# SPEAKER-ADAPTATION FOR HYBRID HMM-ANN CONTINUOUS SPEECH RECOGNITION SYSTEM

João Neto‡§ Luís Almeida‡§ Mike Hochberg* Ciro Martins§ Luís Nunes§
Steve Renals† Tony Robinson*

§*Instituto de Engenharia de Sistemas e Computadores (INESC), Portugal*
‡*Instituto Superior Técnico (IST), Portugal*
* *Cambridge University Engineering Department (CUED), UK*
†*Sheffield University (SU), UK*

## ABSTRACT

It is well known that recognition performance degrades significantly when moving from a speaker-dependent to a speaker-independent system. Traditional hidden Markov model (HMM) systems have successfully applied speaker-adaptation approaches to reduce this degradation. In this paper we present and evaluate some techniques for speaker-adaptation of a hybrid HMM-artificial neural network (ANN) continuous speech recognition system. These techniques are applied to a well trained, speaker-independent, hybrid HMM-ANN system and the recognizer parameters are adapted to a new speaker through off-line procedures. The techniques are evaluated on the DARPA RM corpus using varying amounts of adaptation material and different ANN architectures. The results show that speaker-adaptation within the hybrid framework can substantially improve system performance.

## 1. INTRODUCTION

Automatic speech recognition has been a major goal for a large research community in the last few years. The predominant approach to large-vocabulary, speaker-independent, continuous speech recognition (CSR) has been based on hidden Markov models (HMMs). Connectionist models have been widely proposed as a potentially powerful approach to speech recognition. Despite the very good results achieved in static pattern classification, there is not yet enough knowledge to adequately model the temporal structure of speech through connectionist models. To overcome these difficulties hybrid connectionist-HMM models have been built. The connectionist system acts as a phone probability estimator and is used as the observation model within the HMM framework. This approach brings some benefits. In particular, strong assumptions about the input statistics and the functional form of the output density are not required [1].

One of the major sources of error in speech recognition is inter-speaker variability—typically, speaker-dependent systems have half the error rate of speaker-independent systems. In many cases, however, the development of a speaker-dependent system for each talker is impractical. Large amounts of speech training data may be unavailable or difficult to acquire. In these cases, rapid 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.

This paper presents a study of different speaker-adaptation techniques applied to hybrid connectionist-HMM systems. The speaker-adaptation techniques are implemented off-line by transforming the input and/or weights of a speaker-independent connectionist component in our hybrid system. The starting point is a speaker-independent (SI) system [2]. The aim of this work is to find a set of transformations that adapt the SI system to a new speaker. The approach is evaluated on the DARPA Resource Management (RM) corpus [3] for varying amounts of adaptation material. The results show that speaker-adaptation within the hybrid framework can substantially improve system performance.

This paper is organized into 5 sections. In section 2., we will present our speaker-independent hybrid system. Section 3. presents the speaker-adaptation techniques and the associated results are found in section 4.. The final section presents some conclusions and future perspectives.

## 2. HYBRID ANN-HMM SYSTEM

Throughout this work, we use a hybrid HMM-ANN approach for continuous speech recognition. This hybrid approach (well explained in [1]) combines the temporal modeling structure of HMMs with the pattern classification capabilities of artificial neural networks. In this hybrid system, a Markov process is used to model the basic temporal nature of the speech signal. The connectionist structure is used to model the acoustic signal conditioned on the Markov process. This makes use of the result that connectionist networks satisfying certain regularity conditions provide class probability estimates for given input patterns [4, 5].

Here we present results for two connectionist architectures integrated in the hybrid approach:

1. a multilayer perceptron (MLP) [1], using a single hidden layer and incorporating local acoustic context via a multiframe input window; and

2. a recurrent neural network (RNN) [6] which uses a fully recurrent set of hidden state units to model acoustic context.

Both these systems estimate context-independent posterior phone probabilities. In [2], these systems were evaluated on the RM corpus and speaker-independent results were reported.

