Effect of Acoustic Adaptation for A Distant-talking Interactive TV System

Jing Huang1, Mark Epstein1, Marco Matassoni2

1 IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A.
2 Fondazione Bruno Kessler
via Sommarive, 18 - 38050 Trento, Italy
jghg, meps@us.ibm.com, matasso@fbk.eu

Abstract

In this paper we have studied how to adapt a close-talking baseline acoustic model to a distant-talking application developed in an interactive TV dialogue system: distant-talking interfaces for control of interactive TV (DICIT) project. We have shown that in order to have effective adaptation from the out-of-domain data it is better to acquire that data in the same DICIT environment than using contaminated data. By measuring grammar error rate (GER) and action classification error rate (AER) in addition to word error rate (WER), we have shown the best way to adapt the baseline model using available out-of-domain adaptation data (TIMIT) and small amount of in-domain (DICIT) adaptation data. The best approach is to use cascading MAP adaptation. With less than 5 hours of out-of-domain data and 1 hour of in-domain data, the cascading MAP improves WER/GER/AER by 17%/18%/16% relative respectively over the baseline model. The experimental results show that in-domain adaptation data is definitely needed to improve GER and AER.

Index Terms: acoustic model adaptation, distant-talking speech recognition, dialog system

1. Introduction

The DICIT (Distant-talking Interfaces for Control of Interactive TV) project aims to integrate distant-talking voice interaction with a set-top box (STB) as an alternative way to the use of remote control in interactive TV systems. In the targeted application scenario, the DICIT system recognizes freeform commands spoken by multiple users, even in the presence of background noise and TV surround audio propagated in the environment. The (possibly multichannel) sound produced by the TV itself needs to be compensated by acoustic echo cancellation techniques; similarly, the hands-free voice acquisition requires a proper signal enhancement to mitigate the environment disturbances (e.g. by means of microphone array beamforming). This certainly poses challenges to distant-talking automatic speech recognition (ASR) in DICIT, and the quality of the speech recognition would directly affect the performance of the following natural language understanding (NLU) and dialog system.

Although ASR research has been pursuing speaker independent, device/task independent acoustic modeling, the best recognition performance on a specific device for a specific task still comes from an acoustic model trained on tens or hundreds of hours of data recorded on this device for this task — often this data is referred to as in-domain speech data. Collecting the in-domain training data is costly, however. And every time we change a device or task, we have to repeat the process. At the same time, there exist large amount of out-of-domain data available, such as desktop dictation data and telephony data.

Therefore, for the DICIT project we investigate effective adaptation approaches to avoid the need to acquire large amounts of real speech interactions in the target scenario, which is hard to obtain when the design and integration of the whole application system is not ready in the first place. In particular, starting with the already available IBM Embedded ViaVoice (EVV) acoustic models, which have shown very good performance on in-car applications, our goal is to adapt them to the DICIT acoustic domain and task with minimum data collection effort. One possible way to obtain relevant quantities of data suitable for adaptation is the generation of “contaminated” data starting from available out-of-domain corpora [1]. To better match the acoustic characteristics of the real system front-end, another way of collecting data is to play back the same data and record it in the DICIT-like room.

In this study we evaluate the standard adaptation techniques on the two kinds of adaptation data mentioned above, such as maximum a posteriori (MAP) adaptation [2], maximum likelihood linear regression (MLLR) [3] and feature-space maximum likelihood linear regression (fMLLR) [4]. In a typical ASR experiment, word error rate (WER) is the most popular measure of performance. In the DICIT application, however, the final goal of the ASR system is to serve the NLU and dialog system to provide the correct interaction with end users. This would require low action classification error (AER) of the NLU system. Because the action classification depends on the right input of grammars, the grammar error rate (GER) is also an important factor. In order to measure the performance of different ways of adaptation, we measure all WER/GER/AER instead of WER only.

