interspeech 2008 incorporating SST 08

Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer

Alberto Abad and João Neto

L^2F - Spoken Language Systems Lab
INESC-ID / IST, Lisboa, Portugal
{Alberto.Abad,Joao.Neto}@l2f.inesc-id.pt

Abstract

Speech recognition based on connectionist approaches is one of the most successful alternatives to widespread Gaussian systems. One of the main claims against hybrid recognizers is the increased complexity for context-dependent phone modelling, which is a key aspect in medium to large size vocabulary tasks. In this paper, a baseline hybrid system based on monophone recognition units is improved by incorporating acoustical modelling of phone transitions. First, a single state monophone model is extended to multiple state sub-phoneme modelling. Then, a reduced set of diphone recognition units is incorporated to model phone transitions. The proposed approach shows a 26.8% and 23.8% relative word error rate reduction compared to baseline hybrid system in two selected WSJ evaluation test sets. Additionally, improved performance compared to a reference Gaussian system based on word-internal context-dependent triphones and comparable results to cross-word triphone system are reported.

Index Terms: speech recognition, context modelling, connectionist system

1. Introduction

Decades of research and advances in speech and language technology, and more concretely, in automatic speech recognition (ASR) have made possible the development of successful very large vocabulary continuous speech recognition systems in certain constrained conditions.

Regarding the different ASR paradigms proposed during these years, Hidden Markov Models of Gaussian mixtures (HMM/GMM) [1] is doubtless the most widely accepted approach. Alternatively, Artificial Neural Networks (ANN) based framework has also been proposed [2], but despite their high discrimination ability in short-time classification tasks, they have proved inefficient when dealing with long-term speech segments. With the scope of solving the problem of long time modelling of the ANN framework, one of the most successful alternatives to HMM/GMM was later proposed, commonly known as hybrid ANN/HMM or connectionist paradigm [3]. In general, hybrid architectures seek to integrate ANN ability for estimation of Bayesian posterior probabilities into a classical HMM structure that permit modelling long-term speech evolution.

On the one hand, the main advantage of hybrid ANN/HMM are that classification networks are usually considered better pattern classifiers than Gaussian mixtures approaches. Moreover, they are usually more efficient in terms of computational decoding cost. Additionally, an appealing characteristic of the hybrid systems is that they are very flexible in terms of merging multiple input streams.

On the other hand, one of the most significant limitations of hybrid systems is related with the lack of flexibility and increased difficulty when context-dependent phone modelling is desired. Most notable approaches to phonetic context training in ANN are reported in [4, 5] based on factorization of posterior probabilities. In these works, it is shown that a significant performace improvement is achieved in exchange for a more complex architecture that affects training (multiple networks needed augmenting number of parameters) and decoding (increase of explored hypothesis and computational load).

The aim of the present work is to provide an insight on the recent advances carried out to improve the performance of our connectionist speech recognition system based on MultiLayer Perceptron (MLP) classification networks – named AUDIMUS [6] – and to propose an alternative flexible approach for incorporating context phone modelling information in that kind of hybrid systems.

First, conventional single state phone model has been extended to multiple state sub-phoneme recognition units. That is, each phoneme is split into three regions which are represented by independent classes in the MLP network. This approach was already reported in [7] and it is a necessary step for posterior phonetic context modelling. Additionally, we provide some comments on practical issues related to the generation of the initial state-level alignment.

Second, context-dependent modelling is tackled by means of phone transitions modelling. Concretely, a reduced representative set of diphone units is incorporated into the set of sub-phoneme recognition units of the multiple state hybrid system. That is, the MLP network output layer is augmented with a fixed number of classes representing the most frequent phone transitions that appear in the training corpora. The proposed approach achieves better performance than a reference word internal triphone HMM/GMM system and comparable performance to cross-word triphone system, without affecting the architecture of the baseline hybrid recognition system.