## 3. SPEAKER-ADAPTATION

This paper presents different techniques for speaker-adaptation in the context of continuous speech recognition with a hybrid HMM-ANN system. These techniques are applied off-line to a well-trained, speaker-independent, hybrid HMM-ANN system to adapt the recognizer parameters to a new speaker. In all the experiments, supervised adaptation was performed.

The basic approach used in these experiments was to adapt the connectionist component to a new speaker. This was performed by either modifying the parameters (weights) of the SI connectionist system and/or augmenting the structure of the SI connectionist system with additional, speaker-dependent architectures. Adaptation of the connectionist systems is accomplished in a similar fashion to training [1, 6]. Phone labels are assigned to each frame of the adaptation material using Viterbi alignment. Gradient descent on the connectionist parameters is used to minimize the classification error (via a cross-entropy criterion) on the adaptation sentences. Cross validation and early stopping are used to insure that the network parameters do not overtrain and generalize to new data.

The techniques investigated for speaker-adaptation include:

**LIN:** This technique employs a trainable Linear Input Network (LIN) to map the speaker-dependent input vectors (typically PLP cepstral coefficients) to the SI system. This mapping is trained by minimizing the error at the output of the connectionist system while keeping all other parameters fixed.

**RSI:** This technique is designated as Retrained Speaker-Independent (RSI) adaptation. Starting from the SI system, the full connectionist component is adapted to the new speaker. Early stopping of the training process is used to keep the system from over-fitting the training data.

**PHN:** In this technique, Parallel Hidden Network (PHN), additional, trainable hidden units were placed in the connectionist system. These extra units connect to inputs and outputs just like ordinary hidden units. During speaker-adaptation, weights connecting to/from these units were adapted while keeping all other parameters fixed.

**GAMMA:** In this approach, the speaker-dependent input vectors are mapped to the SI system (as in the LIN technique) using a gamma filter [8]. This structure provides a method for temporal as well as spectral adaptation.

These techniques were applied to a hybrid system based on a multilayer perceptron (MLP). In addition, RSI and LIN were also applied to a hybrid system based on a recurrent neural network (RNN).

### 3.1. Linear Input Network (LIN)

In this technique, we create a linear mapping to transform the complete input vector (the current feature vector with three frames of left and right context). During recognition, this transformed vector is used as the input to the SI-ANN. 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-ANN. At that point, the error is calculated and propagated backward through the SI-ANN. As this system is "frozen", there is no weight adaptation of the SI-ANN. Adaptation is performed only in the weights of the new linear input layer.

### 3.2. Retrained Speaker-Independent (RSI)

For this technique we start with a SI model and adapt it to a new speaker. In this case, the set of weights of the SI-ANN are the parameters to adapt. The back-propagation algorithm is used to adapt the weights of the SI-ANN to the new speaker. This is a difficult problem because there are a large number of free parameters with a small amount of adaptation data. Care is necessary to stop the training process before the system over-fits the adaptation data. The stopping criteria is determined by cross validation on independent data.

### 3.3. Parallel Hidden Network (PHN)

In this technique, a parallel network to the SI-MLP is created. This parallel network has the same input layer and the same output layer of the SI-MLP, but with a different hidden layer. This hidden layer is on the same level as the hidden layer of SI-MLP. In this new system, the input is propagated through the SI-MLP and through the parallel network. The output combines the connections from both hidden layers. The error evaluated on the output is just backward propagated through the parallel hidden layer. That means that the SI-MLP is "frozen" and we just adapt the weights that connect the parallel hidden layer to the SI-MLP input and output layer. We try to compensate for the differences between the SI system and the new speaker system through the weights. The parameters of the SI system are not changed, but additional parameters are created and trained to provide speaker-dependent information.

### 3.4. Gamma Network (GAMMA)

The GAMMA network may be regarded as a generalization of the baseline MLP system. Rather than using a multi-frame input (which may be regarded as a delay line) it is possible to use a generalized delay line, or gamma filter [8], with a set of trainable filter coefficients. The motivation for this approach is to allow the gamma filter coefficients to succinctly model (speaker specific) aspects of acoustic context, such as speaking rate. This is described in more detail in [7].

