DYNAMIC ADAPTATION OF VOCABULARY INDEPENDENT HMMS TO AN APPLICATION ENVIRONMENT

C. Vair and L. Fissore and P. Laface

CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - I-10148 Torino, Italy
E-Mail Luciano.Fissore@cselt.it Claudio.Vair@cselt.it

Dipartimento di Automatica e Informatica - Politecnico di Torino
Corso Duca degli Abruzzi 24 - I-10129 Torino, Italy
E-Mail laface@polito.it

ABSTRACT

The paper presents a software architecture allowing to collect, select, and exploit speech data from a specific application field to dynamically generate Hidden Markov Models tailored to that application environment and vocabulary.

The framework we are interested in is, therefore, an already operational voice activated service that allows to collect directly from the field a large amount of speech data.

We propose a procedure for data selection and for incremental training of the units using a strategy of model selection.

Several tests are presented for a train timetable information system, and for a Directory Assistance application with a very large vocabulary of city names showing that significant improvements can be obtained with respect to the laboratory models, keeping the old models and transcribing only the most frequent words in terms of the new units, incrementally trained from the field data.

1. INTRODUCTION

It is well known that the use of speech data collected from the field, during the interaction of real-life users with a specific application, allows the acoustic models of the system to be adapted to the application vocabulary increasing the robustness of the recognizer. It is also well known that the integration of the information derived from real-life data with the knowledge embedded in the so called “laboratory” acoustic models, typically trained with a large amount of read data, is a key factor to achieve high performance in application trials.

Maximum A Posteriori estimation [2, 4] is one of the most successful approach to speaker and environment adaptation. Good results have been reported using MAP as a mean for the combination of laboratory and field data [5, 6]. In these approaches interesting solutions were proposed to the problems raised by MAP estimation, namely:

- the estimation of the a priori distribution parameters
- the lack of field data for the adaptation of the distributions of some models
- the optimization of a forgetting parameter for achieving a good tradeoff between the prior information and the application data.

The aim of this work was to find a procedure for incrementally generating, during the lifetime of a vocal access information system, a set of models that could improve its performance, going beyond the capabilities of MAP trained, field environment adapted, models.

A trivial observation is that several important telephone services exist that are already operational, possibly as prototypes. These systems allow a large amount of user interactions to be easily collected. This reduces the problems related to the lack of training data, rather, it raises the problem of how to filter a huge amount of speech data that are often unevenly distributed, or incorrectly labeled. For example, it is normal for some vocal services, such as timetable information or directory assistance systems, that the requests of real-user for city names are mostly concentrated on some very popular entries like Milano and Roma, in Italy. Moreover, since it would be expensive to use a human operator to label every utterance, the field collected data are automatically labeled. The resulting tokens are affected, thus, by labeling errors due to incorrectly recognized in-vocabulary or out of vocabulary (OOV) words.

The paper presents a software architecture allowing to collect, select, and exploit speech data from a specific telephone application field to dynamically generate system HMM models tailored to that application environment and vocabulary.

We propose a procedure for data selection and for incremental training of the units, using a strategy of model selection that allows to improve the performance of the isolated word recognizer module of an already deployed vocal access information system.

2. SYSTEM ARCHITECTURE

Figure 1 illustrates the proposed architecture for the dynamic generation of the HMMs of a given application. The Recognition and Labeling module, described in more detail in Section 3, automatically labels the field signal collected from the real-user interactions with the system. It rejects the tokens that have not been reliably recognized, while it associates the recognized label to the reliable tokens, and collects and classifies them according to a set of properties, such as length of the token, sex of the speakers, reliability score, and others features that can be derived from the application knowledge, or estimated from the data. The field tokens together
Signal from the application
Recognition and Labelling Token Database

Subword training
Knowledge Database
Training
Whole word training

Decision Model Training
Decision Model Maker

Deployment Decision

Figure 1: Model adaptation system architecture

In a second step, a Filtering module makes queries to the Token Database, according to a given selection rule, and updates the training lists and the occurrence statistics of the selected words in the Knowledge Database. The system model creation or updating is performed by the sequence of steps 3 and 4 in Figure 1. The first one, the Training Decision maker, decides, on the basis of the occurrence counts of the tokens in the Knowledge Database, among four possible actions: to create a new model, to update a model, to remove a model, or to do nothing. In the first two cases a HMM training process is activated to perform the requested action. It is worth noting that the decision may activate the training of sub-word or even of whole word models.

