COMBINATION OF ACOUSTIC MODELS IN CONTINUOUS SPEECH RECOGNITION HYBRID SYSTEMS

Hugo Meinedo *João P. Neto*

INESC - IST
Rua Alves Redol, 9 1000-029 Lisboa - Portugal
Hugo.Meinedo@inesc.pt, jpn@inesc.pt
http://neural.inesc.pt/NN/RFC

ABSTRACT

The combination of multiple sources of information has been an attractive approach in different areas. That is the case of speech recognition area where several combination methods have been presented. Our hybrid MLP/HMM systems use acoustic models based on different set of features and different MLP classifier structures. In this work we developed a method combining phoneme probabilities generated by the different acoustic models trained on distinct feature extraction processes. Two different algorithms were implemented for combining the acoustic models probabilities. The first covers the combination in the probability domain and the second one in the log-probability domain. We made combinations of two and three alternative baseline systems where was possible to obtain relative improvements on word error rate larger than 20% for a large vocabulary speaker independent continuous speech recognition task.

1. INTRODUCTION

For some time now researchers have been working in the combination of different classifiers following the "divide and conquer" thesis as a way to improve overall classifier performance. In fact, the combination of multiple sources of information is an attractive approach to speech recognition where frequently there is a high number of classes and the input is affected by noise or other sources of variability resulting in hard to classify data. In speech recognition, there are at least three potential levels where combination may be introduced: i) we can combine the probability estimates of several acoustic models before they are fed into the decoder [1, 2], ii) it is possible to combine several language models estimated on different corpora by interpolating their probabilities estimates and iii) it is also possible to run several independent recognisers in parallel and combine their output word hypotheses based on voting [3].

In our work we have been developing hybrid systems that combine the temporal modelling capabilities of hidden Markov models (HMMs) with the pattern classification capabilities of multilayer perceptrons (MLPs). In this hybrid HMM/MLP system [4], a Markov process is used to model the basic temporal nature of the speech signal. 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. We developed hybrid MLP/HMM context independent systems for continuous speech recognition using a different set of databases and applied to distinct tasks [5, 6].

We noted that some of our baseline recognition systems that use acoustic models with different properties gave emphasis to distinct portions of the speech signal and caused different responses to difficult situations. Through the observation of the output produced by these recognition systems for a common task we concluded that they tended to make errors with a low degree of correlation. This evidence suggested that an appropriate method for combining the baseline recognisers might produce a final system more accurate than either of the constituents alone.

In this work we developed a combination method of different context independent acoustic models through an average in the probability and in the log-probability domain of the phonetic probabilities estimated at the output of the classifiers. We combined simultaneously two and three baseline recognition systems and compared them based on the total number of parameters and word error rate (% WER).

On section 2 we present the combination algorithms used for the experiments. The baseline speech recognisers are described on section 3. On section 4 the evaluation of the combined recognition systems is summarised. Finally, some conclusions and future work are presented in sections 5 and 6.

2. COMBINATION OF ACOUSTIC MODELS

In this work we developed a method that combines phoneme probabilities generated by different acoustic

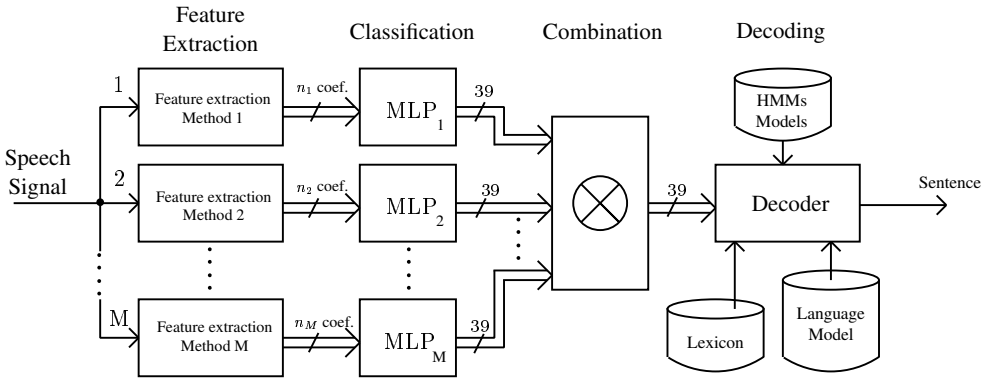

Figure 1: Combination of acoustic models used.

models trained on distinct feature extraction processes. These probabilities are taken at the output of each model's MLP classifier and combined using an appropriate algorithm. The processing stages are represented in Figure 1. All acoustic models use the same phoneme set constituted by 38 phonemes for the Portuguese language plus the silence. The combination algorithms merge together the probabilities associated to the same phoneme. If we use M acoustic models the new probability value for phoneme y will result from the merging operation of M probability values, each one resulting from a different acoustic model. Two different algorithms were implemented for combining the acoustic models probabilities. The first one covers the combination in the probability domain and the second one in the log-probability domain.

