COMPARISON OF A NEW HYBRID CONNECTIONIST-SCHMM APPROACH WITH OTHER HYBRID APPROACHES FOR SPEECH RECOGNITION

Hans-Peter Hutter

Speech Processing Group, Computer Engineering and Networks Laboratory,
ETH-Zentrum, CH-8092 Zurich, Switzerland
E-mail: hutter@tik.ee.ethz.ch

ABSTRACT

This paper compares a newly proposed hybrid connectionist-SCHMM approach [5] with other hybrid approaches. In the new approach a multilayer perceptron (MLP) replaces the conventional codebooks of semi-continuous HMMs. The MLP is therefore trained on so-called basic elements (phones and phone parts) in such a way that the outputs of the network estimate the a posteriori probabilities of these elements, given a context of input vectors. These a posteriori estimates are converted into scaled likelihoods, which are then used as observation probabilities in the framework of classical SCHMMs. The remaining parameters of the SCHMMs are trained with the well-known Baum-Welch algorithm using the estimated likelihoods of the MLP. This approach compared favorably with other recently proposed hybrid systems and classical approaches on an isolated German digit recognition task over telephone lines. It exhibited the highest recognition rate of all systems, followed by an approach using LVQ3 optimization of the codebook.

1. INTRODUCTION

Over the past decades, the most successful approaches to speech recognition tasks have been based on hidden Markov models (HMM). In spite of their strong theoretical foundation and their predominant role, HMMs have their weaknesses, too. One of their major drawbacks is the lack of discrimination that HMMs have due to the maximum likelihood training. For the last few years several research groups have tried to incorporate connectionist ideas into the statistical framework in order to improve the discrimination capability of the HMMs:

A larger group of researchers replaces or optimizes the classic codebooks of discrete density HMMs (DDHMMs) with connectionist approaches (e.g., [2, 8, 6, 4]). ANNs, typically trained on phonemes, are used in [8] as frame labelers. In [6, 4] learning vector quantization procedures, based on the algorithms proposed by T. Kohonen in [7], are used to optimize classically generated codebooks.

Other research groups are following the ideas of Bourlard et al. [1], who applies MLPs to estimate a posteriori probabilities. The MLPs are therefore trained to classify one or more frames into phoneme classes. The a posteriori probability estimates at the outputs of the MLPs are converted into scaled likelihoods, which substitute for the phoneme observation probabilities in the HMMs. This approach imposes the modeling of each distinct phoneme in a single HMM state.

In this paper a new hybrid approach based on semi-continuous HMMs is described and compared with the other hybrid approaches. Section 2 elaborates the new connectionist-SCHMM approach, Section 3 describes the training and test material, Section 4 outlines the MLP training, and Section 5 compares the different approaches.

2. CONNECTIONIST-SCHMM APPROACH

Semi-continuous HMMs (SCHMMs) [3] calculate the observation probability $b_i(x) = P(x|q_t = S_i)$ of a feature vector x in one of the SCHMM states S_i as the combination of a discrete observation probability $b_i(c_j)$ of the codebook class c_j and a class-dependent probability density function $p(x|c_j)$ according to the equation

$$b_i(x) = \sum_{j=1}^{M} p(x|c_j) b_i(c_j) , \qquad (1)$$

where M denotes the number of classes or partitions in the codebook. The pdfs of the codebook partitions are modeled as Gaussian densities, whose parameters are estimated together with the other SCHMM parameters within an extended Baum-Welch training procedure.

This work was pursued in the framework of COST 232 and was supported by the Swiss Federal Administration (BBW)

0-7803-2431-5/95 $4.00

Digit		Transcription	
		standard	adapted
0	null	nʊl	[?nʊl?]
1	eins	a͡ins	[?ains?]
2	zwei	ʦva͡i	[?tsvai?]
	zwo	ʦvo:	[?tsvo?]
3	drei	dra͡i	[?drai?]
4	vier	fi:ɐ̯	[?fiər?]
5	fünf	fʏnf	[?fʏnf?]
6	sechs	zɛks	[?sɛ?ks?]
7	sieben	zi:bn̩	[?si?bən?]
8	acht	axt	[?ax?t?]
9	neun	nɔyn	[?nɔʏn?]

Table 1: Adapted transcription of the German digits into the phonetic elements (IPA notation)

In the new hybrid approach, the codebook of the classic SCHMMs is replaced by an MLP, which directly estimates the pdfs of the codebook classes in Eq. 1. These classes correspond to phonetic elements (phones or subphones). For that purpose, an MLP is trained to classify the 20 phonetic elements of the German digits. The transcription of the German digits into these phonetic elements is given in Table 1. [?] is the symbol for silence at the word boundaries or for closures. For the identification of the classes, we use a "1-from-N" coding scheme at the output layer of the MLP. The error function to be minimized is the mean-square error (MSE) between the actual MLP outputs and the desired outputs. It has been proven by several authors (e.g., in [1]) that an MLP trained in such a manner estimates the a posteriori probabilities $P(c_j|x)$ of the phonetic element classes c_j given the input vector x, if the MLP has enough parameters and reaches the global minimum of the error function. In order to get probability estimates, these a posteriori probabilities are converted into scaled likelihoods using

$$\frac{p(x|c_j)}{p(x)} = \frac{P(c_j|x)}{P(c_j)} \,. \qquad (2)$$

The a priori probabilities of the classes $P(c_j)$ in Equation 2 are estimated from the training data. The division by the a priori class probabilities was already applied by Bourlard et al. with quite success in order to compensate for the a priori class probabilities in the training set, which may not be representative ([1, pp. 174,180]).

