ABSTRACT

Is it possible to use out-of-domain acoustic training data to improve a speech recognizer’s performance on a specific, independent application? In our experiments, we use Wallstreet Journal (WSJ) data to train a recognizer, which is adapted and evaluated in the Phonebook domain. Apart from their common language (US English), the two corpora differ in many important respects: microphone vs. telephone channel, continuous speech vs. isolated words, mismatch in speaking rate. This paper deals with two questions. First, starting from the WSJ-trained recognizer, how much adaptation data (taken from the Phonebook training corpus) is necessary to achieve a reasonable recognition performance in spite of the high degree of mismatch? Second, is it possible to improve the recognition performance of a Phonebook-trained baseline acoustic model by using additional out-of-domain training data? The paper describes the adaptation and normalization techniques used to bridge the mismatch between the two corpora.

1. INTRODUCTION

A recurring problem in speech recognition is the lack of sufficient training material. Although general acoustic data for a given language is usually available, the special characteristics of the actual recognition task at hand make it often necessary to collect a substantial amount of task-specific material.

This paper deals with the question how task-independent speech material can be utilized to train a recognizer for an application with mismatching characteristics. To be more specific, we do recognition experiments within the Phonebook domain assuming that only minutes or a few hours of domain-specific training data are available (see [2] for a description of the Phonebook data and their definition of training and test subcorpora). In the experiments, we will use only a part of the available training data to adapt a recognizer that has been trained on totally different material, in our case WSJ data. WSJ qualifies for this task for mainly two simple reasons: it represents a relatively large speech corpus of good acoustic quality and it contains US English as does the Phonebook target application. On the other hand, both corpora show a substantial mismatch in several important respects:

- telephone (Phonebook) vs. microphone (WSJ) channel
- mismatch in speaking rate: 13 phonemes/s (WSJ) vs. about 9 phonemes/s (Phonebook)
- isolated words (Phonebook) vs. continuous speech (WSJ)

Our motivation for using WSJ as a training corpus in spite of these kinds of mismatch conditions can be summarized as follows: if it is possible to train a specific recognizer with domain-independent acoustic data, the need for collecting task-specific training data can be greatly reduced, which in turn cuts down on the time and costs for developing new speech recognition applications.

In this context we were interested in two questions:

1. How well performs a recognizer trained on the WSJ corpus when using just a few minutes of Phonebook training data for adaptation? This reflects the case that training is performed on some independent data and almost no domain-specific data is available to adapt the recognizer to the given task.

2. Given a large corpus of domain-independent training material (WSJ0 or WSJ0+1) and the complete Phonebook training material for domain adaptation, is it possible to improve the recognition performance as compared to a baseline recognizer trained exclusively on the Phonebook training material?

This paper was influenced by other work on speaking rate normalization [6] and on the recognition of isolated words with continuous speech recognizers [1].

In the following sections we first comment on the involved corpora. After explaining the applied adaptation and normalization techniques we present some experimental results. In all experiments, we used a recognizer applying MFCC signal analysis with cepstral mean subtraction and LDA to produce 36-dimensional feature vectors. The acoustic model is a Gaussian mixture model with a globally tied diagonal covariance matrix. The training and test lexica were derived from the CMU lexicon.

2. THE CORPORA

2.1. Phonebook Corpus

The Phonebook corpus [3] is a collection of about 23 hours of US English telephone channel acoustic data. The
larger part of this corpus consists of isolated-word utterances based on a vocabulary of about 8000 words. The vocabulary has been selected out of a much larger lexicon (CMU lexicon) to construct a phonetically rich collection of words, each word having an individually unambiguous pronunciation within the United States.

In [2] a partitioning of the Phonebook corpus in training, crossvalidation and test sets has been proposed, which is adopted here. We use the “small” training set of 5 hours (or parts thereof) in the experiments. The test corpus has a vocabulary size of 600 words, which equals the test lexicon size for all the experiments. Training and test vocabulary are disjoint. In a baseline experiment, we achieved a word error rate of 3.80% on the test corpus, when training the acoustic model on the 5-hour Phonebook training set.

### 2.2. WSJ Corpora

Unless otherwise indicated, we used the WSJ0 corpus for training the acoustic model. The corpus contains about 15 hours of microphone-channel read newspaper texts. Since the above described Phonebook baseline system features a disjoint training and test vocabulary, we removed all the sentences from the WSJ0 training corpus that contain words of the Phonebook test vocabulary. This reduces the size of the WSJ0 corpus by about 12%, but it makes sure that in the comparison of the Phonebook baseline system and the WSJ0-trained recognizer there is no hidden bias in favour of the latter system.

In section 4 we also describe experiments using the 80-hour WSJ0+1 corpus for training. Here, we are interested in the optimum recognition performance utilizing a large amount of task-independent training data. Therefore, the complete corpus including sentences with Phonebook test words has been used.

In all cases the WSJ acoustic data had to be downsampled to 8kHz to fit the specifications of the Phonebook corpus.

