AN INCREMENTAL SPEAKER-ADAPTATION TECHNIQUE FOR HYBRID HMM-MLP RECOGNIZER

João P. Neto   *Ciro Martins*   *Luís B. Almeida*

Instituto de Engenharia de Sistemas e Computadores (INESC), Portugal
Instituto Superior Técnico (IST), Portugal
R. Alves Redol, 9, 1000 Lisboa Codex, Portugal
Phone: +351.1.3100315, Fax: +351.1.3145843
E-mails: jpn@inesc.pt cam@inesc.pt lba@inesc.pt

ABSTRACT

One of the problems of the speaker-independent continuous speech recognition systems is their inability to cope with the inter-speaker variability. When we find test speakers with different characteristics from the ones presented in the training pool we observe a large degradation on the system performance. To overcome this problem speaker-adaptation techniques may be used to provide near speaker-dependent accuracy. In this work we present a speaker-adaptation technique applied to a hybrid HMM-MLP system for large vocabulary, continuous speech recognition. This technique is based on an architecture that employs a trainable Linear Input Network (LIN) to map the speaker specific features input vectors to the speaker-independent system. This speaker-adaptation technique will be evaluated in an incremental speaker-adaptation task using the Wall Street Journal (WSJ) database. Both supervised and unsupervised modes are evaluated. The results show that speaker-adaptation within the hybrid framework can substantially improve system performance.

1. INTRODUCTION

Hybrid systems have been presented in the last two to three years as an alternative to Hidden Markov Models (HMMs) based systems for large vocabulary, speaker-independent, continuous speech recognition. This hybrid approach combines the HMM with connectionist models. The connectionist model acts as a phone probability estimator and is used as the observation model within the HMM framework. This hybrid HMM-connectionist system brings some benefits relative to HMM-only recognizers due mainly to the fact that strong assumptions about the input statistics and the functional form of the observation density are not required [1].

One of the problems of the speaker-independent continuous speech recognition systems is their inability to cope with the inter-speaker variability. These speaker-independent system are normally trained on large speech databases. Their speaker-independence cames from the use of a large pool of speakers. When we find test speakers with different characteristics from the ones presented in the training pool we observe a large degradation on the system performance. The problem is more extreme for fast and/or non-native speakers. This drawback is evidenced by the fact that speaker-dependent systems, typically have half the error rate of speaker-independent systems. However, the development of a speaker-dependent system for each talker is normally impractical. Large amounts of speech training data for each speaker may be unavailable or difficult to acquire. In these cases, speaker-adaptation algorithms – starting from a speaker-independent system and using a small amount of additional training data – may bridge the gap and provide near speaker-dependent accuracy. In classical HMM based systems different speaker-adaptation techniques have been used with sucess. Normally these techniques are based on the adaptation of the parameters of the speaker-independent system to maximize the likelihood of the adaptation data of the new speaker.

In this work we present a speaker-adaptation technique applied to a hybrid HMM-MLP system. This technique is based on an architecture that employs a trainable Linear Input Network (LIN) to map the speaker specific features input vectors (typically PLP cepstral coefficients) to the SI system. The LIN speaker-adaptation technique will be evaluated in an incremental speaker-adaptation task using the Wall Street Journal (WSJ) database. Both supervised and unsupervised modes are evaluated. The results show that speaker-adaptation within the hybrid framework can substantially improve system performance. The incremental unsupervised speaker-adaptation mode affords the possibility of incorporation in a real-time speaker-independent system without changing, from the user point of view, the way in which this system works.

2. THE BASIC HYBRID SYSTEM

In our work, we use a hybrid system where the connectionist architecture is based on a multilayer perceptron (MLP), with a single hidden layer and incorporating local acoustic context via a multi-frame input window [1]. This hybrid approach combines the temporal modeling capabilities of HMMs with the pattern classification capabilities of multilayer perceptrons. In this hybrid HMM-MLP system, a Markov process is used to model the basic temporal nature of the speech signal. The Markov process is determined in a hierarchical fashion. The language model is a Markov process on the words and the words are a Markov process on the sub-units used, in our case the phones. The MLP is used as the acoustic model within the HMM framework. The MLP estimates context-independent posterior phone probabilities to be used in the Markov process. This makes use of the fact that MLPs satisfying certain regularity conditions provide class probability estimates for given input patterns [2]. Decoding in the hybrid framework is equivalent to classical HMM decoding with the MLP modeling the observations.