The scaled likelihoods substitute for the observation pdfs in the codebook partitions. While training, these likelihoods are needed to reestimate the remaining parameters of the SCHMMs with the normal Baum-Welch reestimation algorithm. During recognition, the scaled likelihoods are used in the same way to calculate the likelihoods of the SCHMMs.

The advantage of this hybrid approach compared to the classical SCHMMs is that here the pdfs of the codebook classes are not restricted to a simple parametric form (Gaussian) but are trained discriminatively with only weak assumptions about their parametric form. In contrast to the hybrid approach described by Bourlard et al. in [1], a phonetic element may here be modeled in more than a single state. This allows a more accurate temporal modeling and the use of whole-word HMMs. In addition to that, this approach automatically takes into account pronunciation variations of speakers and misclassifications of the MLP, which may occur at phone boundaries. This advantage applies also compared to those hybrid approaches that use an MLP as frame labeler, whose labels are taken as observations for discrete density HMMs. These approaches are vulnerable to misclassifications of the MLP, too.

3. TRAINING AND TEST DATA

The new approach was trained and tested on a speaker-independent isolated German digit recognition task over real telephone lines. The training set for the whole hybrid system comprised 895 utterances spoken by 80 speakers, which originated from different dialect regions in Switzerland. Although the speakers read the digits in standard German, the influence of the Swiss dialect was quite audible. To account for that, the transcription of the German digits has been adapted as shown in Table 1. The speech signal was recorded at an analog telephone connection. It was sampled at 7.2 kS/s with a 12 bit A/D-converter. As for the test set, another 835 utterances from 5 sessions with each of 15 additional speakers were recorded in the same way. Again, a different telephone line was selected for each session.

For the following experiments and comparisons 13 weighted LPC cepstral coefficients (0th coeff. was always 0) were derived from 8 autoregressive coefficients. The autoregressive coefficients were extracted every 10 ms from a signal window of 30 ms. Before the feature extraction, a Hamming window and a preemphasis of 0.97 were applied to the signal.

4. TRAINING OF THE MLPS

In order to estimate the a posteriori probabilities of the phone element classes, fully connected MLPs with two hidden layers were investigated. The tanh function was used as nonlinearity, with an offset of 0.5 added to the

output layer activities. For the training and evaluation of the MLPs, the whole training set was labeled semi-automatically with the phonetic elements. The MLPs were trained with three quarters of the training corpus, the rest was set aside for cross validation. This resulted in 3024 training templates with a total of 42,707 frames. The evaluation set counted 993 templates with 13,372 frames.

In each training iteration 42,707 randomly selected training patterns were applied to the MLPs as follows: A random template of a randomly chosen phonetic element class was first selected. Then, again randomly, a context of successive frames was picked out of the template. A context of 1 up to 9 frames of successive feature vectors was applied to the net inputs at a time. At the same time the outputs of the MLP were set to the code corresponding to the center frame. The MSE between the net outputs and the desired outputs was back-propagated through the network by means of the standard Error Back-Propagation algorithm. The weights of the networks were updated after each pattern presentation using a learning rate of 0.01 with no momentum term. After every 5 iterations, the MLPs were cross validated with the evaluation set. Typically, the MLPs reached the highest performance after 40 iterations, albeit the performance degraded only slightly with additional iterations.

MLPs with input contexts ranging from 1 to 9 frames were investigated. The first hidden layer had approximately the same number of nodes as the input layer and the second hidden layer the same as the output layer. Hence, the number of parameters in the MLPs ranged from about 1600 to 17,000. Frame classification results on the evaluation set of the best MLPs for the different input contexts are given in Table 2. While the MLP with the best absolute frame recognition rate averaged over all classes was an MLP with 3 frames context, the best absolute frame recognition rate exhibited an MLP with a context of 5 frames.

Num. of frames context	MLP topology: Nodes in layer 1	2	3	4	Recognition rate on eval. set: frame	mean ov. classes
1	13	20	20	20	45.2	38.3
3	39	40	20	20	47.2	44.1
5	65	60	20	20	47.6	43.6
7	91	90	20	20	47.5	43.2
9	117	120	20	20	45.3	39.3

Table 2: Recognition rates of MLPs with different contexts on the evaluation set (raw frame rate and averaged over all classes)

5. COMPARISON WITH OTHER APPROACHES

For the comparison of the different hybrid systems, we decided to use the MLP with the best frame recognition rate averaged over all classes, the MLP with a 3 frames context (MLP3). The following list gives a short description of each of the different approaches that have been considered:

DDHMM: This is the base system with discrete density HMMs. The models are left-right HMMs with no skipped states. All HMMs have 17 states with one non-emitting entry and exit state, resp. The codebook size is 64 and it has been generated with the well known LBG algorithm. The HMMs were trained with the standard Baum-Welch algorithm.

