SPEAKER ADAPTATION FOR HYBRID MMI/ CONNECTIONIST SPEECH RECOGNITION SYSTEMS

J. Rottland, Ch. Neukirchen, G. Rigoll

Gerhard-Mercator-University Germany
Faculty of Electrical Engineering
Department of Computer Science
D-47057 Duisburg, Germany
e-mail: {rottland, chn, rigoll}@fb9-ti.uni-duisburg.de
http://www.fb9-ti.uni-duisburg.de

ABSTRACT

In this paper we present a new adaptation technique for our hybrid large vocabulary continuous speech recognition system. In most adaptation approaches the HMM parameters are reestimated. In our approach, however, we train a speaker independent continuous speech recognizer, then we keep the HMM parameters fixed and we train a second network, which transforms the features of the adaptation data to fit the HMM parameters. Thus, less parameters have to be estimated, and therefore this approach performs well even for a small number of adaptation data. With this approach we achieve relative improvements in recognition rates on the Wall Street Journal (WSJ) task of 16.5%.

1. INTRODUCTION

Over the last years we developed a high performance speech recognition system based on a new hybrid approach [5][6]. Recently [7] we could show that the performance of our hybrid connectionist/HMM speaker independent continuous speech recognition systems is very close to the performance of standard continuous HMM systems for the Wall Street Journal database (WSJ0). The WSJ is a large vocabulary speech recognition task for read speech. In this paper we will describe how to extend our approach in order to perform very effective speaker adaptation using novel ideas based on information theory algorithms, which fit well to the basic principle of our hybrid approach. We will do this without changing any HMM parameters in opposite to the common adaptation techniques like MAP and MLLR [2]. We perform some kind of feature transformation to adapt the new speaker to the speaker independent system. The reason for this is, that in speaker independent systems a large amount of parameters is estimated. These parameters would have to be reestimated with a relatively small amount of adaptation data, if we would adapt the HMM parameters. In our approach we will only estimate a relatively small amount of approximately 2800 weights.

2. SYSTEM DESCRIPTION

2.1 BASELINE SYSTEM

The baseline system we use is a system similar to the one presented in [7], which consists basically of a neural vector quantizer (VQ) and discrete HMMs. The VQ is a single layer neural network with the Euclidean activation function:

$$y_m = \sum_N (x_n - g_{mn})^2 \qquad (1)$$

The NN (Figure 1) is trained with the information theory based maximum mutual information (MMI) paradigm described in [5][6]. This paradigm maximizes the mutual information $I(Y,W)=H(W)-H(W|Y)$ between the sequence of the firing neurons y_m and the corresponding phonetic description w_i which is derived from the transcriptions or an alignment. $H(W)$ is not affected by the neural network, thus only

$$H(W|Y) = -\sum_I \sum_M P(w_i, y_m) \cdot \log P(w_i, y_m) \qquad (2)$$

has to be minimized to maximize $I(Y,W)$. The probabilities $P(w_i,y_m)$ are estimated using the firing sequence of the network. The NN is a "winner takes all" network i.e. the output of the network is the number of the neuron with the lowest activation (distance to corresponding prototype). After presenting the full training set to the network, we can compute the probabilities from a matrix in which we count the observations of neuron m firing, when a phoneme w_i was seen on the input layer.

To train the neural networks with a gradient based method we have to calculate the derivative of $\partial H(W|Y) / \partial g$. Therefore during the training procedure the output is smoothed with a softmax function $O_m(x) = e^{\frac{y_m(x)}{T}} \Big/ \sum_{l=1}^{J} e^{\frac{y_l(x)}{T}}$. By using the softmax it is now possible to calculate the derivative of $\partial H(W|Y) / \partial g$ which can be written as:

$$\frac{\partial H(W|Y)}{\partial g} = \sum_{i=1}^{I} \sum_{m=1}^{M} \frac{\partial H(W|Y)}{\partial P(w_i, y_m)} \cdot \sum_{n=1}^{train\ data} \frac{\partial P(w_i, y_m)}{\partial O_m(x(n))} \cdot \sum_{l=1}^{M} \frac{\partial O_m(x(n))}{\partial y_l(x(n))} \cdot \frac{\partial y_l(x(n))}{\partial g} \qquad (3)$$

