HYBRID CONNECTIONIST-STRUCTURAL ACOUSTICAL MODELING IN THE ATROS SYSTEM

M. J. Castro
mcastro@dsic.upv.es

F. Casacuberta
fcn@dsic.upv.es

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46022 València (Spain)

Abstract
In this paper, we introduce several hybrid connectionist-structural acoustic models for context-independent phone-like units in the ATROS recognition system. The structural part of the acoustic models has been modeled with Markov chains, and a multilayer perceptron (or a committee of multilayer perceptrons) is used to estimate the emission probabilities of the Markov chains. We compare the recognition performance attained by these models with the performance obtained by classical continuous density hidden Markov models on a semantic restricted task.

1 Introduction
Acoustic phonetic-decoding for continuous speech recognition is an open problem in speech research, because the final performance of an automatic speech recognition system greatly depends on the acoustic modeling quality. Hidden Markov models (HMMs) of phone-like units are the most popular option for modeling speech sounds.

Under the statistical framework [1], the problem of speech recognition is to search for a word string \( \hat{w} \) that maximizes the a posteriori probability of \( w \) given a sequence of feature vectors \( x \),

\[
\hat{w} = \arg\max_w \Pr(w) \Pr(x|w).
\]

The first term \( \Pr(w) \) can be estimated through a language model, and the second term \( \Pr(x|w) \) can be estimated through acoustic models. In this paper, we focus on acoustic modeling. We have trained and tested hybrid acoustic models composed of Markov chains and neural nets [2–4]. The structural part has been modeled with Markov chains. A multilayer perceptron (MLP) or a committee of MLPs (CMLP) [5] is used to estimate the emission probabilities of the Markov chains.

In the following section, we describe the recognition system that has been used in all the experimentation. Section 3 is devoted to the training of the hybrid connectionist-structural acoustic models. Section 4 presents the experimentation with an application task, that consists of queries to a geographical database. Finally, we summarize the work and draw some conclusions.

2 ATROS: Recognition system overview

We have used in all the experimentation the recognition system developed under the Spanish Senglar project, ATROS (Automatically Trainable Recognizer Of Speech) [6, 7], which assumes the statistical framework.

The system generates a word-string, from the language model, as an hypothesis of the words that have been uttered. The language model is represented in ATROS by a stochastic finite state network that can represent a stochastic regular grammar (i.e. a n-gram). The word acoustic models are generated by using acoustic models of phone-like units and lexical models. Lexical models are easily generated by Spanish orthographic-phonetic rules and they are represented in ATROS by stochastic finite state networks whose transitions are labelled by phone-like units. The phone-like models in ATROS are continuous density HMMs or hybrid connectionist-structural models. The word acoustic models are generated by substituting the transitions in lexical models with the corresponding phone acoustic models.

For decoding, the word acoustic models are dynamically integrated in the language model by substitutions of the transitions in the language model by the corresponding word acoustic models. The decoding process with such an integrated network is performed with a beam-search Viterbi algorithm. Every experiment is performed on a SG12 workstation R10000 with 384 MB of RAM.

3 Hybrid acoustic models

3.1 Experimental framework

The acoustic models were trained with part of the Spanish EUROM.1 database [8] and other acoustic
material which was recorded at the Universities of Valencia and Catalunya. The overall training database is composed of 1,529 utterances from 57 speakers (470,000 acoustic frames and 55,000 phones). A small part of the training material (162 utterances) was manually segmented and the rest of the utterances were automatically segmented.

A test set was used to assess the phonetic recognition performance of different types of acoustical models. This test set was composed of part of eurom.1 and other phonetic databases, for a total of 375 sentences from 20 different speakers (17,625 phones).

The speech was parametrized with mel-cepstrum coefficients, using cepstral mean subtraction since the utterances were recorded differently, resulting in vectors of 12 Cepstral coefficients and energy.

All the sentences were automatically transcribed into sequences of phone-like units. This set was composed of 24 units that roughly corresponded to the Spanish phonemes (plus one unit to model the silence).

3.2 Hybrid acoustic models training

We have trained hybrid acoustic models composed of Markov chains and an MLP to estimate the emission probabilities [2–4], being tied the emission probabilities of the states of each model.

Different topologies of Markov chains were tested for the hybrid HMM/MLP models. An MLP with 24 output units was used to estimate a posteriori probabilities of the classes, given the acoustic input. Each output unit of the MLP was associated to each phone (as emission probabilities of the states of the models were tied, only one output unit was needed for each model). The acoustic input to the MLP was formed by the actual frame plus four frames of left and right context. Different sizes of the hidden layer were tested. A posteriori probabilities estimated by the MLP were used directly as emission probabilities of the models.

In the training process of the MLP, the desired output was 1 if, according to the segmentation of the training data, the acoustic vector corresponded to the phone associated to the output unit; and it was 0 otherwise. The MLP input layer was formed by 117 inputs corresponding to an input window of nine acoustic vectors (normalized to have zero mean and unit standard deviation over the training set). Different sizes of hidden layers have been tested: MLPs with 100, 500, and 1,000 hidden units and with two layers of 100 units each.

The training of the MLP was performed using the on-line scheme of the backpropagation algorithm [9] with a sigmoidal function (in order to verify stochastic constraints, a normalization over all outputs was performed). The criterion function was the mean squared error. To prevent overtraining, after each epoch, the classification performance at acoustic vector level was measured on a validation set (out of the training set, a subset was selected as a validation set –10% of the total training data—) and the training process of the MLP was stopped when no improvement was expected.