In [3], this system and a similar one using a Recurrent Neural Network (RNN) were evaluated on the RM corpus, in speaker-independent mode. To see the last aplications and evaluations of a RNN hybrid system to large vocabulary see [4].

3. SPEAKER-ADAPTATION IN A HYBRID HMM-MLP SYSTEM

The speaker-adaptation technique presented here is based on an architecture that employs a trainable Linear Input Network (LIN) to map the speaker specific features input vectors (typically PLP cepstral coefficients) to the SI system.

In this technique, as represented in Figure 1, we create a linear mapping to transform the complete input vector (the current feature vector with some frames of left and right context). During recognition, this transformed vector is used as the input to the MLP component of the hybrid SI system (SI-MLP). To train the LIN for a new speaker, the weights of the mapping are initialized to an identity matrix. This guarantees that our initial point is the SI model. The input is propagated forward to the output layer of the SI-MLP. At that point, the error is calculated and propagated backward through the SI-MLP. As this system is "frozen", there is no weight adaptation of the SI-MLP. Adaptation is performed only in the weights of the linear input layer.

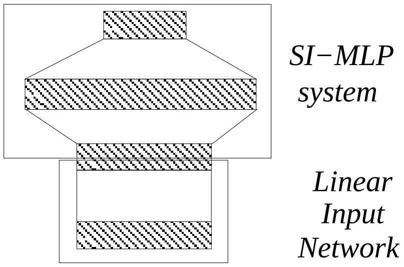

Figure 1: A schematic representation of the Linear Input Network (LIN).

This technique was first presented in [5] where evaluation and comparison of different architectures for speaker-adaptation in the context of hybrid HMM-MLP and HMM-RNN systems were made. Among the techniques presented the Linear Input Network showed to have a better performance when compared to several other altrenatives. These evaluation was made on the Resource Management (RM) corpus in a static supervised speaker-adaptation task. In [5] we showed that this technique achieved similar results as changing all the parameters of the MLP if enough adaptation data is available. However our goal is to adapt the system in a fast way to the new speaker using as few data as possible. In that sense the LIN technique yields better results. Our technique should not be seen just as a spectral mapping in the input feature parameters but as a transformation of the overall speaker-independent system, through the appending of new parameters, maximizing the likelihood of the adaptation data.

In [6] we extended this technique to a static unsupervised speaker-adaptation task on the RM corpus. In this context, by "unsupervised" we mean that there is no previous knowledge of the sentences being used for adaptation. The data used for adaptation are the first sentences pronounced by the speaker in his/her normal use of the system. On the contrary, in supervised adaptation there is prior knowledge of the initial sentences. This means that each speaker will have to go through an initial enrollment phase, in which he/she pronounces certain prescribed sentences. By "static" we mean that there is a separate training/adaptation set and a different test set where we evaluate the adapted system.

In [7] the LIN technique was further extended to comprises evaluation in an incremental unsupervised mode on the Wall Street Journal (WSJ) database. "Incremental" means that the system is only allowed to use any information it can extract from test data that it has already recognized. In this situation there are no separate training/adaptation and test sets. The adaptation procedure is incrementally applied over the test set itself.

In this work we had evaluated this technique in both supervised and unsupervised modes in an incremental speaker-adaptation task using the WSJ database. In the "incremental supervised" mode we can use the correct information after the fact. This means that we can use the correct sentence transcription but after the correspondent sentence recognition.

The incremental unsupervised speaker-adaptation procedure is as follows:

1. let $i = 1$
2. pick a group of T test sentences from the speaker
3. recognize these T test sentences with the SI system (this recognition will be the one used for the recognition score)
4. use the recognition of the $i.T$ sentences to make a Viterbi

alignment (this step generates the phone labels to assign to each frame of the $i.T$ sentences)

5. adapt the SI system as explained above (we use the backpropagation algorithm to minimize the classification error on the adaptation sentences)
6. let $i = i + 1$
7. pick a new group of T test adaptation sentences from the speaker; if there are no more sentences, stop.
8. recognize these new T sentences (this recognition will be the one used for the recognition score)
9. recognize the $(i - 1).T$ sentences (the sentences acumulated so far) with the current system
10. go to 4.