2. Corpora and task description

The Wall Street Journal (WSJ) database [8] is used throughout this work for acoustic model development and training. It is a US English native speakers database that contains high-fidelity speech recordings with excerpts from the Wall Street Journal. Only the SI-84 training material from WSJ0 was used, resulting in approximately 15 hours of speech material.

The November 1992 ARPA WSJ evaluation corpora

This work was funded by PRIME National Project TECNOVOZ number 03/165.

(*Nov92*) containing 330 sentences from 8 speakers is used as development test data, that is, the various systems are tuned on this set. The *si_dt_s6* data from the WSJ1 (202 sentences from 8 speakers) and the *si_dt_05.odd* subset of the WSJ1 (248 sentences from 10 speakers) defined in [9] are used as evaluation test sets.

The systems assessed in this work have been evaluated with the WSJ 5K non-verbalized 5k closed vocabulary set and the WSJ standard 5K non-verbalized closed bigram language model. The language model and the acoustic model weighting is tuned to provide optimum performance in the development data set. The same language model, vocabulary and development and evaluation test sets are kept in all the following experiments both in the HMM/GMM system and in the hybrid ANN/HMM recognizer.

3. HMM/GMM reference system

The reference recognizer is the HMM/GMM system based on context dependent triphones trained with HTK toolkit described in [9]. First, a set of monophone models is trained up with manually annotated data of the TIMIT database to later do forced alignment of the WSJ0 training corpora. Then, training of monophones, triphones, state-tying and mixing-up the number of Gaussians is done until final tied-state context-dependent triphones of 3 states and 8 Gaussians per state are built producing a system of about 2.4 million parameters. The front-end of the HMM reference system consists of a single stream of 39 element feature vectors composed by 13 mel-cepstrum components (included coefficient 0) and its first and second derivatives. Cepstral mean normalization (CMN) is also applied.

Table 1 shows the reference results after [9] depending wether word-internal triphones (*wint*) or cross-word (*xword*) triphones are considered. Notice that the main objective of this work is to tackle some of the limitations of hybrid systems, rather than providing a comparison between both modelling paradigms (ANN an GMM). Thus, some state of the art approaches for Gaussian based systems, such as discriminative training, were not considered. Nevertheless, this reference system sets the context for the experiments of the next sections.

4. Hybrid ANN/HMM speech recognition

4.1. The AUDIMUS speech recognition system

Figure 1 shows a block diagram of the AUDIMUS speech recognizer. It is based on the hybrid ANN/HMM paradigm for speech recognition [3]. This kind of recognizers are generally composed by a phoneme classification network, particularly a MultiLayer Perceptron (MLP), that estimates the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated to the single state of context independent phoneme hidden Markov models.

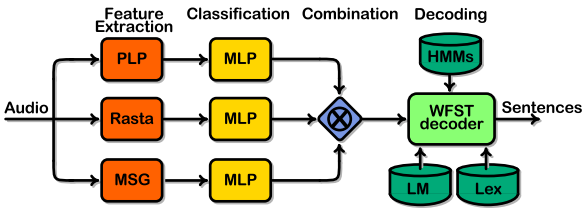

Figure 1: *Block diagram of the AUDIMUS speech recognizer.*

Concretely, the system combines three MLP outputs trained with Perceptual Linear Prediction features (13 static + first derivative), log-RelAtive SpecTrAl features (13 static + first derivative) and Modulation SpectroGram features (28 static). In addition to the feature representation, MLP networks are mainly characterized by the size of their hidden layer(s) and the size of the output layer. In [6], a more detailed description of the AUDIMUS recognizer can be found for the broadcast news transcription task of European Portuguese. The decoder of the recognizer is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [10].

4.2. The hybrid baseline system

As in the case of the HMM/GMM system, the TIMIT database with manually annotated phonetic alignments is first used to train a MLP that allows the generation of frame-to-phone alignments of the WSJ0 data using word-level transcriptions.

The WSJ0 data is split into training (TR) and cross-validation (CV) sets. The CV data ($\sim$ 10%) is used to define the stopping criteria in the updating process of the MLPs as it is usually done in this kind of approaches.

