ADAPTATION AND COMPENSATION FOR SPEECH RECOGNITION – LEARNING FROM EXTRA DATA TO IMPROVE ROBUSTNESS

Chin-Hui Lee
Dialogue Systems Research Department
Bell Laboratories, Lucent Technologies
Murray Hill, NJ 07974, USA
chl@research.bell-labs.com

The performance of speech recognition algorithms often degrades drastically when the testing environment is somewhat different from the training conditions. Although many approaches have been proposed to cope with this mismatch problem, learning from extra data collected in operational conditions is always an attractive option if more processing is permitted. Supervised adaptation is the process of learning from an additional set of labeled adaptation data. On the other hand, unsupervised compensation is often referred to as learning from testing data directly. Adaptation and compensation share many similar principles and techniques. We examine some recent advances in this active research area and discuss their applications to improve recognition performance in hands-free communications.

1. INTRODUCTION
Recent advances in automatic speech recognition (ASR) are usually attributed to the use of statistical pattern recognition paradigm. Assuming the true joint distribution of a word sequence, $W$, and its corresponding sequence of acoustic vector observations, $X$, can be modeled by a true parametric probability density function $p(W, X) = p(X|W) \cdot P_T(W)$; and the full knowledge of the parameters $(\Lambda, \Gamma)$ of the above distributions is known, then an optimal decoder (speech recognizer) which achieves the expected minimum word error rate and gives the recognized string, $\hat{W}$, is the following maximum a posteriori (MAP) decoder:

$$\hat{W} = \arg\max_W p(W|X) = \arg\max_W p_A(X|W) \cdot P_T(W).$$

However, in practice, neither do we know the true parametric form of $p(W, X)$, nor do we have the knowledge about its true parameter values. Therefore, the above optimal speech recognizer is never realizable. Instead, the recognizer parameters are estimated from a large set of labeled speech and text training data. Even though this paradigm has worked well, in most real applications, there always exists some form of mismatch which causes a distortion between the trained models and the test data. A conceptual illustration is shown in Figure 1. $D_1$, $D_2$ and $D_3$ characterize the possible distortion in the signal, feature and model spaces, respectively. These mismatches may arise from inter- and intra-speaker variabilities; transducer, channel and other environmental variabilities; and many other phonetic and linguistic effects caused by mismatch in training and testing task definitions. In the case of hands-free ASR, the mismatch might even be more severe.

Figure 1: Mismatch between training and testing

Many recent studies have illustrated some of the potential difficulties [13, 14, 15, 19, 26]. Due to missing channel characterization in speakers and speaking environments in the current paradigm, the degradation is expected.

Traditional approaches to robust speech recognition include finding invariant or robust features and developing better modeling and learning techniques. In addition, three major classes of statistical techniques to improve ASR robustness can be defined:

- adapting recognition parameters with labeled data;
- compensating signal, feature and model distortion with only testing data;
- using robust decision strategies.

Adaptation and compensation are two closely related topics that have been intensively studied. Many techniques that were originally developed for adaptation can be extended to compensation and vice versa (e.g., [16, 17]). Robust decision is a rather new research area (e.g., [12]).

2. MOTIVATION FOR LEARNING MORE

In the following we show two sets of results to characterize some of problems in mismatch especially in hands-free speech recognition of data recorded in a moving vehicle and how adaptation and compensation could potentially be used and combined to improve the performance.

2.1. Training and Testing Mismatches

The impact of mismatch in training and testing conditions on connected digit recognition performance with testing data recorded in a moving vehicle is illustrated first. Three training conditions were used:

- TrainCT: 6249 digit utterances recorded with close-talking microphone;
- TrainTel: over 50,000 digit utterances from 21 digit corpora recorded in landline and wireless channels;
Table 1: SER in various training and testing conditions

<table>
<thead>
<tr>
<th></th>
<th>TrainCT</th>
<th>TrainTel</th>
<th>TrainB16</th>
</tr>
</thead>
<tbody>
<tr>
<td>TestCT</td>
<td>6.5</td>
<td>8.8</td>
<td>-</td>
</tr>
<tr>
<td>TestM8</td>
<td>16.5</td>
<td>20.7</td>
<td>-</td>
</tr>
<tr>
<td>TestB16</td>
<td>-</td>
<td>13.2</td>
<td>9.5</td>
</tr>
</tbody>
</table>

**TrainB16** 6249 digit utterances from each channel recorded with a distant linear array of 16 microphones, processed by delay-and-sum beamforming.

Using the models trained from the above three conditions, we evaluated them on testing utterances recorded in the following three conditions:

**TestCT** 1720 digit utterances recorded with close-talking microphone;

**TestM8** 1720 digit utterances recorded on the channel 8 of the linear array of 16 microphones;

**TestB16** 1720 digit utterances from each channel recorded with a distant linear array of 16 microphones, processed by delay-and-sum beamforming.

