SEMI-SUPERVISED ADAPTATION OF ACOUSTIC MODELS FOR LARGE-VOLUME DICTATION

Colin W. Wightman  
Department of Computer Science  
Minnesota State University, Mankato  
Mankato, MN 56082 USA  
colin.wightman@mankato.msus.edu

Ted A. Harder  
Department of Electrical Engineering  
Duke University  
Durham, NC 27708 USA  
tah@ee.duke.edu

ABSTRACT

Using a Large-Vocabulary, Continuous Speech Recognizer in a high-volume application such as a commercial transcription service presents a different set of challenges and constraints than in a laboratory setting. We examine these differences with regard to acoustic model adaptation and find serious shortcomings in both the supervised and unsupervised approaches. We then examine a new method, semi-supervised adaptation, which overcomes the limitations of the other methods and can reduce recognition error rates by as much as 15% more than the reduction obtained through unsupervised adaptation.

1. INTRODUCTION

Speaker adaptation has become an important method for improving the recognition accuracy of Large-Vocabulary, Continuous Speech Recognition (LVCSR) systems. Although a great deal of research has been focused on the adaptation of acoustic models [c.f., 1, 2, 4, 6], the techniques developed to date can be broadly classified into two categories: (1) supervised methods which utilize both speech samples from the new speaker and associated transcriptions of those samples, and (2) unsupervised methods which utilize only the speech samples. In the context of a large-volume transcription operation, however, both of these approaches become undesirable because of differences between the laboratory and commercial environments.

In Section 2, we briefly describe a commercial, large-volume transcription application and examine the constraints and challenges that such an environment presents. In Section 3, we examine the ways in which both supervised and unsupervised adaptation fail to meet the challenges of this commercial environment and propose a hybrid adaptation method which better meets these requirements. We then present the results of experimental comparisons between the various methods before summarizing our conclusions in Section 4.

2. HIGH-VOLUME TRANSCRIPTION

Figure 1 shows a block diagram of a high-volume transcription system. In this system, human transcriptionists are needed to edit the recognition output both to correct recognition errors and to format and rephrase the dictation so that the report contains only polished text. In this environment, the authors producing the dictation are dictating off-line and most are not aware of the use of speech recognition in the transcription process. The structure of this system, and its commercial nature, produces some particular characteristics that are not generally addressed by laboratory systems:

- Utterances are very long: a single dictation is typically a minute or two in length while 10-minute utterances are not uncommon.
- Recordings are not of high-quality: typically speech is recorded on telephone-grade equipment and digitized at 64-kbps.

![Figure 1. Block diagram of a commercial transcription operation.](image-url)
• Speech contains numerous filled pauses, word fragments, restarts, and other speech errors.
• Background noises are often present and can include other conversations, PA announcements, paper rustles, etc.
• Many parts of the dictation do not appear in the final report. These parts may be duplicating information (such as the date) which is captured elsewhere, or may be verbal instructions to the transcriptionist (e.g., “change that to ‘left’ ”).
• Data is cheap, people are not. With hundreds of hours of dictation coming through the system each week, making more data available for adaptation is straightforward. Consuming human resources to process that data, however, is expensive.
• Total human effort is the relevant evaluation metric. If adaptation requires more human effort than will be saved by improved recognition accuracy over a few weeks, then it is too costly to justify.

3. ADAPTATION METHODS
Adaptation seeks to modify a speaker-independent set of acoustic models by using a limited amount of speech from the new speaker, along with phone-level labels for that speech, to improve the match between the acoustic models and the new speaker. In our experiments, the basic adaptation method used was the maximum likelihood linear regression (MLLR) technique described in [3]. This method, which estimates a set of linear transformations to map the model parameters, requires not only sample speech from the new speaker, but phone-level labels for that speech as well. The fundamental difference between the three adaptation methods described here is in how those labels are obtained.

3.1 Supervised Adaptation
In supervised adaptation, the phone-level labels for the speech samples are obtained by manually transcribing each sample. To produce accurate labels, these transcriptions must accurately reflect all speech sounds including word fragments, hesitations, stutters, filled pauses, etc. Alternatively, the speech samples can be edited to remove these phenomena and the remainder transcribed. Preparing these labels, whether transcribing or excising disfluencies, requires a considerable expenditure of human effort. Moreover, the process requires some specialized skills either in phonetic transcription, or in using a waveform editor to select only fluent regions.

In a commercial environment, this is simply too high a price to pay. Even with specialized tools and training, producing accurate labels for even a modest fifteen minutes of speech requires several hours of a person’s time. An expense that is difficult to justify based on the few minutes of transcription time that will be saved by the improved recognition accuracy.

