ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 73-78

A Discriminative HCNN Modeling

Bojan Petek
University of Ljubljana
Faculty of Electrical Engineering and Computer Science
61000 Ljubljana, Slovenia

Abstract. Several systems following a concept of the dynamical systems approach to automatic speech recognition [10] have been developed so far. Most of the initial system evaluations have been carried out on smaller speech recognition tasks, such as speaker-independent digit recognition experiments [4], or speaker-dependent continuous speech recognition evaluations having constrained domains with identical training and test set vocabularies [9]. Several extensions to the Hidden Control Neural Network (HCNN) architecture [4] have been proposed and evaluated on such smaller tasks [5, 6, 7]. This paper reports the initial performance analysis of the recently proposed acoustics-error predictive context-dependent HCNN system [6] on a standarized task, i.e., speaker-independent countinuous speech recognition tests using the DARPA TIMIT acoustic-phonetic continuous speech corpora.

1 Introduction

Recently, a non-linear predictive approach has been proposed for automatic speech recognition (ASR), and used in several systems, e.g., "Neural Prediction Model", NPM, by Iso and Watanabe [2, 3], "Hidden Control Neural Network", HCNN, by Levin [4] and "Linked Predictive Neural Networks", LPNN, by Tebelskis and Waibel [8, 9]. 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 [10]. Initial evaluations of these models were carried out on small vocabulary recognition tasks, such as speaker-independent digit recognition [2, 4], yielding high recognition performances, and large vocabulary continuous speech recognition extensions [3, 9, 5].

One of the most important problems of these large vocabulary system extensions was found to be a poor discrimination among predictive models of the system [9]. To address this problem, we have recently investigated if the prediction error vectors generated by the acoustic HCNNs could be used to increase the discrimination power of the system. Specifically, by using the results from discriminant analysis, an *acoustics-error predictive HCNN modeling* has been proposed to address the discriminatory problem of the system [6, 7].

Initial evaluations of the proposed HCNN system extensions have been carried out on Slovenian translations of the CMU's Conference Registration Dialogs using the speaker dependent continuous speech recognition experiments. At perplexity 100 without using the grammar, the proposed solutions yielded an increase in word recognition accuracy from 76% (baseline system) to 87% (extended system, i.e., using the acoustics-error predictive HCNN modeling) on the test set database [7].

In order to make more general judgement about a maturity of the extended

HCNN system, this paper presents its evaluation on the standarized TIMIT speech database. The reported speaker independent continuous speech recognition experiments used the Core Test Set of the TIMIT database.

The paper is organized as follows. The Section 2 presents a brief overview of the concept of acoustics-error predictive HCNN modeling and the current training and testing procedures used. Descriptions of the task, the training and test sets used in the experiments are given in Section 3. Next, the Linear Discriminant Analysis (LDA) results are presented and discussed. By using the insights from the LDA, a dimensionality of the error predicting part of the extended HCNN model is determined and the performance evaluation of the system on the core test set is given. Finally, the conclusions and the future work are summarized in the Section 4.

2 Acoustics-Error Predicting HCNN Model

It is well known that a variance modeling may improve the performance of a speech recognition system. In contrast to the traditional technique of variance modeling we proposed a different, i.e., ***predictive approach*** [6]. We supported each of the acoustic predicting HCNN models with a separate, squared error predicting HCNN network. This composite model now models ***dynamics*** in acoustic and in error signal space (Figure 1).

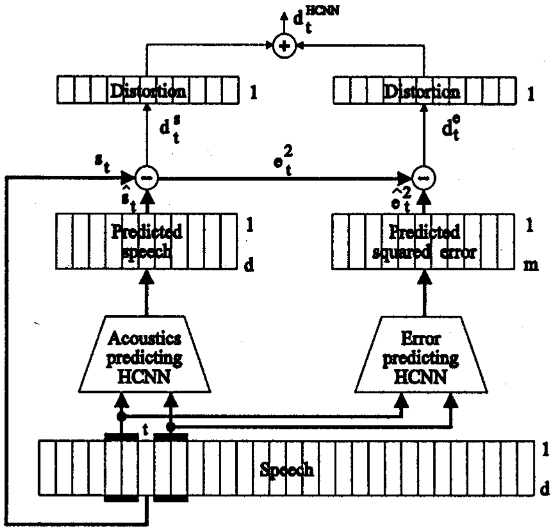

Figure 1: Acoustics-error predictive HCNN modeling.

The dimensionality of the vector in the error predicting part of the HCNN model (i.e., the parameter m in Figure 1) is determined by the LDA. First, the error vector signal of the acoustics predicting HCNN model on an optimal alignment path using the training set is saved. All misrecognition alignments are excluded from the LDA. Out of the d-component error signal only the most important vector components which are identified by the LDA, (i.e., m of them), are selected for modeling by the error predicting HCNN (see Figure 1). The important vector components in the LDA are those which maximize the separability among the phoneme classes.