3.3 Phonetic recognition performance

The acoustic models were used to perform an acoustic-phonetic decoding task with the database described previously. This task consists of building a finite state network in which all acoustic models are in parallel. There is an initial state connected to every initial state of each acoustic model and a merged final state of the final states of every model. The added initial and final states are connected by a back transition with probability one.

Experiments with different topologies of Markov chains were carried out and the best performance was obtained using left-to-right Markov chains of three states with self-loop transitions in every state and no skips.

After that, we combined this kind of Markov chains with an MLP of increasing size of the hidden layer (100, 500, 1,000). The performance of the acoustic models was incrementing as the size of the hidden layer did. Nevertheless, the MLP with two hidden layers of 100 units each significantly outperformed the other acoustic models.

Details of this experimentation can be found in [4], where it was shown that the best acoustic models which used an MLP in isolation was the HMM/MLP2×100 system (composed of Markov chains and an MLP of two hidden layers of 100 units each), with a percentage of phone errors of 29%.

The following series of experiments were carried out using a committee of MLPs instead of a single MLP in order to obtain a better classifier [5]. A simple form of committee of MLPs can be built by combining two or more networks, being the output of the committee the average of the outputs of the MLPs which comprise the committee. The importance of such an approach is that it can lead to significant improvements in the prediction on new data, while involving little additional computational effort. In fact, the performance of a committee can be better than the performance of the best single network used in isolation [5]. A scheme of a committee of three MLPs is illustrated in Figure 1.

We performed experiments with different committees of two and three MLPs, showing that every system with committees achieved better performance than any system using only one MLP. The very best result, a percentage of phone errors of 27%, was achieved with a committee of three MLPs.
4 Experiments with the GDQ task

We carried out some experiments to evaluate the performance of the recognition system on an application task and to establish the influence of each acoustic model which was presented in the previous section.

4.1 Experimental framework. The GDQ task

A task-oriented Spanish speech corpus devoted to the training and testing of automatic speech understanding systems in a constrained semantic context was defined and recorded by means of the Albayzin Spanish project [10]. The selected application was queries to a geographical database (GDQ task). According to the conceptual scheme of the semantic universe (which can be consulted in [10]), queries can be made about regions, rivers, mountains, seas, etc. After the written acquisition, a small number of the sentences were selected to be pronounced by different speakers.

The training set used for the estimation of the language model for the task was defined and consisted of 8,221 written sentences (78,200 words), with a vocabulary of 1,264 words. A test set of 1,138 different written sentences (11,200 words) was used to measure the perplexity of the obtained model.

The vocabulary of the GDQ task (1,264 words) has been represented by the concatenation of sublexical models (according to orthographic-phonetic rules) to form word acoustic models.

A trigram language model (smoothed with bigrams and unigrams with back-off) was estimated using the second version of the Stochastic Language Model toolkit [11]. The perplexity of the test set with the trigram model was 10.22.

Table 1: Word error rate (wer) obtained for the test set of the GDQ task using a trigram language model. The real time factor (rtf) is shown in the third column.

<table>
<thead>
<tr>
<th>System</th>
<th>wer</th>
<th>rtf</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM/MLP</td>
<td>7.48</td>
<td>0.5</td>
</tr>
<tr>
<td>HMM/C2MLP</td>
<td>7.36</td>
<td>0.7</td>
</tr>
<tr>
<td>HMM/C3MLP</td>
<td>7.23</td>
<td>1.2</td>
</tr>
</tbody>
</table>
of the HMM/C3MLP experiment is that most of the errors (∼ 80%) are on monosyllabic words (prepositions, articles, etc.). These kinds of errors produce very low recognition rates at sentence level (66%). On the other hand, it is interesting to note that these errors do not affect the meaning of the query. Thus, in a real application when the answer must be given to the user, recognized sentences with those errors would still provide a correct answer.

In order to compare the results of our experiments with a system using Continuous Density HMMs acoustic models, several experiments were performed with HMMs (CDHMMs of 16, 32, and 64 gaussians) [7] on the same task. The best result was a word error rate of 6.93%, obtained with CDHMMs of 32 gaussians, and a real time factor of 49.4 (close to 50 seconds to process one hundred frames). In real time, and by using speed-up techniques (histogram pruning) in the ATROS system [7], the best achieved performance was 7.10%, with the system that also used 32-gaussian CDHMMs.

5 Summary and conclusions

Hybrid connectionist-structural models are used in an acoustic-phonetic decoding task. We trained and tested acoustic models where the structural part was modeled with Markov chains. An MLP was used to estimate the emission probabilities of the models. We followed with acoustic models that used committees of MLPs instead of a unique MLP to estimate the emission probabilities of the structural models and showed that this significantly increased the phonetic performance.

Secondly, we worked with the GDQ task, composed of 1,264 words and a trigram language model with a test set perplexity of 10.22. We performed experiments with the hybrid acoustic models and obtained a word error rate of 7.23% on the task. These results were equal when compared with those obtained with 32-gaussian CDHMMs and with speed-up techniques in the ATROS system.

Our next goal is to model contextual-dependent units using hybrid acoustic models. Moreover, experimentation with hybrid HMM/MLP models where MLPs are used to estimate emission probabilities for each state of the Markov chains are needed in order to fully evaluate the technique.

In addition, we are planning to incorporate speed-up techniques in the search process of this version of the ATROS system (fast-phoneme look-ahead and histogram pruning as in [7]) in order to reduce the search effort and to allow to use wider beams and more powerful and computationally acoustic models (hybrid contextual-dependent units or more detailed hybrid acoustic units).

References