MLP-HMM Two-Stage Unsupervised Training for Low-Resource Languages on Conversational Telephone Speech Recognition

Yanmin Qian¹, Jia Liu²

¹ Institute of Intelligent Human-Machine Interaction, MOE-Microsoft Key Lab. Of Intelligent Computing and Intelligent Systems,
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
² Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China
yanmingqian@gmail.com, liuj@mail.tsinghua.edu.cn

Abstract
This paper focuses on speech recognition applications where there is a limited amount of manually labelled training data in the target language, but plentiful unlabelled data. We investigate approaches based on unsupervised training following the traditional method, we proposed a more effective and efficient data selection principle considering confidence scores as well as phone frequency. In addition, we transfer the HMM-based unsupervised training to MLP feature level at the first time, and obtain much more robust MLP-based features. Taking into account that HMM or MLP based unsupervised trainings are focused on model or feature level of speech recognition systems, we combined these two approaches finally, and proposed a more optimized strategy to get further improved unsupervised trained system in the low-resource scenario. In our experiments, we get significant improvements of about 12% relative versus a conventional baseline in this low-resource scenario.

Index Terms: low resource speech recognition, unsupervised training, multi-layer perceptron, data selection

1. Introduction

Rapid development of ASR systems for resource insufficient languages has recently attracted much interest [1-2]. Several approaches have recently been proposed to address this, including SGMM based acoustic modeling [3-4] and articulatory attribute based multilingual MLP features [5-6].

The transcription of the speech corpus involves significant manual labor and is very expensive and time consuming. Unsupervised training can significantly reduce the human effort required. The typical procedure for unsupervised training includes using a seed model, trained on a small amount of a manually transcribed corpus, to recognize a large quantity of unlabelled speech data. The recognized hypotheses are filtered and then the chosen hypothesis transcriptions are combined with manual transcriptions to retrain the HMM acoustic model. This approach has been demonstrated to be effective in Broadcast news and conversational speech scenarios [7-10], and is also a promising approach for multilingual tasks [11].

In this paper, we follow this HMM-based unsupervised training approach, and propose a more effective data filtering method. We also present a NN-based unsupervised training method for robust MLP feature extraction. The final system consists of MLP-HMM two-stage combined unsupervised training to get most improved performance.

The paper is organized as follows: in section 2 and section 3 we describe the HMM-based and NN-based unsupervised training respectively. Then we combine these two methods and introduce the MLP-HMM two-stage combined unsupervised training strategy in section 4. Section 5 gives our experimental setup and presents experimental results. Finally, we conclude in section 6.

2. HMM-based unsupervised training for acoustic modeling

2.1. Lattice posterior probability based data selection

We use a word lattice for compact representation of multiple hypotheses. Applying the forward-backward algorithm on the lattice, lattice-based posterior probability confidence scores have previously introduced for spoken term detection [12] and unsupervised data filtering [13]. Normally the word-level posterior is used representing the confidence score that word \( W \) occurred in the time \( [t_s, t_e] \) [13]:

\[ conf(W) = P(W[t_s, t_e] | O_{t_s, t_e}) \] (1)

Based on this word level posterior probability, an utterance level posterior confidence can be calculated as (2)

\[ conf(U) = \sum_{i=1}^{N} conf(W_i) \cdot T(W_i) / \sum_{i=1}^{N} T(W_i) \] (2)

where \( N \) is the number of words in the utterance, \( T(W_i) \) is the duration of each word, and \( conf(W_i) \) is the word confidence score which could be computed using equation (1).

Usually a confidence threshold \( \epsilon \) is chosen to control the amount of selected data. The following algorithm summarizes the conventional procedures using the confidence score.

---

Data selection using posterior confidence score

**Calculate the word or utterance level posterior confidence score by the Forward-Backward algorithm**

for each data unit \( S_i \) (word or utterance) do

if the posterior confidence is higher than the threshold \( \epsilon \) then select this data unit \( S_i \) for later model retraining

end if

end for
2.2. Phone frequency principle based data selection

We also tried a second approach that makes use of the phone frequencies. The basic intuition is that if a corpus has a large quantity of training samples that is significantly unbalanced across phones, parameters will not be robustly estimated, and the only point of data selection is to overcome data sparsity. Table 1 illustrates the phone occurrence rate for randomly chosen one hour of manually labelled CallHome English training data [14], including both monophones and triphones. We find that the phone coverage problem is significant with some phone occurrences very low in the low-resource scenario.