The expected benefit of modeling the error signal dynamics is the following. Since the current training algorithm enables learning to the HCNNs only in a within class regions of the continuous speech utterances, several models may undesirably develop very similar predictions of acoustic realizations of the different classes. Despite producing similar acoustic predictions (i.e., similar values of the Euclidean distances between the predicted and the observed speech), the models are expected to produce different values of the prediction error vector ***components***. By virtue of modeling dynamics of the most important error vector components, the discrimination power among the acoustics-error predictive HCNNs is expected to increase. The experimental results reported in [6] support this hypothesis.

2.1 Training and testing the HCNN system

A training of the HCNN system starts with the acoustic predicting parts of the HCNN models. The error predicting networks start learning after the convergence of the acoustic parts of the models (Figure 1). All error predicting models are trained on the optimal alignment path determined by the acoustic part of each model on the training token.

During testing, each acoustics-error predictive HCNN model outputs two distortion scores at time t, i.e., one for acoustic prediction (denoted by d_t^s) and one for the error prediction (d_t^e)

$$\begin{aligned} d_t^s &= \|\mathbf{s}_t - \hat{\mathbf{s}}_t\|^2 \\ d_t^e &= \|\mathbf{e}_t - \hat{\mathbf{e}}_t\|^2 \end{aligned} \qquad (1)$$

where $\|\cdot\|$ denotes the Euclidean distance between the observed and predicted vectors. The meanings of $\mathbf{s}, \hat{\mathbf{s}}, \mathbf{e}, \hat{\mathbf{e}}$ are presented in Figure 1.

The final score of the acoustics-error predictive HCNN model, d^{HCNN}, is simply the sum of the distance scores of acoustic and error predicting parts, i.e., $d_t^{HCNN} = d_t^s + d_t^e$ (Figure 1).

3 Experimental results

The acoustics-error predictive HCNN system was trained on the DARPA TIMIT acoustic-phonetic continuous speech corpus. Out of the training portion of the database, 4 male and 2 female speakers per dialect region were selected. Therefore, the resulting training set consisted of 48 speakers (32 male and 16 female). From each of the speaker, 5 SX and 3 SI sentences were used, thus yielding a 384 sentence training set containing 3159 words.

From the test portion of the database, 24 speaker Core Test Set was used for the system evaluations. In this set, each of the 2 male and 1 female speakers per dialect region contributes 5 SX and 3 SI sentences, thus yielding a 192 sentence test set having 1570 words. SA sentences were excluded from either the training or test set.

The acoustic prototype 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, $d = 16$ in Figure 1). The error prototype network didn't have a split first hidden layer. The modeled error vector components were $e_1^2, e_2^2, e_5^2, e_6^2$, thus $m = 4$ (Figure 1 and Section 3.1). The networks were fully connected. Both networks had a direct input-output connections (i.e., weights).

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.

A total of 46 phoneme models (45 found in the pronunciation dictionary plus a silence) were represented by a 92 acoustics-error predictive HCNNs (2 alternates per phoneme model). Before beginning of the training, 61 different phonetic notations found in the sentence label files were rewritten to the 46 phoneme classes modeled by the HCNN system.

3.1 Discriminant analysis results

Initially, the LDA was carried out on the acoustics predicting HCNN system using the training set data. In this experiment, the LDA was applied to the squared prediciton error vector signal $\mathbf{e}^2$ (Figure 1). The main goal of this analysis was to identify the most important components of the error vector

$\mathbf{e}^2$ vector component	Stand. canon. coefficient	Structure coefficient
e_1^2	0.56218	0.70298
e_2^2	0.16332	0.53274
e_3^2	0.09358	0.48643
e_4^2	-0.05877	0.45253
e_5^2	0.12024	0.53618
e_6^2	0.16823	0.53976
e_7^2	0.05664	0.49316
e_8^2	0.10968	0.51236
e_9^2	0.10528	0.50047
e_{10}^2	0.05423	0.45958
e_{11}^2	0.08552	0.45149
e_{12}^2	0.03058	0.40911
e_{13}^2	0.13971	0.45065
e_{14}^2	0.10257	0.40456
e_{15}^2	0.04212	0.30926
e_{16}^2	0.04760	0.27855
Eigenvalue	0.1240	
Canonical corr.	0.3322	

Table 1: *Summary of the LDA results on the squared prediction error signal for the first canonical discriminant function,* F_1.

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 73-78

to additionally support discrimination among the models. These components were later modeled by the error predicting part of the expanded HCNN model. Only the first canonical discriminant function among statistically significant functions found by LDA, F_1, was considered. These LDA results are summarized in Table 1. The canonical correlation coefficient found by the LDA experiment is statistically significant (Table 1). Its value is rather high for the assumption on $\mathbf{e}^2$ being the white noise. This justifies the use of prediction error to additionally support discrimination power of the system.