The remainder of the paper is structured as follows: Section 2 describes the data resources used in our experiments. Section 3 presents our baseline system and the dialog system, and Section 4 briefly reviews the adaptation approaches MAP, MLLR and fMLLR. Section 5 is devoted to the experimental study and discussions, and Section 6 concludes the paper.

2. Data Acquisition

All data in our experiments are collected inside an acoustically controlled room which is set up at FBK for the DICIT project, comprising a 15-microphone nested array [5], a 5+1 surround audio system, and a 46-inch LCD display. Figure 1 shows the room and the position of TV, array and loudspeakers.

The front-end processing of acoustic echo cancellation
(AEC) and beamforming (BF) are done before passing the audio to the ASR system. The AEC algorithm is based on a sub-band NLMS based compensation approach, that presents advantages in terms of convergence, robustness against noise and reduced computational load [6]. The monophonic TV-audio is diffused in the room by two loudspeakers and the recorded audio channel serves as reference for the algorithm. The BF technique is based on time-delay compensation: the actual speaker is located and delay-and-sum beamforming is applied. For these batch experiments a unique position is estimated for every utterance, and the resulting delay vector is applied to the multi-channel input so that a monophonic enhanced signal is provided.

We choose the TIMIT database as our adaptation data because TIMIT was designed to provide phonetic balanced speech data for the development and evaluation of ASR systems. It consists of utterances of 630 speakers that represent the major dialects of American English. Among 630 speakers, we use 542 speakers (about 276 minutes of data) as the real adaptation data (we refer to it as TIMIT-train), and the remaining 88 speakers (about 45 minutes of data) as the held-out data to tune parameters in the adaptation process (we refer to it as TIMIT-test).

TIMIT data was first contaminated to produce the first set of the adaptation data (TIMIT1), which is split into TIMIT-train and TIMIT-test accordingly. In addition to the simulation of the room acoustics, the contamination process also takes into account AEC and BF. TIMIT data was then played back and recorded in the DICIT room, which is referred as TIMIT2 data, and split into TIMIT2-train and TIMIT2-test accordingly.

We have two sets of DICIT data available, each with different acoustic setup. DICIT-dev1 has 4 English speakers with close-talking microphone, uttering 200 sentences related to the DICIT domain. The close-talking data was then played back and recorded in the DICIT room with a high-quality loudspeaker. DICIT-dev2 has another 4 English speakers reading the same DICIT transcripts as DICIT-dev1, but recorded in the real DICIT acoustic environment. With this setup, 4 speakers have been recorded together with a close-talking microphone and with the microphone array. During part of the recordings, some TV audio has been diffused by two frontal (left and right) loudspeakers of the surround system. The TV signal is hopefully removed by the AEC process.

DICIT-test has 4 non-native English speakers, each uttering 150 DICIT-related scripts which do not overlap with the scripts in DICIT-dev data. The data is recorded in the DICIT room, while half of the data is recorded with speakers allowed to move around the room. This is different from the data recorded for DICIT-dev where speakers talk in a fixed position. There is also another difference: the TV audio is directly taken from live channel. The TV volume is less controlled and more realistic than the previous data setup, resulting in a high variability in energy ratio between read speech and diffused audio. This would affect the AEC processing since the algorithm exhibits a different behavior depending on the variable position of speakers and actual levels of diffused audio and speech.

3. The IBM Embedded System and NLU/Dialog System

Our baseline acoustic model is the IBM EVV model. Its front-end generates 39-dimensional feature vectors, with 12 mel frequency cepstrum coefficients (MFCCs) and the frame log-energy, augmented by the first- and second-order derivatives, with 150 frames per second. The hidden Markov models (HMMs) used are word-internal context-dependent triphones. HMM state output densities are Gaussian mixture models (GMMs), estimated by maximum likelihood. The GMMs have 826 word-internal context dependent states, modeled by 11k Gaussians, which are gender-dependent. The decoder then uses gender-dependent labeling (GDL) [7] to estimate Gaussian likelihoods. The final GMMs are MPE (minimum phone error) [8] trained from more than 900 hours of close-talking dictation data and close-talking car data at various speed and noise levels.