## 4. RESULTS

### 4.1. Data

The DARPA RM 1 corpus [3] was used in this study. This database of read speech contains a speaker-independent portion and a speaker-dependent portion. In the speaker-dependent part there are 12 speakers. For each speaker there is a training set (600 sentences), a development test set (100 sentences) and an evaluation test set (100 sentences). All results are reported on the evaluation test set.

### 4.2. Speaker-Independent System results

The baseline speaker-independent (SI) recognizer used in this study was a hybrid HMM-ANN system using an MLP or RNN probability estimator and a

| ANN | Test Set | Sub | Del | Ins | Err |
|---|---|---|---|---|---|
| MLP | RM-feb89 | 3.5% | 1.4% | 0.2% | 5.1% |
| MLP | RM-oct89 | 4.0% | 1.7% | 0.3% | 6.1% |
| MLP | RM-feb91 | 3.7% | 1.4% | 0.5% | 5.7% |
| MLP | RM-sep92 | 7.6% | 3.1% | 1.7% | 12.3% |
| RNN | RM-feb89 | 4.3% | 1.8% | 0.9% | 7.0% |
| RNN | RM-oct89 | 4.5% | 1.9% | 0.9% | 7.2% |
| RNN | RM-feb91 | 4.4% | 1.5% | 1.3% | 7.3% |
| RNN | RM-sep92 | 7.8% | 2.6% | 1.9% | 12.3% |

Table 1. Speaker-Independent system performance on standard SI test sets.

set of single state context-independent phone models with a pseudo-Poisson duration model. Decoding was performed using the Viterbi criterion.

The front end to the ANN used a 12th order perceptual linear prediction (PLP) analysis to produce 12 PLP cepstral coefficients plus energy, for each 20ms frame of speech using a 10ms frame step. For the MLP, the temporal derivatives of these coefficients were estimated using linear regression over nine frames, to give a 26 coefficient feature vector. Three frames of left and right context were appended at the MLP input, giving a total of 182 inputs and the resulting network used 1,000 hidden units and 68 output phone classes (about 250,000 weights). Because the RNN implicitly models temporal context, only the 12 PLP cepstral coefficients plus energy are used as input. The RNN system employed 256 state units and 68 outputs with approximately 90,000 parameters[1].

Table 1 shows the baseline SI performance over the four standard SI test sets. The column labeled SI in tables 2 and 3 show the SI performance for each speaker in the speaker-dependent portion of the corpus for the MLP and RNN systems, respectively.

### 4.3. Speaker-Dependent Results

For a correct evaluation of the results obtained in the speaker-adaptation process, it was important to have a comparison point. We trained a set of speaker-dependent (SD) MLP models for each speaker in the SD corpus from RM 1 database (see Table 2 - column SD). These SD models used the same approach as for the SI system. The same front end with PLP analysis and the same phone set were used. The number of hidden units in this case were 350. The set of SD models were trained from scratch (initial random weights).

### 4.4. Speaker-Adaptation Results

The different techniques investigated for speaker-adaptation were trained and evaluated with the speaker-dependent (SD) part of the Resource Management 1 corpus. We split this SD data in two sets: one as adaptation data with 700 sentences (resulting from the training and development test data of the SD RM 1) and the other for test data with 100 sentences (resulting from the evaluation test data of the SD RM 1). All the results in Tables 2 through 4 were evaluated over the same test data (100 sentences).