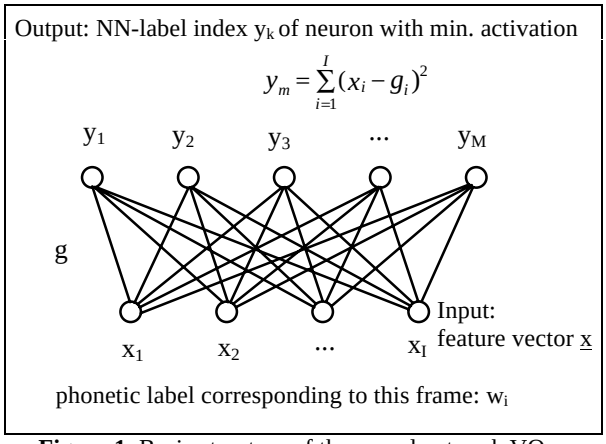

Figure 1: Basic structure of the neural network VQ

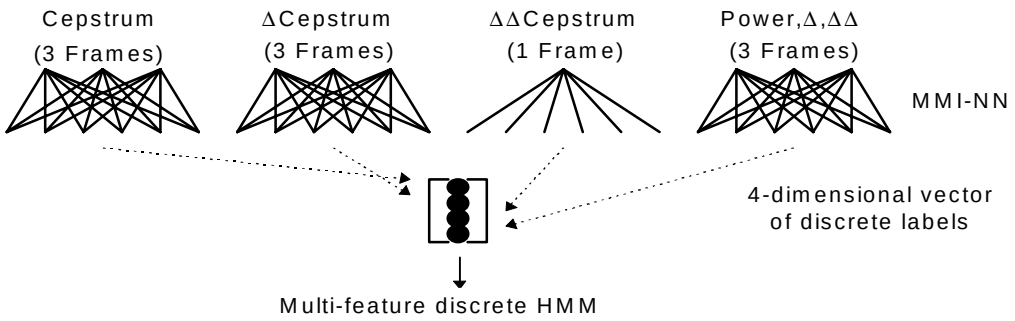

Figure 2: Context dependent connectionist vector quantizer with four features

The system is partitioned into four streams. Each of the four streams is trained separately. The first stream consist of 12 Mel-scaled cepstral features which are computed every 10 ms. A hamming window with a size of 25 ms is applied for each frame. The second and third stream are the first and second order derivatives of the cepstral features. The last stream is computed from the power in each frame plus first and second order derivatives. For a better context dependency of the neural network three adjacent frames are used as input vector for each network. This results in an input size of 36 for the first three streams (networks) and an input size of 9 for the fourth network. The complete structure of the network is shown in Figure 2. The output size for all networks was equally chosen to 300. This size leads to a number of 27900 neural weights. The output of each of the four networks is a single integer representing the index of the nearest prototype. All outputs are calculated for each frame. The output of the network is the input of a discrete HMM system. The HMMs are context dependent (cross word triphones) which are state clustered to approximately 6000 states. This results in a total number of 7.2 million HMM parameters. Because of the large number of HMM parameters smoothing techniques are applied [1][8]. The system is gender independent , i.e. there is only one set of HMMs for both genders.

Recognition is performed in two steps. In the first step the bigram language model is used. The result of this first recognition is a lattice, which is then rescored in the second step with the trigram language model. For the 5k speaker independent task (si-84) the error rate for this system is 5.7% [7].

2.2 ADAPTATION SYSTEM

As outlined in the previous sections, our large vocabulary speech recognition system differs substantially from standard systems, concerning its architecture as well as the basic theoretical framework it relies on. We have been able to exploit these facts in order to develop new principles for speaker adaptation which make extensive use of these special conditions and are presented here for the first time. Therefore, in compliance with our basic approach using MMI neural networks as neural vector quantizers in a discrete HMM system, these new speaker adaptation techniques are based on information theory algorithms and make use of the fact that it is possible to represent our discrete system structure as one large neural network trainable with gradient descent methods.

The latter fact is exploited in our adaptation scheme by extending the MMI neural networks by an additional layer with a scalar product activation function, as outlined in Figure 3. Our goal is then to train this additional layer in order to transform the speaker dependent adaptation data to match the distribution obtained from the original speaker independent training. For this purpose, we apply a novel information theory-based training scheme to the weights of the additional layer, while the weights of the original MMI neural network are frozen and therefore remain unchanged. In this way it is possible to model the speaker characteristics using only a relatively small number of parameters and to adapt the new speakers' data to the existing Markov models which remain unchanged during this procedure, rather than adapting the numerous Markov model parameters to the new speaker data. This is only possible because our hybrid system makes use of trainable vector quantizers implemented in form of the MMI neural networks. Therefore, if we can adapt such a vector quantizer to produce the same neural firing as in the original speaker independent case for a new speaker, the speaker independent probabilities of the underlying HMMs are still valid for this new VQ and can be retained. This leads to a dramatic reduction of the parameters that have to be adapted, which now consist only of the 2800 weights of the four input networks, compared to the large amount of about 7.2 million HMM parameters of the hybrid system.