Our language model is a free-format model that uses a trigram statistical language model (SLM) with 50 embedded grammars. In an embedded-grammar language model [9], most words are in a class by themselves, but certain concepts like DATE, TIME, or NUMBER are implemented using a grammar to compute class probability $p(word|class)$, where words are generated by the probabilistic finite-state grammar corresponding to the class.

After ASR has generated sentences of decoded words and embedded grammars, the dialog system then uses a hierarchical classification model to assign one of 114 actions to the sentences. The classifiers at each level of the classification tree are built with maximum entropy models that combine unigram, bigram, and trigram features. Once an action is determined, the action and any identified embedded grammars are passed to the dialog manager, which generates proper response to the users.

Normally, the ASR community uses the word error rate (WER) to measure system performance. Consider the sentence: I would like to fly on May One Ten AM. This could likely be generated in two ways based on the embedded grammars: I would like to fly on [DATE May] [TIME One Ten AM] or I would like to fly on [DATE May One] [TIME Ten AM].

From a WER point of view these two are the same. However, to a dialog system that uses the embedded grammars as parameters to determining actions in the dialog manager, there's a huge difference between the two. Words that align to the embedded grammars are more relevant to the dialog system. Even though words and classes are correct, the alignments of words to classes are not. What matters most to a dialog system is whether the system understands each input sentence and responds correctly. We call this "turn accuracy". A decoded sentence is counted as correct if the action classification of the decoded utterance matches that of the expected truth. In addition, the values of the decoded embedded grammars, if there are any, must also match those of the truth as well. Thus the action classification error rate (AER) highly depends on the embedded grammar error rate (GER). Therefore when we measure the ef-
fectiveness of an adapted acoustic model, we must take into account WER/GER/AER that affect the performance of the NLU system and the dialog system.

4. Adaptation Approaches

In our experiments we use MAP [2] and MLLR [3] as acoustic model adaptation with correct transcripts (supervised adaptation), and use fMLLR as on-line unsupervised feature space adaptation, leaving the acoustic model untouched. We briefly review each of these approaches in the following.

In MAP adaptation Gaussian means and covariances are updated as follows:

\[ \mu_{s,i,\text{map}} = \frac{c_{s,i,\text{a}} \mu_{s,i,\text{a}} + \lambda c_{s,i} \mu_{s,i}}{c_{s,i,\text{a}} + \lambda c_{s,i}} \]  

(1)

where \( \mu_{s,i,\text{a}} \) and \( c_{s,i,\text{a}} \) are the means and counts of the adaptation data, \( c_{s,i} \) and \( \mu_{s,i} \) are the Gaussian component counts and means, and \( \lambda \) is a parameter which controls how much contribution from the baseline model to the adapted model. The bigger \( \lambda \) is, the larger effect of the baseline model to the final adapted model. If the in-domain adaptation data is small, a large \( \lambda \) is recommended because gaussian estimation from the out-of-domain data is more reliable; otherwise use a value of \( \lambda \) to get the best performance on held-out data.

In MLLR adaptation, a linear transform \( A_{g} \) is applied to the Gaussian means:

\[ \mu_{s,i,\text{mllr}} = A_{g} \mu_{s,i} + b_{g} \]  

(2)

where subscript \( g \) denotes the regression class for state \( s \), and \( b_{g} \) is a bias term. \( A_{g} \) and \( b_{g} \) are estimated to maximize the likelihood of the adaptation data.