2.1. Combining in the probability domain

The first algorithm involves an average in the probability domain for the new *a posterior* probability value of phoneme y given the acoustic vector $\mathbf{x}$, represented by $\hat{P}(y|\mathbf{x})$. This value is obtained by averaging the probability for that particular phoneme calculated by each one of the M acoustic models that are being combined.

$$\hat{P}(y|\mathbf{x}) = \frac{1}{M}\sum_{k=1}^{M} P_k(y|\mathbf{x}) \tag{1}$$

The *a priori* probability value, $P(y)$, used internally by the decoder to calculate the scaled likelihood, as usual in a hybrid system [4], is not affected by the average operation. This value is obtained by estimating the relative frequency of phoneme y in the training set and thus is equal for all acoustic models.

2.2. Combining in the log-probability domain

The second algorithm multiplies the probability values, which internally to the decoder corresponds to perform an average in the log-probability domain. The new *a posteriori* probability value for phoneme y given the acoustic vector $\mathbf{x}$ is defined as,

$$\hat{P}(y|\mathbf{x}) = \prod_{k=1}^{M} P_k(y|\mathbf{x}) \tag{2}$$

In this case the *a priori* probability value is obtained by raising the original value to the M^{th} power.

$$\hat{P}(y) = \prod_{k=1}^{M} P_k(y) = [P(y)]^M \tag{3}$$

The scaled likelihood is calculated as,

$$\frac{\hat{P}(y|\mathbf{x})}{\hat{P}(y)} = \frac{\prod_{k=1}^{M} P_k(y|\mathbf{x})}{\prod_{k=1}^{M} P_k(y)} = \frac{P_1(y|\mathbf{x})}{P(y)} \cdots \frac{P_M(y|\mathbf{x})}{P(y)} \tag{4}$$

In order to obtain exactly the average in the log-probability domain it was necessary to adjust the acoustic scale factor, ac, that internally to the decoder appears multiplied by the log scaled likelihood. The log-probability values are M times bigger due to the merging of M probabilities and so the new acoustic scale has to be reduced by an equal factor, $\widehat{ac} = \frac{ac}{M}$ in order to maintain the scaled likelihood's magnitude. Using equation (4) and transforming the multiplication into a sum of logs we have for the log scaled

likelihood,

$$\widehat{ac}\ \log\left[\frac{\hat{P}(y|\mathbf{x})}{\hat{P}(y)}\right] = \frac{ac}{M}\sum_{k=1}^{M}\log\left[\frac{P_k(y|\mathbf{x})}{P(y)}\right] \quad (5)$$

3. BASELINE SYSTEMS

In our previous work we have developed several context independent continuous speech recognition systems based on hybrid MLP/HMM architectures [5, 6]. For the present work we chose some baseline recognisers that have been trained using our Portuguese continuous speech database, the BD-PUBLICO [7]. The main difference between these systems lays in the acoustic modelling. They use different feature extraction methods and MLPs with different structures. The PLP (or Log-RASTA) baseline recognition systems uses log-energy and PLP (or Log-RASTA) 12^{th} order cepstral coefficients and their first temporal derivatives summing up to 26 parameters per frame. The ModSpec baseline recognition system was developed using Modulation Spectrogram features [8]. This feature extraction process outputs 28 coefficients in each frame. The PLP+Sil and Log-RASTA+Sil use an additional coefficient in each frame provided by a syllable segmentation system which acts as an alternative feature extraction method [9]. The MLP classifier uses 7 input frames, 500 or 1000 units in the hidden layer and 39 output units. They were all evaluated using a development test set from the BD-PUBLICO database and applying a 5k or 27k vocabulary set. The evaluations were based on the word error rate percentage (% WER). In Table 1 we summarise the performance obtained by these baseline systems.

System	5k vocabulary % WER	27k vocabulary % WER
ModSpec	16.2	17.7
PLP	14.2	15.8
Log-RASTA	13.7	15.7
PLP+Sil	13.3	14.7
Log-RASTA+Sil	12.3	14.2
PLP 1000	12.4	14.1

Table 1: Evaluation results for several developed baseline recognition systems. All of these systems use 500 hidden units in the MLP, except the last one that has 1000 units.

4. COMBINATION EXPERIMENTS

A set of experiments was performed making some combinations using the baseline systems with 500 hidden units in the MLP and without mixing together similar feature extraction processes. By combining two different systems having 500 hidden units it is possible to make a direct comparison with the PLP 1000 because in the end both recognition systems will have approximately the same number of parameters. The results obtained are summarised in Table 2.

System	5k vocabulary % WER	27k vocabulary % WER
<PLP, ModSpec>	12.9	14.2
<Log-RASTA, ModSpec>	12.4	14.0
<PLP, Log-RASTA>	11.3	13.1
PLP x ModSpec	11.2	13.3
Log-RASTA x ModSpec	11.8	13.2
PLP x Log-RASTA	10.9	12.8
PLP+sil x Log-RASTA	10.9	12.6
PLP x Log-RASTA+sil	10.8	12.3
PLP+sil x Log-RASTA+sil	10.7	12.6

Table 2: Combinations of two baseline systems with 500 hidden units. <A, B> represents the average in probability domain while A x B indicates the combination using the multiplication of probabilities algorithm (average in log-probability domain).