<table>
<thead>
<tr>
<th>Monophone</th>
<th>Occurrences</th>
<th>Triphone</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1345</td>
<td>l-Y+k</td>
<td>108</td>
</tr>
<tr>
<td>I</td>
<td>1652</td>
<td>@-n+d</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>53</td>
<td>E-l+i+h</td>
<td>3</td>
</tr>
<tr>
<td>O</td>
<td>16</td>
<td>k+s&lt;E</td>
<td>1</td>
</tr>
<tr>
<td>Z</td>
<td>3</td>
<td>i+v+b</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Phone occurrence on randomly selected 1 hour manually labelled CallHome English data

We propose a phone frequency based data selection approach. When considering the accuracy of the recognized transcriptions, we will be more inclined to select the data with lower phone frequency in the limited manually labelled data set. This idea is motivated by the cross-lingual data borrowing strategy proposed in [4] and [15]. The way we apply this intuition is to select a frequency cutoff, and to treat data unit with frequency above and below this value differently, in that we apply two different confidence thresholds: a small threshold \( e_2 \) for data units with “small” phone frequency and a larger threshold \( e_1 \) for data units with “large counts”. We calculate each unit’s phone frequency by considering both monophone and triphone frequency factors, and adding them together with the same weight.

The modified data selection algorithm is as follows:

Data selection using phone frequency principle

Calculate posterior confidence score and phone frequency for all the data units

\[
\text{for each data unit } S_i \text{ (word or utterance)} \text{ do} \\
\quad \text{if the posterior confidence is higher than the threshold } e_1 \text{ then select this data unit } S_i \text{ for later model retraining} \\
\quad \text{else if the posterior confidence is higher than the threshold } e_2 \text{ then select this data unit } S_i \text{ for later model retraining} \\
\quad \text{end if} \\
\text{end for}
\]

3. NN-based unsupervised training for MLP feature extraction

Previous unsupervised training methods have focused on HMM models [7-10], borrowing data from the unlabelled data at an acoustic model level. Considering the promising gains from unsupervised training, we extend this idea to the feature-level and get further improvements for low-resource language recognition. We use a NN as a bridge, and propose to implement this approach for MLP feature extraction. (We need to clarify that the unsupervised NN training described here is different from the recently interested RBM training [16]. The RBM training is class irrelevant and is an initialization method for NN training, while the unsupervised NN training here is still class-based, and it uses hypothesis as the training transcriptions since lack of manual labels. The detailed algorithm is described as bellows.)

Similar to HMM unsupervised training process, we also need to generate hypotheses and then filter the data, then utilize an enlarged transcribed speech corpus with phone-level transcriptions to train a neural network, rather than a HMM as usual. We use a 9 frame spliced PLP as NN\#1 inputs. The NN\#1 with unsupervised training can utilize significantly more data, so it can use more hidden nodes than the normal supervised approach, with robust parameters estimation, as shown in the left dashed box of Figure 1. The unsupervised trained NN\#1 is not very accurate due to the label errors in the hypotheses, so we use the limited manually labelled data to adapt the original NN\#1 to refine the performance of the classifier. We adapt the NN\#1 by retraining it using limited manually labelled data after initializing it with the previous NN\#1’s weights, to obtain more refined and accurate classifier NN\#2, shown in the right dashed box of Figure 1.

We could use these two NNs, unsupervised NN\#1 or adapted-unsupervised NN\#2, to generate MLP features for HMM training individually, by applying a typical tandem processing [17] on the NN outputs.

4. MLP-HMM two-stage combined unsupervised training strategy

The model-level HMM approach in section 2 and the feature-level approach in section 3 are focused on distinct levels of the speech recognition system, so we combine these two unsupervised trainings to create a more integrated strategy to get further improved performance. The two-stage combined unsupervised training is summarized as Algorithm I.

Algorithm I: MLP-HMM two-stage combined unsupervised training strategy

1. PLP seed model training: use small amounts of manually transcribed speech data and PLP feature to train an initial PLP-GMM-HMM model.
2. Transcription auto-generation #1: use the PLP seed
model to recognize the unlabelled speech utterances, and obtain initial 1-best hypotheses and word lattices.

3. Data selection #1 for NN-based unsupervised training: use phone frequency principle to filter the hypotheses transcription #1, and getting reliable and efficient recognized transcriptions.

4. NN-based unsupervised training for feature extraction: pool the filtered hypotheses data #1 with the manually labelled corpus, and follow the approach described in section 3 to train an MLP on the enlarged speech corpus #1.

5. MLP seed model training: use the MLP features obtained from procedure 4 to train initial MLP-GMM-HMM model, which is better than the PLP seed model.