Table 2 shows the MLP results for the RSI, PHN, and LIN adaptation techniques as well as the SI and SD results. The adaptation process used 80 sentences as adaptation data and 20 sentences as cross-validation data. From the results we can observe the improvement resulting from the application of speaker-adaptation techniques. The best results came from the LIN with a reduction of 35.3% on the word error compared to the SI model. Note, however, the performance has still not reached the level of a speaker-dependent system.

| Speaker | SI | RSI | PHN | LIN | SD |
|---|---|---|---|---|---|
| BEF0 | 7.9% | 8.2% | 8.2% | 7.1% | 4.6% |
| CMR0 | 15.5% | 10.1% | 12.0% | 8.7% | 4.7% |
| DAS1 | 6.7% | 2.7% | 4.0% | 2.9% | 2.6% |
| DMS0 | 6.1% | 4.9% | 5.1% | 4.5% | 2.5% |
| DTB0 | 7.3% | 5.8% | 6.8% | 4.8% | 3.8% |
| DTD0 | 8.1% | 6.1% | 10.9% | 4.8% | 4.3% |
| ERS0 | 8.7% | 7.3% | 8.3% | 7.0% | 5.9% |
| HXS0 | 10.9% | 6.4% | 10.5% | 4.9% | 3.2% |
| JWS0 | 5.3% | 2.7% | 4.9% | 4.6% | 2.0% |
| PGH0 | 6.9% | 4.9% | 5.8% | 3.3% | 3.6% |
| RKM0 | 14.6% | 11.2% | 12.2% | 9.6% | 6.4% |
| TAB0 | 4.1% | 3.9% | 3.1% | 3.9% | 3.5% |
| Mean | 8.5% | 6.2% | 7.7% | 5.5% | 3.9% |

Table 2. MLP word error percentage results for the different Speaker-Adaptation techniques under study and results for the SI and SD systems. Speaker-adaptation was performed using 100 adaptation sentences (80 training and 20 cross-validation).

One characteristic of the hybrid approach has been the normalization of the MLP inputs. This process normalizes each input channel to have zero mean and unit variance. The appropriate normalization transformation is usually determined during training and fixed for testing. We have found that the speaker-adaptation performance can be greatly enhanced by estimating a new input normalization transformation from the speaker-adaptation training data and using that same transformation for testing. In Table 2 we used 80 sentences for adaptation. The normalization transformation for each speaker was estimated over their 80 adaptation sentences.

Table 3 shows the RNN results for the RSI and LIN adaptation techniques as well as the SI system. The adaptation process used 80 sentences as adaptation data and 20 sentences as cross-validation data. The results show that the RSI approach does improve the system, although not as significantly as for the MLP. Also like the MLP system, the LIN approach does a better job of adapting to the new speaker than the RSI approach. This is certainly due to the limited amount of adaptation data. The RSI approach must adapt all 90,000 RNN parameters and can very easily over-fit the data. The LIN approach, while only adapting 169 parameters, is able to capture speaker-dependent information. Another benefit of the LIN approach was that the computational requirement was less than half of that required for the RSI technique.

To further evaluate the properties of LIN, the speaker-dependent LINs developed with the MLP system were used to compensate the the data used in the RNN system. The goal of this experiment was to determine if the LIN adaptation compensation was

[1] In the interest of consistency with the MLP system, a non-optimal configuration was used for the RNN. See [2] for more state-of-the-art performance values with the RNN system

| Speaker | SI | RSI | LIN |
|---|---|---|---|
| BEF0 | 9.4% | 7.6% | 8.3% |
| CMR0 | 9.5% | 8.4% | 8.1% |
| DAS1 | 8.1% | 4.5% | 3.9% |
| DMS0 | 6.8% | 6.4% | 5.0% |
| DTB0 | 7.9% | 6.0% | 5.5% |
| DTD0 | 10.9% | 9.0% | 7.1% |
| ERS0 | 7.3% | 6.6% | 6.8% |
| HXS0 | 10.5% | 8.7% | 5.5% |
| JWS0 | 6.3% | 4.8% | 5.8% |
| PGH0 | 5.3% | 4.2% | 4.0% |
| RKM0 | 12.0% | 10.9% | 11.2% |
| TAB0 | 6.5% | 6.2% | 5.1% |
| Mean | 8.4% | 6.9% | 6.4% |

Table 3. RNN word error percentage results for the different Speaker-Adaptation techniques. Speaker-adaptation was performed using 100 adaptation sentences (80 training and 20 cross-validation).

