural

Table 2
Speaker-dependent word recognition accuracies of LPNN and HCNN systems

System	Speaker A (mjmt), perplexity 111		
	LPNN	HCNN-CD	HCNN-CDF
Substitutions	28%	20%	18%
Deletions	8%	6%	3%
Insertions	4%	2%	4%
Word accuracy	60%	72%	75%

Network architecture for large vocabulary continuous speech recognition and evaluated its performance on the Conference Registration Database.

Speaker-dependent performance evaluation of the LPNN and HCNN systems on the same task with perplexity 111 showed that the systems achieve 60% and 75% word recognition accuracy, respectively.

Additionally, we have shown that the baseline performance of the HCNN system increased from 72% to 75% when modeling three most frequently misrecognised function words.

One of the remaining drawbacks of the system is still the insufficient discrimination power among the predictive models of the system. The main reasons for this seem to be the following:

The lack of "negative" examples of prediction. As the models get trained only on "positive" examples along the optimal alignment path, no explicit mechanism enforces them to become discriminant among each other. The consequence of training on only "positive" examples is that the predictors have *undefined regions*, in which no training examples were given to define a controlled, i.e. non-confusable, response when compared to the other models of the system.

The problem of identity mappings. The current implementation of the system predicts successive speech frames in an utterance. Some subword speech units, e.g. vowels, exhibit very small change between successive speech frames, thus training the model to make very similar (i.e. identity) mappings. The response of the model trained in this manner increases the confusability among the predictive models of the system, since it predicts fairly well also the data of the other models.

One possibility to address this problem would be a design of an additional *corrective training* algorithm. This algorithm should enable the training on examples outside the optimal alignment path, thus enforcing discriminant behaviour of the models outside the regions of "positive" training examples. The design of this procedure for the predictive models of the system is at present still an open and important research issue.

Acknowledgment

The authors gratefully acknowledge the support of DARPA, the National Science Foundation, ATR Interpreting Telephony Research Laboratories, and NEC Corporation. B. Petek also acknowledges support from the Ministry of Science of Slovenia.

References

H. Bourlard (1991), "Neural nets and hidden Markov models: Review and generalizations", *Proc. Eurospeech-91*, Vol. 2, pp. 363–369.

H. Bourlard and C.J. Wellekens (1990), "Links between Markov models and multilayer perceptions", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 12, No. 12, pp. 1167–1178.

D.S. Broomhead and D. Lowe (1988), "Multivariable functional interpolation and adaptive networks", *Complex Systems 2*, pp. 321–355.

M.A. Franzini, A.H. Waibel and K.F. Lee (1991), "Recent work in continuous speech recognition using the connectionist Viterbi training procedure", *Proc. Eurospeech-91*, Vol. 3, pp. 1213–1216.

K. Iso and T. Watanabe (1990), "Speaker-independent word recognition using a neural prediction model", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 441–444.

K. Iso and T. Watanabe (1991), "Large vocabulary speech recognition using neural prediction model", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 57–60.

K.F. Lee (1988), Large vocabulary speaker independent continuous speech recognition: The SPHINX system, PhD dissertation, Computer Science Department, Carnegie-Mellon University.

E. Levin (1990), "Word recognition using hidden control neural architecture", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 433–436.

E. Levin (1991), "Modeling time varying systems using a hidden control neural network architecture", in *Advances in Neural Information Processing Systems 3*, pp. 147–154.

J.L. McClelland, D.E. Rumelhardt and the PDP research group (1986), *Parallel Distributed Processing* (MIT Press, Cambridge, MA), Vol. 2, Chapter 18, pp. 217–268.

N. Morgan, H. Bourlard, C. Wooters, P. Kohn and M. Cohen (1991), "Phonetic context in hybrid HMM/MLP continuous speech recognition", *Proc. Eurospeech-91*, Vol. 1, pp. 109–112.

M. Niranjan and F. Fallside (1988), "Neural networks and radial basis functions in classifying static speech patterns", Technical report CUED/F-INFENG/TR22, University Engineering Department, Cambridge, UK.

B. Petek, A.H. Waibel and J.M. Tebelskis (1991), "Integrated phoneme-function word architecture of hidden control neural networks for continuous speech recognition", *Proc. Eurospeech-91*, Vol. 3, pp. 1407–1410.

T. Poggio and F. Girosi (1990), "Networks for approximation and learning", *Proc. IEEE*, pp. 1481–1497.

S. Renals and R. Rohwer (1989), "Phoneme classification experiments using radial basis functions", *Proc. IJCNN*, pp. 461–467.

J. Tebelskis and A. Waibel (1990), "Large vocabulary recognition using linked predictive neural networks", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 437–440.

J. Tebelskis, A. Waibel, B. Petek and O. Schmidbauer (1991), "Continuous speech recognition using linked predictive neural networks", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 61–64.

N. Tishby (1990), "A dynamical systems approach to speech processing", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 365–368.

A. Waibel and K.F. Lee, eds. (1990), *Readings in Speech Recognition* (Morgan Kaufman, Los Altos, CA).

R. Watrous (1989), "Context-modulated discrimination of similar vowels using second-order connectionist networks", Technical report CRG-TR-89-5, University of Toronto.