6. Transcription auto-generation #2: use the MLP seed model to recognize the unlabelled speech utterances a second time, giving new 1-best hypotheses and word lattices #2.

7. Data selection #2 for HMM-based unsupervised training: use phone frequency principle to filter hypotheses transcription #2, and getting reliable and efficient recognized transcriptions.

8. HMM-based unsupervised training for acoustic modeling using MLP feature: pool the filtered hypothesis data #2 with manually labelled corpus, and follow the method described in section 2 to retrain the MLP-GMM-HMM seed model on the enlarged speech corpus #2.

9. Unsupervised training process iteration: repeat the steps 1~8, and train iteratively, the system can be refined.

5. Experiments and results

5.1. Experimental data and baseline system

Our experimental setup is similar to our previous work [4-5], using conversational telephone data. We use two setups to simulate the low-resource application, with English as the target low-resource language: in setup #1, we use 1 hour of randomly chosen speech from the entire 15 hours of CallHome English corpus as the manually labelled training data, and the remaining 14 hours are used as unlabelled data. In setup #2, we use all 15h CallHome English data as the manually labelled data set, and randomly choose 100 hours of speech from the Switchboard I corpus [18] as the unlabelled corpus. In both setups we use the CallHome English evaluation set, roughly containing 1.8 hours of speech, as our test set.

We use 39-dimensional PLP parameters, plus per-speaker mean and variance normalization, to train cross-word triphones, and use a trigram language model with a word-list of 62K words obtained by interpolating individual models trained from the English Callhome corpus, the Switchboard corpus and the Gigaword corpus [19].

In this study, the parameter estimation criterion used in unsupervised or supervised training is restricted to the Maximum Likelihood framework, but the approach can be extended to discriminative training [9-10].

We tuned the model size of 550 states with 4 Gaussians per state in setup #1, 1930 states with 16 Gaussians per state in setup #2, to get the best baseline systems. The performances of the baseline supervised PLP-GMM-HMM systems are summarized in Table 2. It is clear that the ASR systems built with low resources perform poorly, and that availability of more transcribed data could improve performance.

<table>
<thead>
<tr>
<th>System description</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>setup #1: using 1hour manually-labelled data</td>
<td>72.6%</td>
</tr>
<tr>
<td>setup #2: using 15 hour manually-labelled data</td>
<td>55.2%</td>
</tr>
</tbody>
</table>

5.2. HMM-based unsupervised training evaluation

To assess the impact of using different confidence score based data selections (word or utterance posterior). Table 3 shows the hypothesis transcriptions WER on different amounts of chosen data by varying threshold e for setup #1, obtained by comparing the 1-best hypothesis with true transcriptions. It can be observed that the WER of the hypotheses are much better than the baseline despite the poor baseline performance, and the accuracy of the selected data using word level confidence score is better than that of the utterance level confidence score.

<table>
<thead>
<tr>
<th>Amount of selected data</th>
<th>1 h</th>
<th>2 h</th>
<th>3 h</th>
<th>4 h</th>
<th>5 h</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>word-conf</td>
<td>25.6</td>
<td>38.8</td>
<td>44.3</td>
<td>48.1</td>
<td>51.0</td>
<td>…</td>
</tr>
<tr>
<td>utterance-conf</td>
<td>29.1</td>
<td>40.1</td>
<td>46.4</td>
<td>50.3</td>
<td>53.3</td>
<td>…</td>
</tr>
</tbody>
</table>

Figure 2: Performance comparison as the amount of selected data increases, using different data selection methods (setup#1)

Pooling the filtered data and the original manually labelled data, we retrain the acoustic model and build ASR systems. Figure 2 shows the systems WER as the amount of selected data increased, using different data selection strategies. We get large improvements from acoustic model unsupervised training compared to the baseline system. Surprisingly the performance using utterance level confidence score is consistently better than the word one, the opposite of the hypothesis accuracy results described above. Since an
utterance is a longer unit than a word, and it contains more useful context information, which is particularly important in the context-dependent acoustic modeling (e.g. triphone).

Using the phone frequency principle with the better utterance-level confidence score, shown in Figure 2, gives further large improvement, especially with small amounts of hypothesis data, and it converges to the best point more quickly than the confidence score one. We can see that a reliable and efficient data selection approach is very important in unsupervised training.