(more) a function of the spectral characteristics or the connectionist architecture. The error rates achieved in this case were substantially greater (over 50%) than the SI performance, indicating that the LIN is architecture dependent. Further investigation of this is planned.

To study the influence of the quantity of adaptation data we tested the different speaker-adaptation techniques with a variable number of adaptation sentences. The results are presented in Table 4 where on the first column we present the number of adaptation sentences. We began with 30 sentences for adaptation training and 10 sentences for cross-validation. In the next step we switched to 80 sentences for adaptation training and 20 for cross validation which is the situation presented in Table 2. The next step was to use 100 sentences for training and 100 sentences for validation. The objective was to make the cross-validation process more accurate but there was just a small decrease of the word error comparing with the case of 80 adaptation sentences.

The last line of Table 4 shows a case where we use all the available data for adaptation training. This is equivalent to our SD training and for the RSI and LIN techniques the result is the same. The difference is in the adaptation process beginning with a well trained SI model and on the SD training starting from scratch. In terms of computing time, the adaptation process is fast because it needs a shorter number of iterations of the backpropagation algorithm.

Experiments using the GAMMA approach were reported in [7]. The baseline GAMMA system had a poorer recognition performance compared with the baseline MLP system, and speaker-adaptation performed by re-estimating the gamma filter coefficients was not successful, with no consistent decrease in error rate.

## 5. CONCLUSION

Different techniques for speaker-adaptation of a hybrid HMM-ANN speaker-independent system were described and evaluated on the DARPA RM corpus. The results show that speaker-adaptation within the hybrid framework can substantially improve system performance. It was found that adaptation of the MLP resulted in greater reductions in error rate than that of the RNN. One possible factor for this is that the RNN is more difficult to train than the MLP and a better optimization scheme could lead to improved results.

| # Sentences | RSI | PHN | LIN |
|---|---|---|---|
| 30+10 | 7.3% | 8.2% | 6.4% |
| 80+20 | 6.2% | 7.7% | 5.5% |
| 100+100 | 5.8% | 6.8% | 5.3% |
| 600+100 | 3.9% | 4.1% | 3.9% |

Table 4. MLP Speaker-Adaptation results for different amounts of adaptation data.

In the future, we plan to extend this work to on-line and unsupervised procedures and to evaluate these techniques on large vocabulary tasks such as the Wall Street Journal corpus. It is also possible that these techniques may generalize to adaptation to different acoustical environments. Experiments are planned to apply the LIN approach to microphone adaptation.

## 6. ACKNOWLEDGEMENTS

This work was funded by ESPRIT project 6487 Wernicke.

## REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994.

[2] 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, pp. 1941-1944, 1993.

[3] P. Price, W. M. Fisher, J. Bernstein and D. S. Pallett, *The DARPA 1,000-Word Resource Management Database for Continuous Speech Recognition*, Proceedings ICASSP 1988, p. 651-654, 1988.

[4] M. D. Richard and R. P. Lippmann, *Neural Network Classifiers Estimate Bayesian* a posteriori *Probabilities*, Neural Computation, vol. 3, pp. 461–483, 1991.

[5] S. Santini and A. Del Bimbo, *Recurrent Neural Networks can be Trained to be Maximum A Posteriori Probability Classifiers*, Neural Networks, vol. 8, no. 1, pp. 25–29, 1995.

[6] T. Robinson, *An application of recurrent nets to phone probability estimation*, IEEE Trans. on Neural Networks, vol. 5, no. 2, pp. 298-305, 1994.

[7] S. Renals and M. Hochberg, *Using Gamma Filters to Model Temporal Dependencies in Speech*, Proceedings of ICSLP-94, pp.1491–1494, 1994.

[8] B. de Vries and J. C. Principe, *The Gamma Model—a New Neural Model for Temporal Processing*, Neural Networks, vol. 5, pp. 565–576, 1992.