In fMLLR adaptation, a linear transform \( A \) is applied to the feature vectors of a speaker, i.e., \( y = Ax \), and the transformed feature vector is fed to the recognizer. \( A \) is again estimated to maximize the likelihood of the adaptation data. For details see [4]. Note that we use fMLLR on speaker-based adaptation data. This is more effective than using a single global fMLLR transform for all adaptation speakers. MLLR transforms can also be computed per speaker to adapt the acoustic model, but the need for storing the transformed acoustic model makes speaker-based MLLR not practical in the resource-constrained devices such as a set-top box. The advantage of fMLLR on such devices is that we only need to store the single transform matrix \( A \) per speaker. In the real application where the speaker boundary is not given beforehand, fMLLR is estimated by an incremental stochastic optimization approach [10].

5. Experiments and Results

Our first set of experiments are to check the effectiveness of out-of-domain adaptation data \( \text{TIMIT1} \) and \( \text{TIMIT2} \). We use MAP adaptation on the baseline IBM EVV model. The language model is just built with TIMIT training data alone, with perplexity of 15 and 40% prediction error rate on the TIMIT test data. Table 1 shows the word error rate (WER) results on the corresponding TIMIT test data. MAP has given both adapted systems about 10% relative improvement over the baseline.

Similarly, in Table 2, Table 3 present the NLU action classification error (AER) on the baseline acoustic models and two adapted acoustic models on \( \text{DICIT-dev1} \) and \( \text{DICIT-dev2} \) respectively. Obviously using N-best list of 5 provides an easy gain of more than 30% over the 1-best. Apparently \( \text{DICIT-dev1} \) data matches the baseline system better because it is recorded first using close-talking microphone and then played back and recorded again in the DICIT room, while \( \text{DICIT-dev2} \) is recorded in the DICIT room directly.

On the \( \text{DICIT-dev1} \) test set, the \( \text{TIMIT2} \) model is better than the \( \text{TIMIT1} \) model because \( \text{TIMIT2} \) data matches the \( \text{TIMIT1} \) data set than the artificially contaminated data \( \text{TIMIT1} \), with 22% relative gain over the baseline AER on 1-best, and 13% relative gain over the baseline AER on 5-best; on \( \text{DICIT-dev2} \) test data, both \( \text{TIMIT2} \) and \( \text{TIMIT1} \) sets do not match the test data. While \( \text{TIMIT1} \) degrades a little bit over the baseline, \( \text{TIMIT2} \) provides 23% relative gain over the baseline on 1-best, and 29% relative gain over the baseline on 5-best. These results recommend that the \( \text{TIMIT2} \) model is better suited for adaptation purpose than the contaminated speech data \( \text{TIMIT1} \).

Table 4 compares the AER of 1-best on close-talking (CT) data and distant-talking (DT) data of \( \text{DICIT-dev2} \). Since both CT and DT data are recorded in the DICIT room, AEC is applied to CT data as well. Note that there is 13% absolute difference between CT and DT on the baseline. The gap reduces quite a lot on the two adapted systems, with only 5% gap on \( \text{TIMIT2} \) model. This is because both models only adapt on the DT data.

### Table 1: WER comparison of adapted model \( \text{TIMIT1} \) and \( \text{TIMIT2} \)

<table>
<thead>
<tr>
<th>System</th>
<th>Baseline</th>
<th>TIMIT1</th>
<th>TIMIT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT1-test</td>
<td>33.7</td>
<td>30.0</td>
<td>–</td>
</tr>
<tr>
<td>TIMIT2-test</td>
<td>28.2</td>
<td>–</td>
<td>25.2</td>
</tr>
</tbody>
</table>

### Table 2: AER comparison of adapted model \( \text{TIMIT1} \) and \( \text{TIMIT2} \) on \( \text{DICIT-dev1} \)

<table>
<thead>
<tr>
<th>System</th>
<th>Baseline</th>
<th>TIMIT1</th>
<th>TIMIT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>nbest=1</td>
<td>18.5</td>
<td>17.0</td>
<td>14.4</td>
</tr>
<tr>
<td>nbest=3</td>
<td>13.4</td>
<td>13.7</td>
<td>11.1</td>
</tr>
<tr>
<td>nbest=5</td>
<td>11.6</td>
<td>12.5</td>
<td>10.1</td>
</tr>
</tbody>
</table>