5.3. NN-based unsupervised training evaluation

We use NN outputs (monophone posteriors) to do viterbi phone recognition as the ANN-HMM system [20] to evaluate the unsupervised trained NN-classifier. The performance comparison (Phone Error Rate, PER) of the traditional supervised trained MLP and unsupervised trained MLPs using word and utterance confidence scores are shown in Table 4 (setup #1). We tuned the amount of selected data to achieve the best unsupervised NN classifier performance. We can see that the unsupervised MLPs obtain nearly a 10% PER relative decline. Contrary to the results of the acoustic model training, the word level confidence score is better than the utterance one. NN’s training emphasis accuracy of training data, and the context information is not as crucial as in HMM training, so more accurate hypotheses give better classifier performance.

Table 4. Phone Error Rate of NN classifiers (setup#1)

<table>
<thead>
<tr>
<th>System description</th>
<th>Supervised</th>
<th>Unsupervised (word-conf)</th>
<th>Unsupervised (utter-conf)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>67.2%</td>
<td>60.9%</td>
<td>62.6%</td>
</tr>
</tbody>
</table>

After completing the unsupervised MLP training, we adapt the NN using the limited manually labelled data to refine the performance. System results with various MLP features are shown in Table 5 (using the better word-confidence in data selection). Line 1 shows the traditional supervised trained MLP system. Unsupervised trained MLP feature is effective with better performance than supervised one. The adapted MLP gets better results, and obtains abstract 4% improvement over the baseline PLP-GMM-HMM low-resource system.

Table 5. Performance of MLP-based unsupervised training (setup#1)

<table>
<thead>
<tr>
<th>System description</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised MLP using 9frame spliced PLP as input</td>
<td>71.2%</td>
</tr>
<tr>
<td>Unsupervised training MLP-GMM-HMM</td>
<td>69.7%</td>
</tr>
<tr>
<td>Unsupervised training + adapted MLP-GMM-HMM</td>
<td>68.2%</td>
</tr>
</tbody>
</table>

5.4. Two-stage combined final unsupervised training system

Following the Algorithm I described in section 4, we apply the MLP-HMM two-stage combined unsupervised training and build our final systems in the low-resource scenario. Table 6 presents the system performances for both two experimental setups, including baseline systems, proposed single-stage (MLP feature or HMM model) unsupervised systems and the two-stage combined unsupervised training system. We can see that HMM or MLP unsupervised training obtains significant improvement individually. Moreover the gains from different methods are additive, and the final two-stage combined strategy has the best result with a 12% relative WER improvement compared to the original baselines in both setups.

Table 6. Performance comparison of proposed unsupervised training strategies (WER)

<table>
<thead>
<tr>
<th>System description</th>
<th>setup#1</th>
<th>setup#2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline supervised PLP-GMM-HMM</td>
<td>72.6%</td>
<td>55.2%</td>
</tr>
<tr>
<td>Paper proposed HMM-based unsupervised training system</td>
<td>66.2%</td>
<td>50.9%</td>
</tr>
<tr>
<td>Paper proposed MLP-based unsupervised training system</td>
<td>68.2%</td>
<td>51.6%</td>
</tr>
<tr>
<td>Paper proposed final two stage combined unsupervised training system</td>
<td>64.3%</td>
<td>48.6%</td>
</tr>
</tbody>
</table>

6. Conclusions

In this paper, we have presented ideas and experimental results for using unsupervised training in the low-resource speech application where untranscribed data is plentiful. The work focusing on the unsupervised training in [7-8] selects data only considering the confidence score, and most reported work till now do the unsupervised training only on the HMM acoustic model [7-10]. The work presented here is based on the conventional HMM-based unsupervised training [7-10, 13], using a new data selection principle based on confidence score and phone frequency. In addition, we introduce an unsupervised method on the neural network for MLP feature extraction, which is not considered in the earlier studies, to obtain more robust and improved MLP features. Experiments show that the level selection of confidence score is crucial in both kinds of training: utterance level confidence score is more applicable for HMM modeling due to having more contextual information, while the word level one is more useful for the NN unsupervised training as the more local accurate hypothesis.

The final proposed two-stage combined unsupervised training strategy incorporates distinct advantages from both feature and acoustic model refinements, giving a significant improvement of relative 12% versus the baseline supervised system. Experimental results also demonstrated that using such a suggested unsupervised training strategy can significantly reduce the manual effort of transcribing speech data in the low-resource scenario.

7. Acknowledgements

This work was supported by the Project 2009BAH41B00 and 2009BAH41B01 of National Science and Technology Pillar Program of China, the Project 90920302 and 61273268 of NSFC (National Natural Science Foundation of China), and the Project 60931160443 of NSFC and RGC.
8. References