The decision of promoting online a set of new models is taken by the Deployment Decision module on the basis of the number of tokens that have contributed to the model training. The actual deployment is performed by the Model Maker module that builds up the application models from the components selected by the Deployment Decision module, and also produces the acoustic-phonetic units information necessary to create the new transcriptions of the vocabulary words. The Knowledge Database acts like a blackboard where all the statistics are collected: in particular, it stores the current set of HMMs, and the accumulated parameter counts of every old or newly created HMM unit, together with the occurrence counts of the selected tokens and units.

Using this architecture, we have tested several procedures for incremental training of the units, and for model creation and selection that will be illustrated in the following Sections.

3. TASK AND CORPUS

The test-bed for the study of the best training and selection strategies has been a vocal access train timetable information system operational in several call centers of the Italian railway service provider “Ferrovie dello Stato”. Its recognition module uses a CSELT speech recognizer for flexible vocabularies, that is based on a set of 391 vocabulary and gender independent acoustic-phonetic units including stationary context-independent phonemes models and diphone-transition coarticulation unit models [1]. The units are modeled by left-to-right continuous density HMM with a maximum of 32 Gaussian densities per state. The observation vector for the experiments performed includes 39 parameters: 12 liftered and high-pass filtered cepstral coefficients ($C_i$, $C_{i+2}$), and their first and second order derivatives, the log-energy, and its first and second order derivatives.

The set of the 391 vocabulary independent models was trained using about 30 hours of PSTN phonetically balanced read speech, collected from more than 5000 speakers. The field data set, instead, was collected during the interaction of real-users with the information systems located in the call centers.

The reported experiments were performed on an isolated word recognition task with a vocabulary of the 664 city names corresponding to the main connections of the Italian railway network. For the generation of the new models different subsets of more than 38000 tokens have been used, while 6983 tokens have been collected and labeled for the test set.

4. DATA SELECTION

The Recognition and Labeling module selects the field samples that are useful for training by labeling the pronounced words and
Training number of word transcriptions using size laboratory sw sel. field sw sel. field ww
2000*2 664 18 5
2000*4 664 31 5
4000*1 664 19 10
4000*2 664 35 13
23156 664 40 20

Table 1: Number of word transcriptions. Sel. field sw and Sel. field ww refer to the transcription of the most frequent sub-word and whole word models respectively.

associating to them a reliability score. Unfortunately, the reliability score is not itself “reliable” in the case of out of vocabulary words, thus part of field data still remains incorrectly labeled. The effect of these tokens on the model quality will be discussed in Section 6.

5. MODEL TRAINING AND SELECTION

As briefly introduced in Section 2, the Training Decision maker monitors the content of the Knowledge Database. When the number of the available tokens is greater than a given threshold it activates the updating of the models. Only the unit models that occur frequently according to another threshold are actually reestimated. For the models that do not yet satisfy this condition, only the accumulator counts are stored in the Knowledge Database.

Training of the field models is performed by an incremental MAP based approach similar to the one proposed in [3] using the laboratory models, or the previously computed models, for obtaining the necessary a priori knowledge.

It is worth noting that the new field dependent unit models are added to the original, not adapted, laboratory models. The field trained units are used to transcribe only the most frequent application words. Thus, the number of the system units slightly increases, and the most frequent words are transcribed in terms of two different set of units: one vocabulary independent, the other one application dependent.

If the occurrence of a word is much larger than the other ones, the Training Decision maker may activate a whole word training process. For these words the system will have yet another very specific model.

Table 1 reports the number of word transcriptions obtained by using training sets of different size randomly selected from the field data. In the table, the caption “Laboratory sw” refers to the transcriptions of the 664 vocabulary words using the laboratory models. The captions “Sel. field sw (ww)” refer to the number of words, selected by the Training Decision and Deployment Decision modules on the basis of their frequency in the field corpus, that are transcribed in terms of field trained sub-word (whole word) models.

Since the training data are unevenly distributed, the risk of penalizing infrequent words is reduced by keeping the laboratory models. Moreover the laboratory models, with their baseline performance, reduce the danger that this completely unsupervised training process can lead to unpredictable results.

It is possible that some words, for which field adapted models have been produced, during the lifetime of a system appear less and less frequently while other ones emerge. In this case, the system can add the new word models and remove the least frequent ones.

6. RESULTS

From the 38000 tokens collected from the field, 26246 were selected by the Filtering module according to their high reliability score. The percentage of OOV words in this set is 11.7 % for a total of 3813 incorrect labels. Notice, please, that the OOV rejection strategy has not been used for these experiments.

In Figure 2, the performance of the models trained using MAP estimation, with a forgetting factor optimized to give the best weighting between field data and laboratory models, can be compared with Maximum Likelihood estimation performed by means of the Forward-Backward algorithm (FB). The training set includes different subsets of the 23516 tokens, containing in-vocabulary words only. The labeling error rate of this subset is 3.1%. For this first experiment, however, the correct transcriptions of all the tokens have been used.