This system yields a 97.5 % recognition rate on the test set if delta cepstrum, log energy, and delta energy are used in addition to the weighted cepstrum.

Labeler: MLP3 is used as frame classifier, which produces a sequence of labels for DDHMMs. The DDHMMs have the same structure as in **DDHMM**, but the codebook size has to be reduced to 20. This approach was applied by Compernolle et al. (see, e. g., [8]) for the recognition of the Flemish digits.

FuzzyVQ: In this approach, the outputs of MLP3 are considered as scaled likelihoods of the phonetic element classes. The DDHMMs trained in **Labeler** are used as a semi-continuous decoder. The observation probability of a feature vector is calculated with Equation 1 using the observation probabilities trained in **Labeler** and the scaled likelihoods estimated by MLP3. This approach is described in [8], too.

SCHMM/MLP: This is the connectionist-SCHMM approach described in Section 2. The structure of the SCHMMs is the same as in **Labeler**.

LVQ3: Under this label, the codebook generated with the LBG algorithm was optimized on the phonetic element classes using the LVQ3 algorithm proposed in [7]. The performance on the evaluation set was monitored during the optimization using the quality measure introduced in [4]. $\alpha(t)$ started with $\alpha(0) = 0.02$ and decreased linearly

to 0 within 100,000 steps. ϵ was chosen 0.2 and the relative width of the "window" was set to 20 %. The best performance on the evaluation set was achieved after two iterations over all 42,707 training vectors.

Bourlard: This is the approach of Bourlard et al., which has been reported on for example in [1]. Each phonetic element is modeled in a single state, whose (scaled) observation pdf is directly estimated by MLP3. The transition probabilities are trained with the standard Baum-Welch algorithm.

SCHMM1/2: This label stands for the standard SCHMM approach proposed by Huang et al. with codebooks of size 32 and 64, resp. The codebooks were estimated together with the other SCHMM parameters following the formulae in [3]. The trained models and the generated codebooks of **DDHMM** were taken as starting point of the SCHMM training.

The recognition rates of the different approaches on the test set are shown in Table 3. As can be seen, the proposed connectionist-SCHMM system outperforms all approaches considered in this experiment. The improvement relative to the other systems is significant with the exceptions of **LVQ3** and **SCHMM2**. These two, however, use codebooks that have three times the size of the ones used in the MLP based systems. After all, it is worth noting that the LVQ3 optimization of the codebooks on the phonetic element classes is also a promising approach.

Approach	Recognition rates on test set [%]
DDHMM	90.4
Labeler	89.7
FuzzyVQ	90.4
SCHMM/MLP	93.2
LVQ3	92.9
Bourlard	87.3
SCHMM1	89.3
SCHMM2	92.5

Table 3: Recognition rates of the different classic and hybrid approaches

6. CONCLUSION

The new connectionist-SCHMM approach seems to be a promising alternative to existing hybrid systems. Instead of taking early decisions at the frame level, the MLP passes all information to the SCHMMs. In addition to that, the approach is very flexible and can adapt to misclassifications of the MLP. A similar performance was attained by optimizing the classic codebooks with the LVQ3 algorithm on phonetic elements. We are currently investigating how additional features may best be incorporated in these hybrid systems.

7. REFERENCES

[1] H. Bourlard and N. Morgan. *Connectionist speech recognition: a hybrid approach.* Kluwer Academic Publishers, Boston, Dordrecht, London, 1994.

[2] M. A. Franzini, A. H. Waibel, and K.-F. Lee. Recent work in continuous speech recognition using the connectionist viterbi training procedure. In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, volume 3, pages 1213–1216. ESCA, September 1991.

[3] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh Information Technology Series 7. Edinburgh University Press, 22 George Square, Edinburgh, 1990.

[4] H.-P. Hutter. Bericht über die Wort-Erkennungs-Systeme RECO1 und RECO2. Institut für Elektronik, ETH Zürich, May 1991.

[5] H.-P. Hutter and B. Pfister. Neuartiger hybrider SKHMM/KNN-Ansatz für die Spracherkennung. In K. Fellbaum, editor, *Elektronische Sprachsignalverarbeitung*, Studientexte zur Sprachkommunikation, Heft 11, pages 90–97. Technische Universität Berlin, October 1994.

[6] H. Iwamida, S. Katagiri, and E. McDermott. Speaker-independent large vocabulary word recognition using an LVQ/HMM hybrid algorithm. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 553–556, May 1991.

[7] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, September 1990.

[8] Ph. Le Cerf, W. Ma, and D. Van Compernolle. Multilayer perceptrons as labelers for hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 2(1):185–193, January 1994.

[9] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, pages 257–285, 1989.