The string error rates (SERs) in various training and testing combinations are shown in Table 1 [15]. It is interesting to note that the testing results with the close talking microphone were good even in the mismatch conditions. However, when testing on data recorded from one of the distant microphones, the string error rates increased by more than 100%. When a delay-and-sum beamforming technique is used to process the testing data, the SER is reduced to 13.2% from 20.7% when using only one of the array elements. The error rate can further be reduced if the training condition matched the testing condition at 9.5% SER. But this performance is still much worse than that in the matching conditions when close-talking microphone is used, i.e., at 6.5% SER. It is obvious that this is a wide performance gap to be bridged when distant microphones are used in hands-free speech recognition.

**2.2. Combining Adaptation and Compensation**

In the next experiment, we examined ways to combine batch supervised adaptation typically used for fast enrollment of new speakers with on-line unsupervised compensation in order to raise recognition rates to a level high enough for practical usage. New data recorded for the Resource Management task were used for adaptation and testing. This combined adaptation process was carried out in two steps:

**Step 1.** Supervised adaptation to generate seed models;

**Step 2.** Unsupervised compensation using the test data.

The word error rates (WERs, averaged over the five speakers each speaking 75 testing utterances) are listed in Table 2 [23]. The number of utterances used in Step 1 varied from 1 to 300. Five non-native speakers, recorded two sets of stereo data, one using a close-talking microphone (labeled MIC) and the other using a telephone handset through dial-up lines (labeled TEL). Supervised adaptation was carried out using the MIC adaptation data (labeled as SUP in Table 2) and unsupervised compensation and recognition were done for both MIC and TEL test sets (labeled as TEST).

The values listed in the column labeled S1 are the WERs with Step 1 only and the values listed in the column labeled S2 were obtained after Step 2. Although adaptation improved only slightly when the acoustic conditions for the SUP data and that for the TEST data were similar, its effectiveness when acoustic conditions were different was clearly shown. For example, when the MIC set was used for SUP and the TEL data was used for TEST, the combined method required only three utterances to achieve 60% recognition accuracy, while the supervised adaptation (Step 1 only) needed 100 utterances. Scenario S2 is also shown effective when starting with not so well-adapted models (S1 with S1 or 1-10 utterances) while maintaining performance in better matched conditions (S1 with 25-300 utterances). Adaptation and compensation here is based on the structural maximum a posteriori (SMAP) algorithm which is designed to handle any data size including very small amount of data using hierarchical prior evolution [23]. Using SMAP on linear regression parameters (e.g., SAMPLR) has also been proposed and shown equally or more effective in small data sizes [24].

**3. TECHNIQUES AND STRATEGIES**

It is clear from the results in Table 1 shown earlier that it will be difficult to achieve the same level of performance as in the close-talking scenario with the same amount of training data in the distant-talking situation. It is also too expensive to collect a large amount of training each new acoustic condition, such as the TrainB16 and TestB16 conditions in order to match the accuracies obtained by well-trained models in the matched conditions, such as in the case of TrainTel in Table 1. A practical way is to start with a reasonable set of seed models, adapt to the new adverse condition with a small amount of adaptation data, then reduce the signal, feature and model mismatch through compensation with only the test data. The results and scenarios demonstrated in Table 2 seem promising.

Although the most effective way to handle mismatch seems to be finding invariant features so as to minimize the acoustic mismatch between training and testing environments. We have not yet discovered any useful feature that is invariant across all adverse conditions yet still discriminative and easy to model. To circumvent this difficulty, a straightforward solution is to collect additional training data in a specific testing condition and adapt the recognizer parameters accordingly. Compensation can then be applied. Of course robust decision strategies, such as minimax and Bayesian predictive classification algorithms (e.g., [12]), can always be applied after all the available techniques have been considered. In this study we will focus our discussion only on adaptation and compensation.
3.1. Classifier Parameter Adaptation

For HMM-based speech recognition systems, adaptation is usually accomplished in two ways [17]: (1) direct adaptation of the HMM parameters; and (2) indirect adaptation of a set of transformation parameters which induces the adapted HMM parameters through transformation. Bayesian parameter learning is the dominant approach to direct or local adaptation (e.g. [8]). It provides an optimal mathematical framework for combining information in a general set of stochastic models and a specific set of adaptation data. It also has the nice asymptotic property that the more adaptation data we use the better the recognition performance we achieve. It is often combined with other techniques to improve its performance especially for adaptation of large models with only a small set of adaptation utterances [17].