### 3. TECHNIQUES

This section deals with the techniques applied to bridge the mismatch between the two involved corpora. The techniques include adaptation and normalization. The issue of channel mismatch and differing speaking rates are investigated separately.

No special measures have been taken to handle the mismatch of continuous speech in the training data (WSJ) and isolated words in the test data (Phonebook). The article [1] suggests to augment the continuous-speech training material with isolated words to better train phonemes at word boundaries and thus improve recognition performance on isolated words. In several preliminary experiments mixing the WSJ and Phonebook training data in one single training corpus did not prove helpful. Probably, the characteristics of the two corpora are too different to handle them in a unified way. Instead, we used the WSJ material for training and Phonebook data for adaptation. This approach also makes it easier to control the relative influence of both corpora on the final acoustic model. Using the bigger WSJ corpus for training makes it possible to generate a large number of context-dependent phoneme models. Adaptation with domain-specific data is used to transform this fine-grained model for the application domain.

#### 3.1. Channel Mismatch

As mentioned in the introduction, we applied cepstral mean subtraction during signal analysis. Since this linear normalization technique is definitely not sufficient to eliminate the channel mismatch between WSJ and Phonebook, we additionally performed supervised MAP and MLLR adaptation of the WSJ-trained acoustic model. The adaptation data was taken out of the 5-hour Phonebook training corpus. The optimum value for the MAP adaptation parameter $\alpha$ ranges from 0.1 to 0.4 dependent on the amount of adaptation data. Monophone-specific MLLR regression classes have been used.

The table below presents Phonebook word error rates (WER) depending on the amount of Phonebook training material used for adaptation. The last column lists experiments applying linear discriminant analysis (LDA).

Not surprisingly, LDA without adaptation increases the influence of the corpus mismatch on the recognition performance. However, with 10 seconds of adaptation data the recognizer using LDA is already superior. Note that especially with LDA the recognition performance rapidly improve when increasing the adaptation material. For example, using only 3 minutes of Phonebook material a WER of 8% can be achieved.

<table>
<thead>
<tr>
<th>Data</th>
<th>MLLR, MAP</th>
<th>MLLR, MAP, LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 sec</td>
<td>25.6%</td>
<td>87.6%</td>
</tr>
<tr>
<td>10 sec</td>
<td>13.26%</td>
<td>10.23%</td>
</tr>
<tr>
<td>30 sec</td>
<td>12.73%</td>
<td>9.52%</td>
</tr>
<tr>
<td>1 min</td>
<td>10.88%</td>
<td>8.34%</td>
</tr>
<tr>
<td>5 min</td>
<td>9.53%</td>
<td>7.97%</td>
</tr>
<tr>
<td>10 min</td>
<td>8.88%</td>
<td>7.26%</td>
</tr>
<tr>
<td>30 min</td>
<td>7.62%</td>
<td>6.00%</td>
</tr>
<tr>
<td>60 min</td>
<td>7.11%</td>
<td>5.34%</td>
</tr>
</tbody>
</table>

Table 1: WER for the phonebook test corpus without adaptation (first line) and with various amounts of Phonebook adaptation data using MLLR and MAP with and without LDA.

#### 3.2. Speaking Rate

We measure the speaking rate for an utterance as its average phoneme length, which is calculated as the number of speech frames divided by the number of phonemes
within the utterance. Figure 1 presents the speaking rate distribution for the Phonebook training corpus (5 hours), the Phonebook test corpus and the WSJ0 corpus, based on a frame shift of 10 ms. The figure shows clearly the large difference of the mean values for the WSJ0 and Phonebook phoneme length. The mean phoneme length is 7.7 frames for the WSJ0 corpus. For the Phonebook training and test corpus the mean phoneme length is 11.0 and 11.1 frames, respectively.

It can also be seen that the Phonebook corpus shows a much higher variance in the speaking rate. However, this is mainly due to the short duration of individual utterances (single words), whereas for the WSJ corpus the utterances are whole sentences and thus the phoneme length is averaged over complete sentences.

![Phoneme length histograms for Phonebook and WSJ0 corpus](image)

**Figure 1:** Phoneme length histograms for Phonebook and WSJ0 corpus. The x-axis denotes the utterance-wise mean phoneme length (measured in frames, frame shift: 10 msec)

Since this substantial speaking rate mismatch can be expected to impair recognition performance, we developed an algorithm to perform speaking rate normalization (SRN). Similar to an algorithm described in [6], we apply a modified acoustic feature extraction that adjusts the frame shift parameter to normalize the average number of frames per phoneme. The algorithm described here additionally involves speaking rate normalization for the adaptation data. The normalization takes place in six steps:

1. Train the acoustic model on the unnormalized task-independent data (i.e. WSJ). The standard frame shift value $S$ is used for feature extraction.
2. Use forced alignment to calculate the average number $F_{i}^{train}$ of frames per phoneme for the complete training corpus.
3. Using the same acoustic model, perform a forced alignment of the adaptation data (Phonebook training data) to calculate the average number of frames $F_{i}^{adapt}$ per phoneme for each utterance $i$ in the adaptation corpus.
4. Redo feature extraction for the adaptation corpus using frame shift
   \[
   S_i = S \frac{F_{i}^{adapt}}{F_{i}^{train}} \quad (1)
   \]
   and adapt the acoustic model using the thus normalized adaptation utterances.
5. For each utterance $j$ in the (Phonebook) test corpus, do a first recognition with the standard frame shift $S$ to calculate the speaking rate measure $F_{j}^{test}$.
6. Do a final recognition run on the test utterances using the normalizing frame shift
   \[
   S_j = S \frac{F_{j}^{test}}{F_{j}^{train}} \quad (2)
   \]

Note that the application of SRN not only brings the Phonebook speaking rates in line with the WSJ speaking rate, but it also normalizes the varying speaking rates among the individual Phonebook utterances.

The next table shows the influence of speaking rate normalization on the word error rate. Again, the experiments have been conducted for various amounts of adaptation data, as indicated in the first column. The last column shows that the relative WER improvement gained through SRN is generally about 20%. Continuing the above example, the WER for 3 minutes of Phonebook adaptation material goes down to 6.18%.

<table>
<thead>
<tr>
<th>adapt</th>
<th>no SRN</th>
<th>SRN</th>
<th>rel. improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 sec</td>
<td>87.04%</td>
<td>53.33%</td>
<td>33%</td>
</tr>
<tr>
<td>10 sec</td>
<td>10.23%</td>
<td>8.11%</td>
<td>21%</td>
</tr>
<tr>
<td>30 sec</td>
<td>9.52%</td>
<td>7.41%</td>
<td>22%</td>
</tr>
<tr>
<td>1 min</td>
<td>8.34%</td>
<td>6.40%</td>
<td>23%</td>
</tr>
<tr>
<td>3 min</td>
<td>7.97%</td>
<td>6.18%</td>
<td>22%</td>
</tr>
<tr>
<td>10 min</td>
<td>7.26%</td>
<td>5.75%</td>
<td>21%</td>
</tr>
<tr>
<td>30 min</td>
<td>6.00%</td>
<td>5.02%</td>
<td>16%</td>
</tr>
<tr>
<td>60 min</td>
<td>5.34%</td>
<td>4.55%</td>
<td>15%</td>
</tr>
<tr>
<td>300 min</td>
<td>4.47%</td>
<td>3.33%</td>
<td>26%</td>
</tr>
</tbody>
</table>

**Table 2:** Impact of speaking rate normalization on the Phonebook WER. All experiments with MAP and MLLR adaptation and with LDA.

Figure 2 summarizes the adaptation and normalization experiments. Obviously, the WSJ model adapted using the complete Phonebook training corpus outperforms the Phonebook baseline model described in section 2.1. This proves that not only can task-independent data be used to reduce the need for collecting task-specific data. It also helps to improve recognition performance even if a substantial amount of task-specific data is available.

## 4. WSJ0+1 TRAINING DATA

In order to find out, if the WER improvements gained by utilizing the WSJ0 acoustic data can be further increased, we did experiments using the WSJ0+1 corpus instead. This enlarges the amount of task-independent training data by a factor of 5. The table compares the results achieved with and without speaking rate normalization. In both experiments, the 5-hour Phonebook training corpus has been used for adaptation.
Figure 2: Summary of Phonebook experiments using WSJ0 for training. Note the performance of the Phonebook baseline system, indicated by the isolated data point '+' at (18000 seconds, 3.80% WER).

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>rel. Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>no SRN</td>
<td>3.55%</td>
<td>7%</td>
</tr>
<tr>
<td>SRN</td>
<td>2.97%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 3:Recognizer trained on WSJ0+1 and adapted on the 5-hour Phonebook corpus. The improvements are calculated in relation to the baseline WER of 3.80%.

5. CONCLUSION

The experiments show that based on a WSJ-trained acoustic model a few minutes of Phonebook data are sufficient to achieve a reasonably good recognition performance on the Phonebook test data: as an example, a word error rate of 6.18% has been achieved using only 3 minutes of Phonebook training data. This compares to a baseline word error rate of 3.80% for a recognizer trained exclusively on the 5-hour Phonebook training set and demonstrates that the techniques of acoustic adaptation and speaking rate normalization can be powerful enough to make the extensive collection of domain-specific training data unnecessary.

In further experiments, the complete Phonebook training data has been used to adapt an acoustic model trained on WSJ0+1. As a result, a word error rate of 2.97% was measured for the Phonebook test data. This is a 22% improvement compared to the Phonebook baseline system, which was directly trained on the Phonebook data. The experiments prove that the utilization of task-independent training data can lead to improved recognition performance, even if a substantial amount of task-specific training data is available.

6. REFERENCES