The process for the estimation of the final phone classification networks consists of several iterations of re-alignment and re-training until a stable performance is obtained in the development test set (*Nov92*).

In this work, the MLPs trained are composed of an input layer with a context window of 7 frames, 2-hidden layers of 700 weights each one and 40 units output layer (40 phonemes including the silence pattern). The resulting networks are relatively large for the available amount of data ($\sim$650K parameters and $\sim$7.5 patterns per weight), however it has been experimentally found to be the most addequate size.

Table 1 shows the performance of ANN/HMM multiple stream system. It can be observed that the hybrid system performs below both the word-internal triphone system and the cross-word recognizer, despite multiple stream decoding.

System	Nov92	si_dt_s6	si_dt_05.odd
HMM/GMM wint	8.11	10.39	12.40
HMM/GMM xword	6.86	9.52	10.48
ANN/HMM	9.73	13.13	14.37

Table 1: *WER results of HMM/GMM systems after [9] (word internal and cross-word) and of the baseline ANN/HMM system.*

5. Multiple-state hybrid system extension

In this section the baseline ANN/HMM speech recognizer based on single state HMM models (1 network output per phoneme) is extended to multiple state modelling. The underlying idea that justifies this extension comes from the characteristics of the phone production process. Each phone is usually considered to be constituted by three regions or portions: an initial transitional region, a second central steady region known as phone nucleus, and a final transitional region. Thus, it is expected that modelling each one of these portions independently will produce an improvement of the acoustic phone modelling and consequently and improvement in the recognition performance.

The advantage of this approach is its simplicity, since the architecture of the hybrid ANN/HMM speech recognizer is not

altered. The drawback is an increase of the number of outputs and consequently of the size of the network (∼700K parameters). Consequently, sub-phoneme classification networks have 118 outputs (silence is kept as a single state monophone) instead of the 40 of the baseline system. The difference between conventional phone units and the multiple state model can be seen in the following transcription example of word "able":

```
ABLE  ey b ah l
ABLE  -ey ey ey+ -b b b+ -ah ah ah+ -l l l+
```

where `-ey , -b , -ah , -l` are left state units and `ey+ , b+ , ah+ , l+` are right state units of the corresponding phoneme.

The main difficulty of this extension is on the need for state level alignments. This fact is discussed in the next sections.

5.1. Initial HMM/GMM based alignment

In order to validate this approach, state-level alignments were initially obtained with the HMM/GMM speech recognizer. Then, this new sub-phoneme alignment was used to train classification networks for the several streams. Results obtained are shown in the first row of Table 2.

Relative to the hybrid baseline system, a 18.81 % relative WER reduction is obtained for the development test set (*Nov92*). This considerable improvement is due not only to the extension to multiple state modelling, but also to the generation of a better alignment indeed. Actually, single state monophone networks were trained with HMM/GMM based alignment and a slight improvement was also observed.

5.2. Initial blind state-level alignment

Ideally, state-level alignment generation in hybrid systems should be tackled independently of an external HMM/GMM recognizer.

Hence, we have experimented with various blind initializations in order to build sub-phoneme networks. The forced alignment generated by the well-trained single state system is automatically modified in order to translate the phoneme targets to sub-phoneme targets following some criteria. For instance, we have found that a good initialization would consist on forcing the first target and the last target of each block of identical phonemes to correspond to the left and right states respectively, and the middle frames correspond to the center state. This blindly generated alignment is used for training initial sub-phoneme classification networks. Then, multiple iterations of re-alignment and re-training are followed as in the case of the baseline system.

The last row of Table 2 shows the results obtained with the hybrid multiple state system based on blind initialization of state-level alignments. A considerable degradation caused by the initial blind alignment can be observed when comparing to HMM/GMM alignment results. However, improved performance is still achieved with respect to the hybrid baseline system thanks to multiple state modelling.