### Table 3: AER comparison of adapted model \( \text{TIMIT1} \) and \( \text{TIMIT2} \) on \( \text{DICIT-dev2} \)

<table>
<thead>
<tr>
<th>System</th>
<th>Baseline</th>
<th>TIMIT1</th>
<th>TIMIT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>nbest=1</td>
<td>32.5</td>
<td>34.4</td>
<td>25.1</td>
</tr>
<tr>
<td>nbest=3</td>
<td>27.4</td>
<td>27.7</td>
<td>19.2</td>
</tr>
<tr>
<td>nbest=5</td>
<td>25.5</td>
<td>25.5</td>
<td>18.1</td>
</tr>
</tbody>
</table>

Our second set of experiments is to use \( \text{DICIT-dev} \) data as in-domain adaptation data, and test on \( \text{DICIT-test} \). Note again that the acoustic condition of two sets of \( \text{DICIT-dev} \) is very different from that of \( \text{DICIT-test} \) (see Section 2). The \( \text{TIMIT2-test} \) data is used to tune adaptation parameters in MAP and MLLR. The fMLLR is turned on using the stochastic approach in [10]. Since \( \text{DICIT-dev} \) has 8 speakers, each with 200 sentences of the same scripts, we experiment the following three ways of adaptation:

- Adapt1: use \( \text{TIMIT2-train} \) and \( \text{DICIT-dev} \) as MAP adaptation data on the baseline model;
Adapt2: use DICIT-dev as MLLR adaptation data on TIMIT2 model. This is tried because DICIT-dev is a small set of adaptation data.

Adapt3: use DICIT-dev as MAP adaptation data on TIMIT2 model;

Table 5 presents WER/GER/AER of baseline model and different adapted models with 1-best decoding output. Note GER is only computed on the reference sentences which have embedded grammars, while AER is computed on all reference sentences. Clearly DICIT-test is much noisier than DICIT-dev with AER as high as 37.2 (vs. 32.5).

We see gradual improvement of WER moving from TIMIT2 to Adapt1, Adapt2 and Adapt3. However, TIMIT2 does not improve on GER. While Adapt2 has better WER than Adapt1, its GER and AER are worse than those of Adapt1. Therefore improving WER alone does not guarantee better action classification. For example, Adapt1 decodes sentence do you have British Eurosport as mute on British Eurosport, and Adapt2 decodes the same sentence as you application response. Adapt1 gets the grammar right while Adapt2 gets the grammar wrong.

Among all the adapted models, Adapt3 is the best, improves on WER/GER/AER 17%/18%/16% relative respectively. From the out-of-domain adaptation data TIMIT2, there is relative 12%/5%/ gain on WER/AER, and 17% lose on GER over the baseline; from the in-domain adaptation data DICIT-dev, there is 17%/30%/11% relative gain on WER/GER/AER on top of the out-of-domain adapted model TIMIT2. Therefore, most of the gain on GER/AER comes from the in-domain adaptation data.

Table 5: WER/GER/AER comparison of different acoustic models on DICIT-test

Table 6 presents the AER performance of 1-best on close-talking data and distant-talking data of DICIT-test environment than using contaminated speech data. By playback and recording through loudspeakers we eliminate the need for actual people to read scripts. By considering also grammar error rate (GER) and action classification error rate (AER) in addition to word error rate (WER), we have also shown the best way to adapt the baseline close-talking model using out-of-domain adaptation data (TIMIT) and in-domain adaptation data. The experimental results show that in-domain data is definitely needed to improve more on GER and AER. Therefore in order to have a satisfactory NLU and dialogue application, we need more in-domain data and also improve upon the front-end processing handling the possible acoustic changes during interaction.

7. Acknowledgment

This work was partially supported by the IST EU FP6 IST-034624 research program DICIT.

8. References