It can be noticed that FB and MAP training converge to the same performance when enough training data is provided.

A second experiment has been performed to compare the performance of the laboratory models with respect to the models trained with the complete set of field data, using again the in-vocabulary training set, but looking for the effect of the incorrectly labeled tokens on the quality of the models.

The results reported in the first three columns of Table 2 confirm that the use of field data allows to almost halve the error rate. In this table, the caption “All field sw” means that all the in-vocabulary words are transcribed with field trained sub-word units. Selected field sw (ww) refer, instead, to the transcriptions of the selected most frequent words in terms of their sub-word (whole word) models.

Another interesting result, presented in the last two columns of the table, is that the effect of incorrectly labeled tokens is negligible for our training method, while it produces a sensible increase of the error rate for the FB trained models.

Since the proposed training has shown to be robust with respect to incorrectly label tokens, and since we are interested in a dynamic adaptation of the models to the application environment, we performed another set of experiments of incremental training selecting the training subsets from the entire set of the 26246 tokens, including the OOV words. The number of word transcriptions has been given in Table 1, while the obtained results are reported in...
Table 2: Recognition performance using the complete set of 23156 field tokens

<table>
<thead>
<tr>
<th>Training tokens correctly labeled</th>
<th>Training tokens incorrectly labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription models</td>
<td>Laboratory</td>
</tr>
<tr>
<td>Errors (Error rate %)</td>
<td>583 (8.3)</td>
</tr>
<tr>
<td></td>
<td>270 (3.9)</td>
</tr>
</tbody>
</table>

Table 3: Number of errors (error rate %) for incremental training

<table>
<thead>
<tr>
<th>Training size</th>
<th>Word transcription using</th>
</tr>
</thead>
<tbody>
<tr>
<td>All field sw</td>
<td>Lab. + All field sw</td>
</tr>
<tr>
<td>2000*2</td>
<td>307 (4.4)</td>
</tr>
<tr>
<td>2000*4</td>
<td>285 (4.1)</td>
</tr>
<tr>
<td>4000*1</td>
<td>289 (4.1)</td>
</tr>
<tr>
<td>4000*2</td>
<td>267 (3.8)</td>
</tr>
</tbody>
</table>

Table 4: Results for the Directory Assistance database

<table>
<thead>
<tr>
<th>Models</th>
<th>Errors</th>
<th>Error rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC Vocabulary Independent</td>
<td>2178</td>
<td>24.8%</td>
</tr>
<tr>
<td>RASTA-LDA Vocabulary Independent</td>
<td>1536</td>
<td>17.8%</td>
</tr>
<tr>
<td>+ 40 Vocabulary Dependent city names</td>
<td>1122</td>
<td>12.8%</td>
</tr>
<tr>
<td>+ 20 whole word models</td>
<td>1000</td>
<td>11.4%</td>
</tr>
</tbody>
</table>

Table 3. In the first column, the training set size is given according to the number of token chunks that have been used for the incremental training. Again, our proposed approach outperforms the system using word transcriptions in terms of the complete set of field adapted models. Moreover, it seems that larger training chunks allow to converge faster to the full training set results. It is also worth noting, looking at the second column of Table 3, that using both the laboratory and the field trained model transcriptions is not beneficial in terms of performance and would also increase the memory occupation and the computation time.

Table 4 shows the results obtained on a spontaneous speech database collected from an Italian Directory Assistance service prototype with a 9325 city name vocabulary and a test set of 8775 tokens. From the 38200 tokens collected from the field, 19700 city names and 8800 province city names were selected by the Filtering module according to their reliability score. The use of the vocabulary dependent subwords for the most common city names significantly improves the performance of an LDA RASTA PLP front end (from 17.8% to 12.8%), and further improvements are given by the addition of the most frequent city names whole word models.

7. CONCLUSIONS

It has been shown that it is possible to improve the performance of a speech recognizer by dynamically adding to the set of laboratory units a limited number of units trained from the field data. In particular, the results of our experiments show that improvements are obtained by

- transcribing the system vocabulary with the laboratory models, rather than with the models trained using the field data only. This is true even if the amount of field data is comparable with the one used for training, because the field corpus is unevenly distributed,
- adding the transcriptions of the most frequent application words in terms of the units trained from the field data, rather than transcribing all the vocabulary words,
- adding whole word models for the most frequent application words
- The new transcriptions have a marginal negative effect on the recognition of less frequent words.

8. REFERENCES