While it is obvious that such a procedure is only feasible due to

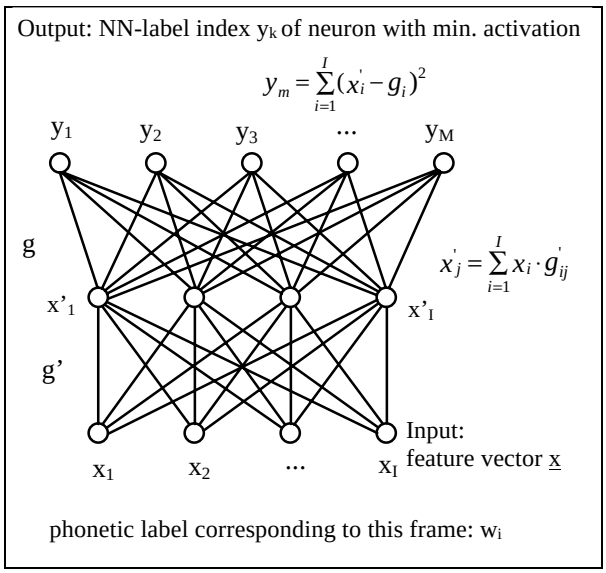

Figure 3: New network topology for speaker adaptation

the hybrid architecture of our speech recognition system, it is furthermore also possible to show that in addition also the MMI training criterion of our hybrid approach can be exploited in a very elegant and constructive manner to positively contribute to our new approach to speaker adaptation. For this purpose, one has to recall Jensens inequality, which is one of the most important relationships used in information theory-based parameter estimation techniques, such as ML or MMI approaches. This inequality states that the expectation value of a logarithmic probability distribution log*f(y)* is always maximum if *f(y)* represents the true distribution $f^*(y)$ used for computing the expectation value. This can be expressed as:

$$\int_Y f^*(y)\cdot\log f^*(y)dy \geq \int_Y f^*(y)\cdot\log f(y)dy \qquad (4)$$

This inequality is also valid if we assume that the function f is a conditional probability function of the form $f = p(w|y)$. In this case, the expectation value of a conditional logarithmic distribution can be computed as:

$$E\{\log p(w|y)\} = \int_W\int_Y p(w,y)\cdot\log p(w|y)dwdy \qquad (5)$$

Thus, Equation (4) can be also expressed as

$$\int_W\int_Y p^*(w,y)\cdot\log p^*(w|y)dy \geq \int_W\int_Y p^*(w,y)\cdot\log p(w|y)dy \qquad (6)$$

This can be exploited for speaker adaptation in the following way: Our MMI neural training approach in [5][6][7] is capable of training the weights of a neural network in order to produce neural firing probabilities that minimize the entropy *H(W|Y)* by using gradient descent techniques. If in the training criterion in Equation (2) we use for the non-logarithmic probabilities - which are used for computing the expectation value - some pre-computed probabilities *p** and for the logarithmic probabilities produced by the neural net the probabilities *p*, then the neural training procedure will try to do its best to maximize the expression

$$\max\{\sum_W\sum_Y p^*(w,y)\cdot\log p(w|y)\} \qquad (7)$$

However, according to (6), this can be only achieved if the probabilities p converge to the probabilities *p**. Therefore, by forcing the neural net to maximize the training criterion in Equation (7), it will automatically train the weights so that the resulting neural net probabilities *p* converge as close as possible to the probabilities *p**. As mentioned earlier, our goal in speaker adaptation was to train the extended neural network for the new speaker so that it produces the same neural firing distributions as in the original speaker independent case. Therefore, for performing speaker adaptation in our hybrid system, we proceed as follows:

1) The speaker-independent training of the HMMs and the neural networks is carried out. This has to be done only once, and the system is now able to recognize speech in speaker-independent mode. The neural firing probabilities obtained from speaker-independent training of our neural vector quantizer are stored. We call these probabilities *p*(y|w).*

2) The weights of the neural vector quantizer are initialized with the weights obtained from the speaker-independent system, and the net is augmented by an additional input network, whose weight matrix is set to the identity matrix. In this way, the neural firing behavior of this extended neural vector quantizer is still exactly equal to the speaker-independent neural acoustic processor.