6. Acoustical modelling of phone transitions

Some previous works seem to support that in most cases a co-articulation effect on one side of the phone is practically independent on the other side. In other words, assuming that the phone nucleus is approximately stationary, the co-articulation effect of the left-context mostly affects to the initial transitional region of the phone and not to the rest of the phone, while the same can be said about the right context co-articulation effect and the final region.

Initial align	Nov92	si_dt_s6	si_dt_05.odd
HMM/GMM	7.9	11.14	12.75
Blind	9.19	11.93	13.31

Table 2: *WER results of the hybrid multiple state system using both alignments generated with HMM/GMM speech recognizer and initial blind alignment.*

Consequently, it is possible to recall the three-state phone model of the previous section as a combination of context dependent and context independent sub-phone units: the sub-phone units corresponding to the phone nuclei (context independent) and left and right context-dependent sub-phone units. In fact, since only one left/right context-dependent unit is trained for each phone, it can be said that they are general class context dependent units.

6.1. Introducing transition modelling: diphones

In this section, it is proposed to replace each adjacent general class right-dependent and left-dependent sub-phone unit by units that explicit model acoustic phone transitions, which are generally known as diphones. In this way, transcription of the word "able" with combined nuclei sub-phone and diphone units would be:

```
ABLE    ey ey_b b b_ah ah ah_l l
```

where `ey_b , b_ah , ah_l` are the diphone units.

Unfortunately, this acoustic phone model is unpractical in a connectionist approach since it imposes modelling each diphone as a classifier output and also due to the amount of data available. For instance, the theoretical number of diphones of any language is up to N^2 (with N number of phones), which in our case would mean more than 1500 phonetic units (although many of them do not exist due to characteristics of each language). An efficient classifier with such a number of outputs is unaffordable without considerably modifying the network architecture (for instance, with a cascade of classifiers).

6.2. A practical solution: reduced number of diphones

In the WSJ training data set considered, there are up to 936 different intra-word diphones. Obviously, not all of them have the same frequency. For instance, only with the first 61 diphones it is possible to give coverage to 50 % of all the intra-word diphones appearing in the training data. Figure 2 shows the coverage rate of the training data depending on the number of diphones.

In order to incorporate phonetic context modelling, more concretely transition modelling, it is proposed to simply add the most significant/frequent diphones as a possible sub-phone unit recognition to the three-state phone model. The three-state model of left and right general class context-dependent and nucleus units is still necessary to give coverage to all the data. With these "new" sub-phonetic units that combine both three-state modelling and phone transition modelling, the transcription of word "able" would be:

```
ABLE    -ey ey ey+ -b b b_ah ah ah_l l
```

where the concatenated units `ey+`, `-b` are not frequent enough in the training corpora to form a separate diphone unit and the three-state model is kept, meanwhile the sub-units `b+ -ah` and `ah+ -l` have their own diphone to model the transition (`b_ah` and `ah_l` respectively).

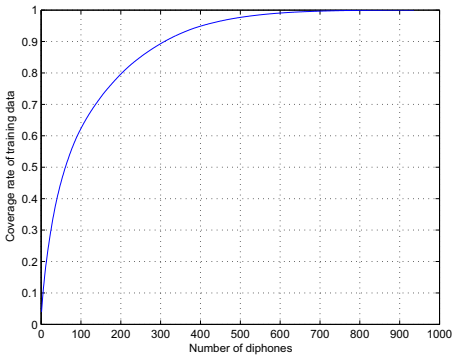

Figure 2: *Coverage rate of the training data depending on the number of diphones*

6.3. Experimental evaluation

In order to generate an initial alignment with phone transition modelling, the sub-phoneme forced alignment provided by the multiple state hybrid system is transformed. Context-right and context-left units that had an equivalent selected phone transition unit (frequent enough) are replaced by the corresponding diphone. This alignment is used to train initial networks. Then, iterations of re-training and re-alignment are done until stable performance in the development test set is achieved.

It is worth noticing that in this work only internal word diphones have been considered for training. The main reason is that our decoder system does not allow cross-word dependency decoding strategies .