The incremental supervised speaker-adaptation procedure follows the same steps as in the incremental unsupervised mode except for step 4. which must be re-write as:

4. use the correct sentences transcription for the $i.T$ sentences to make a Viterbi alignment (this step generates the phone labels to assign to each frame of the $i.T$ sentences)

This adaptation procedure generates for each new group of sentences (in steps 3. and 8.) the final transcription of test sentences. These transcriptions will be used to get the final score.

The incremental unsupervised speaker-adaptation mode affords the possibility of incorporation in a real-time speaker-independent system without changing, from the user point of view, the way in which this system works (the system only needs to know when the speaker changes).

4. EXPERIMENTAL RESULTS

The Linear Input Network technique was first developed and evaluated on the DARPA Resource Management (RM) corpus. In the development phase we tested this technique in both supervised and unsupervised static modes [6, 7]. In the present work this technique was evaluated in an incremental speaker-adaptation task on the Wall Street Journal database. Both supervised and unsupervised modes were evaluated. The task chosen was the Nov. 94 Spoke 4 task. Next we will present the evaluation of the speaker-independent system and the results of the LIN speaker-adaptation technique.

4.1. Evaluation of the speaker-independent system

In previous works, with the RM database, we used a three layer full connected MLP with 1,000 hidden layer units. Our present network results from a scaled version of the RM network. Since the training data had increased aproximatelly in a factor of four (from the RM to the WSJ0-84) we had adjusted the hidden layer size by the same factor. Other point was the increase in the context window on the input of the MLP. Before we were using 7 frames (3 frames of left and right context around the central frame) and in this evaluation we are using a 9 frames window. Each acoustic vector is formed by PLP-12 cepstral coefficients and their first and second temporal derivatives. We use the delta and delta delta energy but not the energy itself. Therefore the feature vector has a total of 38 coefficents. Due to the frames of left and right context which are appended in the MLP input we have a total of 342 inputs. The resulting network has 4,000 hidden units and 61 output context-independent phone classes (about 1.6 million weights). Obviously the increase in the number of parameters results in an increase in the training time. In the work which is reported here we used the *Big Dumb Neural Network* (BDNN) from ICSI [8].

This system was evaluated on the WSJ0-93 Hub 2 test set, using a bi-gram language model and the LIMSI pronunciation lexicon. This speaker-independent system achieved in a 5K words task a result of 16.1% word errors.

4.2. Evaluation on the WSJ corpus

For the evaluation of the LIN speaker-adaptation technique on the WSJ we had chose a task that comprises incremental unsupervised speaker-adaptation (Nov. 94 Spoke 4 task). For this mode the system is only allowed to use any information it can extract from test data that it has already recognized. In this task is also possible to test incremental supervised speaker-adaptation. For this mode we can use the correct sentence transcription but after the correspondent sentence recognition.

Because this is an incremental task there are no separate training/adaptation and test sets. The adaptation and test process will be made with the same set of data. Both unsupervised and supervised process are described in Section 3.

This spoke has 4 speakers with about 100 sentences each. The system should being incrementally adapted to the new speaker and the results should be reported each 25 new sentences. In each evaluation point the results of the system with incremental unsupervised (IUSA) and incremental supervised speaker-adaptation (ISSA) enabled should be reported. Also for each evaluation point the results of the system with speaker-adaptation disabled should be reported. As described in Section 3. we adapt our system based in a group of 5 sentences (T=5). For the LIN speaker-adaptation technique the results are presented in Table 1 for the 4 speakers.