3) For adaptation, the input network of the neural vector quantizer is trained (while all other weights are being kept constant) with the speaker-dependent adaptation data available for the new speaker. As training criterion, we use the criterion in (7), where the values *p*(w,y)* are obtained from the probabilities stored in Step 1 using

$$p^*(w,y) = p^*(y|w)\ .\ p(w) \qquad (8)$$

and the values *p(w|y)* in Equation (7) are obtained from the actual neural firing counts delivered by the neural vector quantizers' output when the adaptation data is presented to its input layer. The probabilities in (8) can be considered as joint probabilities *p*(w,y)* that would have been produced by the speaker-independent network if some speaker-independent data W with the distribution probabilities *p(w)* equivalent to the adaptation data would have been presented to the network. These are exactly the joint probabilities that are desired as output from the adapted neural vector quantizer, because in this case the adapted quantizer would behave similar to the independent quantizer, although now the adaptation data is presented to its input. These probabilities can be considered as target values in the information theory-based training procedure implemented by our MMI learning criterion. If these values are used to compute the expectation of the logarithmic probabilities *p(w|y)*, the training algorithm of the neural network will automatically try to make these probabilities *p* similar to the probabilities *p** in order to fulfill the maximization criterion in Equation (7).

4) After 40-50 iterations using the gradient descend approach described in [6], the adaptation procedure is terminated. This will take only a few minutes per network on a modern workstation due to the small amount of adaptation data. Speaker-dependent recognition is then performed using the new extended neural vector quantizer in combination with the original HMMs which remain unchanged and have been obtained from speaker-independent training in Step 1.

Extension of the gradient descent algorithm as outlined in Equation (3) to the augmented neural network with an additional layer is quite straightforward, and can be carried out using the chain rule in the following manner:

$$\frac{\partial H(W|Y)}{\partial g'} = \sum_{i=1}^{I}\sum_{m=1}^{M}\frac{\partial H(W|Y)}{\partial P(w_i,y_m)}\cdot\sum_{n=1}^{\substack{train\\data}}\frac{\partial P(w_i,y_m)}{\partial O_m(x)}\cdot\sum_{l=1}^{M}\frac{\partial O_m(x)}{\partial y_l(x)}\cdot\sum_{k=1}^{I}\frac{\partial y_l(x)}{x'_k(x)}\cdot\frac{\partial x'_k(x)}{\partial g'} \qquad (9)$$

Additionally, the fact has to be taken into account that the partial derivative of the entropy has to be computed with respect to the probabilities *p*, whereas the probabilities *p** in (7) can be considered to be fixed in this case, which even facilitates the computational effort due to a simpler formula resulting for the derivative. The above described procedure has to be carried out for all four neural codebooks of the hybrid system.

3. EXPERIMENTS & RESULTS

The speaker independent system has been trained with the si-84 part of the 1992 WSJ database. The phonetic transcriptions were derived from the 1993 LIMSI WSJ lexicon. The tests were performed on the 5k speaker dependent evaluation test set (sd_et_05 nvp) of the WSJ0. The language models were the original WSJ0 language models. Table 1 shows the results for

the speaker dependent test set for the baseline system, which was only trained with the speaker independent utterances. Therefore, the average error rate is higher than in [3], where speaker dependent systems were used. All recognition parameters like pruning, language model scale, etc. are kept on the values, which were used for speaker independent tests in [7].

Speaker	Correct	Substit.	Deletion	Insertion	Error
001(m)	93.3	5.3	0.9	0.0	6.1
002(f)	97.8	2.0	0.2	0.2	2.5
00A(f)	90.5	8.8	0.7	1.2	10.8
00B(m)	88.2	8.0	3.8	0.8	12.5
00C(m)	80.9	13.1	6.0	0.5	19.6
00D(m)	90.6	7.2	2.2	0.2	9.7
00F(f)	91.2	7.4	1.4	1.9	10.7
203(f)	96.0	3.2	0.8	0.3	4.2
400(m)	96.4	2.1	1.5	0.4	4.1
430(m)	96.4	3.4	0.3	0.5	4.2
431(m)	94.5	5.0	0.5	0.5	6.0
432(f)	95.4	4.3	0.3	0.3	4.8
Average	92.6	5.8	1.6	0.6	7.9

Table 1: Results in % for the sd_05 task using the si-84 cross word triphone models and the trigram language model (baseline system)

As described in [4] for each of the 12 speakers in the speaker dependent task there are 40 utterances of adaptation data in the WSJ0 database. These 40 sentences were used for the training of the adaptation layers in our approach. Table 2 shows the results for all speakers after the adaptation. The relative improvement of the average error rate is also shown in table 2.