Table 3 shows the results of the proposed approach for different diphone coverage rate. The number of diphones modeled is also provided in brackets. Notice that the size of the output layer of the trained networks is the sum of the number of diphones and the 118 sub-phoneme units of the three-state model, which produces networks of about 850K parameters in the case of the largest number of diphones.

The proposed approach that incorporates modelling of phone transitions outperforms both the baseline hybrid system and the multiple state hybrid system independently on the number of diphone units selected. The best system is the one with 80% of diphones coverage. In this case, compared to the baseline hybrid system (single state phone modelling), a 26.8% and 23.8% relative WER reduction is achieved for the *si_dt_s6* and *si_dt_05.odd* evaluation test sets respectively. A 19.45% and a 17.73% relative WER reduction is obtained when compared to the multiple state hybrid system.

Regarding the reference HMM/GMM systems, the best proposed approach obtains a 7.5% and a 11.7% relative WER reduction in the two evaluation sets with respect to the internal word triphone system. On the other hand, comparable results (slightly worse) are achieved compared to the cross-word triphone system.

7. Conclusions

In this work, an hybrid ANN/HMM speech recognizer based on single state monophone units is extended to incorporate acoustical modelling of phoneme transitions without significantly altering the system architecture. By combining the multiple state sub-phoneme recognition units with a restricted set of the most frequent diphones appearing in the training data, we showed that is possible to obtain remarkable performance improvements compared to the baseline connectionist approach and comparable performance to a reference cross-word triphone Gaussian-based recognizer in a medium vocabulary size task, such as the WSJ 5k. Future research efforts will be focused in validating this approach in larger vocabulary tasks.

Rate diphones (number)	Nov92	si_dt_s6	si_dt_05.odd
20% (14)	8.52	11.53	12.52
40% (40)	8.66	11.23	11.91
50% (61)	7.79	10.27	11.88
60% (91)	8.28	10.09	11.64
70% (137)	7.98	9.22	11.22
80% (203)	7.57	9.61	10.95

Table 3: *WER results of the proposed hybrid system that combines multiple state phone modelling and transition phone modelling for different number of diphones.*

8. References

[1] Rabiner, L.R., "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, 77(2):257–286, 1989.

[2] Lippmann, R. P., "Review of neural networks for speech recognition", Neural Computation, 1(1):1–38, 1990.

[3] Morgan, N. and Bourlad, H., "An introduction to hybrid HMM/connectionist continuous speech recognition", IEEE Signal Processing Magazine, 12(3):25–42, 1995.

[4] Bourlard, H., Morgan, N., Wooters, C. and Renals, S., "CDNN: A Context Dependent Neural Network For Continuous Speech Recognition", In Proceedings of ICASSP'92, II:349–352, 1992.

[5] Franco, H., Cohen, M., Morgan, N., Rumelhart, D. and Abrash, V., "Context-dependent Connectionist Probability Estimation in a Hybrid Hidden Markov Model-Neural Net Speech Recognition System", Computer Speech and Language, 8(3):211–222, 1994.

[6] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I., "AUDIMUS.media: A Broadcast News speech recognition system for the European Portuguese language", In Proc. of Int. Conf. of Computational Processing of Portuguese Language (PROPOR), 2003.

[7] Schwarz, P., Matejka, P. and Cernocky, J., "Hierarchical Structures of Neural Networks for Phoneme Recognition", In Proceedings of ICASSP'06, I:325–328, 2006.

[8] Paul, D. and Baker, J. M., "The Design for the Wall Street Journal-based CSR Corpus", in DARPA Speech and Natural Language Workshop, 1992.

[9] Woodland, P.C., Odell, J.J., Valtchev, V. and Young, S.J., "Large Vocabulary Continuous Speech Recognition Using HTK", In Proceedings of ICASSP'94, II:125–128, 1994.

[10] Mohri, M., Pereira, F. and Riley, M., "Weighted finite-state transducers in speech recognition", In ISCA ITRW Automatic Speech Recognition, pp. 97–106, 2000.