From the results we can observe a significant improvement on

		Sentences				Mean
		1-25	26-50	51-75	+76	
Sp.	SI	13.4	14.8	16.1	22.8	16.8
4TB	IUSA	12.2	11.2	16.6	17.5	14.4
	ISSA	12.0	11.5	12.3		11.9
Sp.	SI	19.7	27.8	18.7	23.3	22.5
4TC	IUSA	19.4	22.3	18.5	18.5	19.7
	ISSA	17.1	20.7	16.8	16.2	17.7
Sp.	SI	35.9	36.9	39.1	47.1	39.8
4TD	IUSA	34.9	31.1	33.0	35.5	33.5
	ISSA	30.7	25.9	24.8		27.2
Sp.	SI	11.5	15.4	14.7	11.2	13.2
4TE	IUSA	10.4	11.6	14.9	11.6	12.2
	ISSA	9.1	9.9	14.2		11.0

Table 1: Word error rate results for the LIN speaker-adaptation technique over the Nov. 94 Spoke 4 task. For each speaker the first row named “SI” presents the speaker-independent results (with speaker-adaptation disabled). The “IUSA” row presents the results with the Incremental Unsupervised Speaker-Adaptation enabled. The “ISSA” row presents the results with the Incremental Supervised Speaker-Adaptation enabled.

the system when the adaptation is enabled, proving the ability of this speaker-adaptation technique to cope with speaker differences. Also we see, as expected, a superior performance of the supervised speaker-adaptation mode.

In Table 2 we see the word error rate mean results for the four speakers.

	Sentences				Mean
	1-25	26-50	51-75	+76	
SI	20.1	23.7	22.1	26.1	23.1
IUSA	19.2	19.1	20.8	20.8	20.0
ISSA	17.2	17.0	17.0		17.1

Table 2: Word error rate mean results for the LIN speaker-adaptation technique over the Nov. 94 Spoke 4 task.

From the results in Table 2 we see an improvement of 10-12% for the incremental unsupervised speaker-adaptation task and an improvement of 20-25% for the incremental supervised speaker-adaptation.

5. CONCLUSIONS

A technique for speaker-adaptation of a hybrid HMM-MLP speaker-independent system was described and evaluated on the WSJ Nov. 94 Spoke 4. This technique was evaluated in both incremental supervised and unsupervised modes. The results show that speaker-adaptation within the hybrid framework can substantially improve system performance. In the incremental unsupervised mode, the improvement is obtained without any extra demands on the speaker, i.e. without an enrollment phase.

6. ACKNOWLEDGEMENTS

This work was partially funded by the Long Term Research RTD project 20077 SPRACH. An acknowledgment goes to ICSI in the person of Nelson Morgan for letting us use the BDNN system. We would like to acknowledge LIMSI-CNRS for providing the pronunciation lexica.

7. REFERENCES

1. H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994.
2. M. D. Richard and R. P. Lippmann, *Neural Network Classifiers Estimate Bayesian* a posteriori *Probabilities*, Neural Computation, vol. 3, pp. 461–483, 1991.
3. A.J. Robinson, L. Almeida, J.-M. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, N. Morgan, J.P. Neto, S. Renals, M. Saerens and C. Wooters, *A Neural Network Based, Speaker Independent, Large Vocabulary, Continuous Speech Recognition System: The* WERNICKE *Project*, Proceedings EUROSPEECH ’93, Berlin, pp. 1941–1944, 1993.
4. G. Cook, J. Christie, P. Clarkson, S. Cooper, M. Hochberg, D. Kershaw, B. Logan, S. Renals, A. Robinson, C. Seymour, S. Waterhouse, P. Zolfaghari, *Real-Time Recognition of Broadcast Radio Speech*, Proceedings ICASSP ’96, Atlanta, Vol.I, pp. 141-144, 1996.
5. J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals and T. Robinson, *Speaker-Adaptation For Hybrid HMM-ANN Continuous Speech Recognition System*, Proceedings EUROSPEECH ’95, Madrid, pp. 2171–2174, 1995.
6. J. Neto, C. Martins and L. Almeida, *Unsupervised Speaker-Adaptation For Hybrid HMM-ANN Continuous Speech Recognition System*, IEEE Signal Processing Society - 1995 Workshop on Automatic Speech Recognition, Snowbird, Utah, pp. 187–188, 1995.
7. J. Neto, C. Martins and L. Almeida, *Speaker-Adaptation in a Hybrid HMM-MLP Recognizer*, Proceedings ICASSP ’96, Atlanta, Vol. 6, pp.3383–3386, 1996.
8. N. Morgan and H. Bourlard, *Neural Networks for Statistical Recognition of Continuous Speech*, to be published in Proceedings of the IEEE.