Speaker	Corr.	Sub.	Del.	Ins.	Error	relative reduction
001(m)	95.4	3.9	0.7	0.0	4.6	24.6
002(f)	97.8	2.0	0.2	0.5	2.7	-8.0
00A(f)	90.7	8.3	1.0	0.7	10.0	7.4
00B(m)	89.0	7.5	3.5	0.3	11.3	9.6
00C(m)	85.1	10.9	4.1	0.8	15.7	19.9
00D(m)	95.2	3.6	1.2	0.2	5.1	47.4
00F(f)	90.7	7.1	2.1	1.0	10.2	4.7
203(f)	97.1	2.4	0.5	0.3	3.2	23.8
400(m)	95.7	2.6	1.7	0.4	4.7	-14.6
430(m)	97.4	2.3	0.3	0.8	3.4	19.1
431(m)	95.9	3.8	0.2	0.5	4.6	23.3
432(f)	96.2	3.8	0.0	0.0	3.8	20.8
Average	93.9	4.8	1.3	0.4	6.6	**16.5**

Table 2: Results in % for the adapted network on the sd_05 task using the unchanged si-84 cross word triphone models and the trigram language model

The improvement in error reduction is up to 47.4% for an individual speaker compared to the results of the baseline system for the same speaker. The average improvement is 16.5% for the adapted system. The total error rate of 6.6% of the adapted system is already very close to the best speaker dependent system in the official ARPA evaluation [3], which was 6.1% for this task.

4. CONCLUSION

In this paper we presented a new and powerful adaptation approach for hybrid connectionist/MMI large vocabulary speech recognition systems. This new, information theory based approach was tested on the Wall Street Journal database. Due to the special structure of our system, only few parameters have to be estimated, and thus the adaptation can be carried out with a small number of utterances. The results on the WSJ database are very promising and the improvement in error rate is comparable to other adaptation techniques like MAP and MLLR on similar tasks.

5. ACKNOWLEDGMENTS

This work was partially founded by the German National Science Foundation (DFG), project #Ri 658/6-1. Responsibility for the contend of this paper is by the authors.

LIMSI provided their 1993 Wall Street Journal lexicon.

6. REFERENCES

[1] Bahl L., Jelinek F., Mercer R. "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. Pami-5, No. 2, pages 179-190, March 1983

[2] Leggetter C., Woodland P. "Flexible Speaker Adaptation for Large Vocabulary Speech Recognition", *4th European Conference on Speech Communication and Technology*, pages 1155-1158, Madrid 1995

[3] Pallett D., Fiscus G., Fisher W., Garofolo J. "Benchmark Tests for the DARPA Spoken Language Program", *Human Language Technology*, pages 7-18, Plainsboro NJ, 1993

[4] Paul D. and Baker J. "The Design for the Wall Street Journal-based CSR Corpus", *DARPA Speech and natural language Workshop*, pages 357-362, February 1992

[5] Rigoll G. "Maximum Mutual Information Neural Networks for Hybrid Connectionist-HMM Speech Recognition Systems", *IEEE Trans. on Speech and Audio Processing, Special Issue on Neural Networks for Speech, Vol. 2, No. 1*, pages 175-184, January 1994

[6] Rigoll G., Neukirchen Ch. "A new approach to hybrid HMM/ANN speech recognition using mutual information neural networks", *Advances in Neural Information Processing Systems 9, NIPS*96*, pages 772-778, Denver 1996

[7] Rottland J., Neukirchen Ch., Willett D., Rigoll G. "Large vocabulary speech recognition with context dependent MMI-Connectionist / HMM systems using the WSJ database", *5th European Conference on Speech Communication and Technology*, pages 79-82, Rhodes 1997

[8] Schwarz R., Kimball O., Kubala F., Feng, M., Chow Y., Barry C., Makhoul J. "Robust Smoothing Methods for Discrete Hidden Markov Models", *IEEE International Conference on Acoustics, Speech and Signale Processing*, pages 548-551, Glasgow 1989