3.2 Unsupervised Adaptation
Unsupervised adaptation seeks to avoid the high cost of preparing accurate labels for the speech samples by using the existing speaker-independent models to do recognition and thus produce the required labels automatically. The price for obtaining these labels automatically, however, is that they will contain some errors.

In laboratory experiments, the recognition errors that occur when generating the labels to be used for adaptation tend to be minor in that, when phones are mis-recognized, the erroneous label is often of the same broad phonetic class as the correct label. Thus, although the labels will contain some errors, the effect of the errors is usually small and can be averaged out by increasing the amount of speech used for adaptation.

In our commercial transcription system however, the limited recording bandwidth, background noises, and the multitudes of speech errors can produce a large number of recognition errors. More problematically, these errors are strongly clustered: there are regions where the recognition is extremely good, interspersed by regions of very poor accuracy. The crucial point is that these regions of poor accuracy often do not result from poor acoustic modeling (which should be addressed by using them to adapt the models), but from other sources such as background noises and conversations, out-of-vocabulary words, etc. If we simply collect speech samples from the new speaker without reviewing them in some way, there is a good chance that the collected samples will not accurately represent the new speaker and the resulting adapted models may not perform as well as they might if these regions of poor accuracy resulting from other sources could be excluded. On the other hand, having a human listener screen the samples to select portions which are representative of the new speaker’s speech, while less costly than preparing labels for supervised adaptation, is still too costly for a system in which dozens of new speakers may need to be added each week.

3.3 Semi-Supervised Adaptation
To reduce the large costs associated with using humans to either prepare labels for supervised adaptation, or to select appropriate regions of samples for unsupervised adaptation, we propose a new, hybrid approach: semi-supervised adaptation. This approach draws upon the completed reports generated by the human transcriptionists. Because these reports are the product of the transcription operation, there is no additional human cost associated with using them as part of our adaptation process.

As noted above, these reports are not exact transcriptions of the dictations. Instead, they are formatted, polished prose based on the dictation. Nevertheless, although there are many parts of the dictation that do not appear in the report, there are many short sections where the text of the report is a verbatim transcription.

Semi-supervised adaptation thus proceeds by using the speaker-independent recognizer to produce a set of word-level labels. These labels are then optimally matched to the report text using dynamic programming. Words that match between the
recognizer output and the report text are thus identified as words that the recognizer has gotten correct. By extracting only these words which are known to be correct, and the corresponding segments of the speech waveforms, we can collect a set of speech samples and word-level labels which can then be used as input to the supervised adaptation process.

There are two concerns raised by this approach: (1) it is wasteful of data, and (2) we are only using speech that the recognizer is already getting right and may thus not be adapting the acoustic models as well as classic supervised adaptation. To address the first concern: data is cheap in the commercial. Even though only 20-50% of the speech may be correctly recognized and appear in the finished report, collecting more speech so that the total amount available for adaptation is sufficient, does not present a problem because, as noted above, data is cheap, people are not. The second concern, regarding the potential performance gain available from this method, must be addressed experimentally and we present those results in the following section.

### 3.4 Experimental Evaluation

To evaluate the effectiveness of the semi-supervised adaptation technique, we conducted a series of tests using a LVCSR system based on the system developed at Cambridge University [5]. The baseline speaker-independent (SI) models were state-clustered, crossword tri-phone models trained on approximately 9 hours of speech from 18 different speakers. The language domain used was family medicine, and a tri-gram language model was trained for a twenty thousand-word vocabulary using previously prepared reports containing approximately ten million words.

We selected a test set of seven speakers that had not been used in developing the SI model set, and collected at least 30 minutes of speech for each of them for use as adaptation data. For each of the seven speakers, a separate set of speech samples containing approximately 300 words was prepared for use as test data. The test data was carefully selected to include fluent speech with only minimal errors and background noises so that accurate baseline transcriptions could be prepared. Finally, true word-level labels corresponding to the adaptation data were manually prepared for three of the speakers.

For each speaker, a recognition baseline result was obtained using the SI models. Adapted models were then generated using the unsupervised and semi-supervised adaptation methods. The resulting models were evaluated using the same recognition test data, and the results are presented in Table 1.