The highest loadings were found to be on $e_1^2, e_2^2, e_5^2, e_6^2$ prediction error vector components (Table 1). The LDA results therefore suggest that the dimensionality of the modeled error vector signal should be reduced, e.g., $m = 4$ (Figure 1).

3.2 Recognition results

A separate word-pair grammar experiments have been completed in each of the 8 dialect regions. The size of the vocabulary in each region was determined by the number of unique words found in the Core Test Set utterances within the particular dialect region. This information together with the word recognition accuracy results is summarized in the Table 2.

Dialect region	*Speakers*	*Total words*	*Vocabulary size*	*S*	*D*	*I*	*Word accuracy*
1	mdab0,mwbt0 felc0	193	146	8%	2%	2%	88%
2	mtas1,mwew0 fpas0	206	158	4%	2%	1%	93%
3	mjmp0,mlnt0 fpkt0	205	170	0%	0%	0%	100%
4	mlll0,mtls0 fjlm0	187	152	5%	1%	2%	92%
5	mbpm0,mklt0 fnlp0	177	147	5%	1%	1%	93%
6	mcmj0,mjdh0 fmgd0	211	163	7%	3%	0%	90%
7	mgrt0,mnjm0 fdhc0	190	149	7%	0%	1%	92%
8	mjln0,mpam0 fmld0	201	154	3%	2%	0%	95%

Table 2: *Performance scores of the HCNN system on the TIMIT Core Test Set. (S = Substitutions, D = Deletions, I = Insertions)*

4 Discussion

The acoustics-error predictive HCNN modeling has been discussed and evaluated on the standarized task, i.e., in speaker-independent countinuous speech recognition experiments using the DARPA TIMIT acoustic-phonetic continuous speech corpora. The word-pair grammar results obtained on this standarized task are encouraging, given the fact that no optimizations on either the acoustics or the error predicting HCNN model architectures were investigated so far. The performance analysis results comparing the baseline and composite (i.e., error predicting) HCNN modeling on a smaller task can be found in references [6, 7].

Future work should therefore address the optimal HCNN architecture issue with the emphasis on a higher perplexity performance evaluations. More theoretical work is also needed on the optimal integration of the acoustic and the error predicting parts of the HCNN model. Finally, one of the most important remaining issues is the application or development of a more efficient (and discriminant) training procedure for the systems following the concept of dynamical systems approach to ASR. The current corrective training procedures tend to be too task specific and lead to generalization problems when porting the system from one task (or vocabulary) to another. Thus, the most appropriate way appears to be the use of efficient learning (e.g., such as [1]).

5 Acknowledgements

The author is grateful to Professor Bogomir Horvat and Dr. Zdravko Kačič from the Faculty of Technical Sciences, University of Maribor, Slovenia, who enabled the experimental part of this work. The valuable comments of the anonymous referees are also gratefully acknowledged.

References

[1] J. B. Hampshire II: *A Differential Theory of Learning for Efficient Statistical Pattern Recognition.* PhD. Thesis, Carnegie Mellon University, Dept. of Electrical & Computer Engineering (September 1993)

[2] K. Iso, and T. Watanabe: Speaker-Independent Word Recognition Using a Neural Prediction Model. Proc. IEEE Int. Conf. on ASSP, 441-444 (1990)

[3] K. Iso, and T. Watanabe: Large Vocabulary Speech Recognition Using Neural Prediction Model. Proc. IEEE Int. Conf. on ASSP, 57-60 (1991)

[4] E. Levin: Word Recognition Using Hidden Control Neural Architecture. Proc. IEEE Int. Conf. on ASSP, 433-436 (1990)

[5] B. Petek, A. H. Waibel, and J. M. Tebelskis: Integrated and Phoneme-Function Word Architecture of Hidden Control Neural Networks for Continuous Speech Recognition. Speech Communication, Special Issue on Eurospeech-91, 11, Nos. 2-3, 273-282 (1992)

[6] B. Petek, A. Ferligoj: Exploiting Prediction Error in a Predictive-Based Connectionist Speech Recognition System. Proc. IEEE Int. Conf. on ASSP, II-267 - II-270 (1993)

[7] B. Petek, A. Ferligoj: On Use of Discriminant Analysis in Predictive Connectionist Speech Recognition. Proc. ESCA Eurospeech-93, 1611-1614 (1993)

[8] J. Tebelskis, and A. Waibel: Large Vocabulary Recognition Using Linked Predictive Neural Networks. Proc. IEEE Int. Conf. on ASSP, 437-440 (1990)

[9] J. Tebelskis, A. Waibel, B. Petek, and O. Schmidbauer: Continuous Speech Recognition Using Linked Predictive Neural Networks. Proc. IEEE Int. Conf. on ASSP, 61-64 (1991)

[10] N. Tishby: A Dynamical Systems Approach to Speech Processing. Proc. IEEE Int. Conf. on ASSP, 365-368 (1990)