Another way is to impose a global constraint over all the model units so that all the parameters can be adapted at the same time even though not all the units have been observed. The most popular technique assumes the adapted models are obtained through a linear transformation of the original models. Known as maximum likelihood linear regression (MLLR) [18], the approach works quite well especially in situations when only a limited set of adaptation data is used. However, the performance often saturates quickly. Another potential limitation of MLLR is a need to determine the number of regression matrices. Structural Bayesian adaptation based on automatic selection of tree structures has been effectively applied to improve estimation efficiency [23, 24]. In addition there is a great potential for a unified Bayesian framework to jointly adapt both the HMM and transformation parameters. Some simplified versions have recently been proposed (e.g. [5, 17]).

Once a recognition system has been designed, it can further be improved based on a dynamic strategy such that new knowledge and information are acquired and incorporated incrementally during the development and use of the ASR system. This adaptive learning algorithm is often referred to as on-line Bayesian adaptive learning (e.g. [3, 11, 28]). It is an important tool for improving compensation.

3.2. Feature and Model Compensation

Compensation can be considered as a form of unsupervised adaptation in which only the testing data are used. Many other names have also been adopted, e.g. self adaptation, auto adaptation or instantaneous adaptation [22, 25, 29]. For robust speech recognition, compensation can be accomplished in the signal, feature and model spaces in order to reduce the distortions shown in Figure 1. The readers are referred to a recent review on the topic of feature and model compensation [16].

One of the early studies on feature compensation is cepstral mean subtraction (CMS) which removes the cepstral mean of each utterance before training and testing and was shown to be robust to microphone and channel distortion in many systems. By making CMS more effective for different sounds in different speaking conditions, codeword-dependent cepstral normalization (CDCN) [1] was then developed. A simplified version using a VQ codebook to compensate cepstral difference with a set of bias was shown to be effective for telephone speech recognition [21]. A natural extension is to use the information embedded in the acoustic HMMs to aid the feature compensation process. Although model-based feature compensation is effective in some situations, there are many types of distortion that can not easily be realized by a simple feature transformation. This is often implemented by introducing some structure to reduce the number of parameters. We will discuss this point and some related recent advances in the following.

3.3. Recent Advances in Structural Compensation

Stochastic matching (SM) [22] is a mathematical framework to perform feature and model compensation to minimize the mismatch between training and testing conditions. The distortion functions shown in Figure 1 often approximates some form of missing structural characterization during modeling. Non-linear SM based on neural network approximation of these distortion functions has been proposed [25]. Neural networks can also be used to estimate the interactions between the original HMMs and the noise to improve performance [7]. SNR levels heavily affect the estimation of compensation parameters and need to be properly handled in noisy speech recognition (e.g. [9]). It can also be used to guide stochastic matching to converge to the right set of models among a collection of SNR-dependent HMMs [10]. Another useful distortion function for SM that has shown some potential is the vector Taylor series (VTS) expansion approximation [20]. It has been combined with an optimal selection of forgetting factors [11] to perform sequential noise estimation and compensation [2].

The use of structure to aid unsupervised adaptation started with the pioneering work by Furui [6]. This hierarchical structure has been used recently to perform compensation [13] in which a collection of compensation mean vectors is used to perform stochastic matching. Similar idea has also been shown useful [4]. It has also been applied to SMAP and SMAPLR [23, 24] using tree-based hierarchical priors such that the priors corresponding to children nodes in the tree is derived from the parent node making it possible to specify all the priors for all the parameters in a large collection of HMMs to perform efficient and effective adaptation and compensation. This specification of correlations between HMM parameters is critical for many future robustness studies. Selection of high confidence segments to improve the effectiveness of compensation is another important topic for making use of as much the limited amount of quality data during testing [27].

4. SUMMARY

We have discussed adaptation and compensation approaches to improving recognition performance. There is a close relationship between these two techniques. Two classes of parameters and related adaptation algorithms can be used, namely direct HMM parameter adaptation via Bayesian learning, and indirect HMM parameter adaptation through structural estimation. Both of these can be applied to unsupervised self-adaptation or compensation. When combined with Bayesian adaptation, transformation parameter adaptation shows both an effective efficiency (for short adaptation data) and a good asymptotic property (e.g. [23, 24]).

We expect this hybrid adaptation and compensation approach together with the discovery of structural models to characterize the interaction between speech and environments to play key roles in future robust ASR research. By combining compensation techniques and on-line unsupervised adaptation of HMM and structure parameters, new
classes of adaptive compensation algorithms are expected to emerge. However, simply relying on the trained model set and the structure constraints is likely to limit our capability. We need to learn more from the given large set of training set by extracting useful auxiliary information and make use of it during adaptation and compensation. By doing so, it will help with enhancing robustness and improving recognition performance.

5. REFERENCES