In the first three cases, where the average in the probability domain was used, we obtained more than 10% improvements in word error rate when comparing the combined systems with the constituents alone. In the middle of Table 2 we see the results for the combinations using the multiplication of phoneme probabilities. This algorithm achieved better results than the average in the probability domain. If we now compare these results with the PLP 1000 recogniser we see in some cases a significant reduction in word error rate. In the last part of Table 2 we show the results for the combination using the acoustic models that take advantage of syllable information. In these cases we obtained a very significant 12% relative reduction in word error rate when compared to the baseline PLP 1000 system.

In the following experiments we evaluated if, in respect to the combination of two acoustic models, there could still be a gain by merging an additional third baseline system. In Table 3 the most representative results are summarised.

System	5k vocabulary % WER	27k vocabulary % WER
PLP x Log-RASTA x ModSpec	10.9	12.4
PLP+sil x Log-RASTA+sil x ModSpec	10.4	11.7
PLP 1000 x Log-RASTA+sil x ModSpec	10.2	11.4

Table 3: Combinations of three baseline systems.

In the first case the three simplest baseline acoustic models

were used. By combining three acoustic models we were able to reduce the WER by 22% relative to the best constituent. In the second case we added the ModSpec acoustic model to one of the combinations of two systems that had obtained the best results. This case represents a 16% relative reduction when compared with the best individual system. Finally, in an attempt to obtain the lowest possible word error rate, we used the PLP 1000 instead of the PLP+Sil because it has better performance due to a MLP with more parameters. We see only a small reduction when comparing with the second case. Compared with the best constituent, the PLP 1000, we got a 15% relative improvement.

We are currently developing recognition systems for more difficult tasks. One such case is broadcast news transcription. Some preliminary experiments were performed using these baseline recognition systems whose acoustic models were only trained with clean read speech. Obviously this is a very demanding task furthermore complicated by frequent OOV words even for a 27k vocabulary. In these experiments we have observed that in situations involving pivot speech the relative improvement obtained in the word error rate by the combined recognition system is slightly above 20%. For situations involving spontaneous speech the improvement approaches 30% in respect to our best baseline system.

5. CONCLUSIONS

In this paper we presented the development of a new large vocabulary continuous speech recognition system for the Portuguese language based on the combination of different acoustic models. These models were merged through two different algorithms combining the phonemes *a posteriori* probabilities, estimated by the several acoustic models MLP's, in the probability and in the log-probability domain.

Starting with absolute word error rates around 14% from the isolated baseline systems we were able to achieve values around 10% which represents a relative improvement larger than 20%. Furthermore the combined systems proved to be even more robust when used in adverse situations that significantly differ from the training ones.

6. FUTURE WORK

We will continue to investigate and develop new ways of improving our recognition systems. One technique that will deserve our attention will be the combination at word level using voting algorithms. Other area of interest will be the development of acoustic models more robust to noise and spontaneous speech in order to cope with more difficult situations. The combined systems described in this paper were obtained by using acoustic models trained only with clean read speech to use in dictation tasks, and by that consequence are not the most suited for the type of situations imposed by broadcast news transcription tasks.

7. ACKNOWLEDGEMENTS

This work was partially funded by the PRAXIS XXI project 1654. An acknowledgement is also given to the PÚBLICO newspaper for making its texts available for this work.

8. REFERENCES

1. Hochberg, M., Cook, G., Renals, S., and Robinson, T., *Connectionists model combination for large vocabulary speech recognition*, in Proceedings IEEE Workshop on Neural Networks for Signal Processing, 1994.

2. Cook, G. and Robinson, A., *Boosting the performance of connectionist large vocabulary speech recognition*, in Proceedings ICSLP 96, Philadelphia, USA, 1996.

3. Fiscus, J., *A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover)*, in Proceedings IEEE Workshop on Automatic Speech Recognition and understanding (ASRU 97), Santa Barbara, USA, 1997.

4. Bourlard, H. and Morgan, N., *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, Massachusetts, EUA, 1994.

5. Neto, J., Martins, C. and Almeida, L., *The Development of a Speaker Independent Continuous Speech recogniser for Portuguese*, in Proceedings EUROSPEECH 97, Rhodes, Greece, 1997.

6. Neto, J., Martins, C. and Almeida, L., *A Large Vocabulary Continuous Speech Recognition Hybrid System for the Portuguese Language*, in Proceedings ICSLP 98, Sydney, Australia, 1998.

7. Neto, J., Martins, C. Meinedo, H. and Almeida, L., *The Design of a Large Vocabulary Speech Corpus for Portuguese*, in Proceedings EUROSPEECH 97, Rhodes, Greece, 1997.

8. Kingsbury, B. E., Morgan, N., and Greenberg, S., *Robust speech recognition using the modulation spectrogram*, Speech Comunication, 25:117–132, 1998.

9. Meinedo, H. and Neto, J., *The use of syllable segmentation information in continuous speech recognition hybrid systems applied to the Portuguese language*, in Proceedings ICSLP 2000, Beijing, China, 2000.