Inspection of Table 1 reveals that both adaptation methods produced increased recognition accuracy in all cases but one (speaker 4). It is also significant that the semi-supervised models outperformed those generated by unsupervised adaptation for all but one speaker (speaker 3). Although the performance advantage of the semi-supervised models is not great for some speakers, it is quite large for others (e.g., speaker 6). These findings become more evident when the results are viewed in terms of changes in the recognition error rates as shown in Table 2.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>SI Baseline</th>
<th>Unsupervised</th>
<th>Semi-Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70.82</td>
<td>73.94</td>
<td>75.35</td>
</tr>
<tr>
<td>2</td>
<td>39.39</td>
<td>42.71</td>
<td>43.73</td>
</tr>
<tr>
<td>3</td>
<td>42.58</td>
<td>50.09</td>
<td>47.41</td>
</tr>
<tr>
<td>4</td>
<td>68.86</td>
<td>67.07</td>
<td>67.37</td>
</tr>
<tr>
<td>5</td>
<td>64.78</td>
<td>73.15</td>
<td>73.89</td>
</tr>
<tr>
<td>6</td>
<td>43.42</td>
<td>52.31</td>
<td>62.28</td>
</tr>
<tr>
<td>7</td>
<td>52.81</td>
<td>57.19</td>
<td>60.63</td>
</tr>
</tbody>
</table>

Table 1. Recognition accuracy for test speakers showing results obtained with SI models as well as models obtained through unsupervised and semi-supervised adaptation.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Unsupervised</th>
<th>Semi-Supervised</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.69</td>
<td>15.52</td>
<td>4.83</td>
</tr>
<tr>
<td>2</td>
<td>5.48</td>
<td>7.16</td>
<td>1.68</td>
</tr>
<tr>
<td>3</td>
<td>13.08</td>
<td>8.41</td>
<td>-4.67</td>
</tr>
<tr>
<td>4</td>
<td>-5.75</td>
<td>-4.78</td>
<td>0.96</td>
</tr>
<tr>
<td>5</td>
<td>23.76</td>
<td>25.87</td>
<td>2.10</td>
</tr>
<tr>
<td>6</td>
<td>15.71</td>
<td>33.33</td>
<td>17.62</td>
</tr>
<tr>
<td>7</td>
<td>9.28</td>
<td>16.57</td>
<td>7.29</td>
</tr>
</tbody>
</table>

Table 2. Percentage reduction in recognition error rate from baseline SI results for unsupervised and semi-supervised models. The difference between the two methods is shown in the last column, with a positive difference corresponding to a greater improvement from the semi-supervised models.

Further investigation revealed an important effect of the semi-supervised procedure. This effect is the result of the removal of portions of the speech samples that are likely to produce recognition errors for reasons other than poorly trained acoustic models (speech errors, background noises, etc.) The effect of these portions on the unsupervised models is detrimental, and removing these portions through the semi-supervised procedure yields better models, even though there is less speech left for adaptation. This effect is illustrated in Figure 2 which plots recognition accuracy against the number of minutes of speech used for adaptation for speaker 1.

From Figure 2, it is clear that the first 10 minutes or so of the speech collected for adaptation contains a sufficiently large proportion of sounds that are not representative of the speaker and that, as a result, the unsupervised adaptation significantly reduces the recognition accuracy. Because the semi-supervised method automatically excludes much of these sounds (for this speaker, the semi-supervised technique excluded an average of 47% of the data), the resulting models suffer much less performance loss. The latter portions of the adaptation data appear to be more representative of the speaker and both adaptation methods yield improved models as a result.
Nevertheless, the models produced by semi-supervised adaptation maintain their performance advantage for any amount of data, even exceeding the performance obtained through supervised adaptation (probably due to compromises made during the preparation of the supervised labels).

The semi-supervised method of adaptation was developed in response to the needs of a commercial transcription operation. It is applicable, however, to any application in which texts based on, but significantly different from, recorded speech inputs are available. Note that when the available texts are not significantly different from the recorded speech, supervised adaptation becomes possible without any further preparation of the data.

4. SUMMARY AND CONCLUSIONS

Large-scale commercial applications of speech recognition technology present new challenges and constraints. Different criteria for evaluating systems, and differences between laboratory and real-world recordings, account for some of the changes, while the need for robust scalability to handle large tasks accounts for others.

In this paper we have briefly outlined a commercial transcription operation utilizing speech recognition and examined the ways in which that operation differs from a typical laboratory system. We looked at the usual approaches to acoustic model adaptation, supervised and unsupervised, and found that both had serious shortcomings in the commercial environment. We then proposed a hybrid method, semi-supervised adaptation, and presented experimental results that indicate that the new method is generally no worse than unsupervised adaptation, and can be substantially better than even supervised adaptation without requiring human involvement in the selection and preparation of the adaptation data.

5. REFERENCES


