MLLR-based accent model adaptation without accented data

LIU Wai Kat, Pascale FUNG
{eekat, pascale}@ee.ust.hk

Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology
Clear Water Bay, Hong Kong

Weniwen Inc.

ABSTRACT

When the user has an accent different from what the automatic speech recognition system is trained with, the performance of the systems degrades. This is attributed to both acoustic and phonological differences between accents. The phonological differences between two accents are due to different phoneme inventories in two languages. Even for the same phoneme, foreigners and native speakers pronounce different sounds. Since accented data is rare but monolingual data is abundant, we propose using the accented speaker’s first language data directly instead of accented data in the second language for our purpose. We propose adapting the native English phoneme models to accented phoneme models using first language data in MLLR adaptation. The baseline performance is 35.25% (phone accuracy) in using native English phone models to recognize Cantonese-accented English speech data. We compare accent adaptation by using accented data and source language data. On the average, using accented data for adaptation improves the phone accuracy by 69.98% while using source language data for adaptation improves the phone accuracy by 70.13%. This shows that both kinds of adaptation data give similar improvements. Therefore non-accented data can be used for adaptation. We can rapidly obtain an accent-adapted acoustic model without the need of collecting accented database.

1. Introduction

In our previous work, we have shown that there are phonological differences between two accents due to different phoneme inventories [1]. In addition to the phoneme inventory difference it is observed that for the same phoneme, foreigners and native speakers pronounce different sounds [2,3]. Consequently, we need to find a way to normalize the acoustic differences between the same phones.

Accent adaptation has been widely studied in recent years in order to reduce error rates. However comprehensive accented data is usually not available. In order to handle accent problem, we propose extracting acoustic characteristics from the source language (the first or the mother language) directly. Speech data of the first language is easy to collect and usually widely available. This can save a lot of effort to collect accented database.

We use Cantonese accented English as an example. We first find a mapping between Cantonese phones and English phones using linguistics knowledge. Then we re-align the source language speech data using English phonemes. Finally we adapt the native English phoneme models to accented phoneme models using MLLR adaptation. We use the first language data of the accented speech to do the MLLR adaptation. The mapping between the first language and the second language can be found from linguistics knowledge. We re-align the source language speech data using English phonemes according to the mapping rules.

This paper is organized as follows: We explain the use of source data in adaptation in section 2. Section 3 describes our method to find the mapping between Cantonese and English phonemes. More details on the acoustic adaptation are given in section 4. Evaluation and results are given in section 5. Conclusion is given in section 6.

2. Adaptation without accented data

One major problem for accented speech recognition is the lack of large accented database. The newly released Foreign Accented English (FAE) database from the OGI does not contain enough data for comprehensive study for one particular accent. Many researchers have to collect accented database by themselves. On the other hand, speech database of the source languages is widely available and easy to collect. Therefore it is more efficient if we could use the source language directly. In our previous work [1], we have already shown that speaking style of the first language affect people in speaking second language. The acoustic features of accents can be derived from the source language. We perform our MLLR adaptation using source language data directly.

We carry out supervise training using source language data. Since the original source language data are transcribed in Cantonese phonemes, we need to choose the speech segment or Cantonese...
phonemes from the source language data to map to English phoneme models. In the next section we present our proposed idea to find the phoneme relationship between two accents using linguistic knowledge.

3. Mapping rules between Cantonese phonemes and English phonemes

The speech data from the source language is originally transcribed using Cantonese phonemes. We obtain a mapping between Cantonese phonemes and English phonemes from linguistics knowledge. The Cantonese transcriptions are then transformed to English transcriptions according to the mapping.

In most countries, many linguistics researchers are interested in studying of the speaking behaviors of their own language and its relationship with Native English. Therefore knowledge between source language phonemes and English phonemes are well studied by some linguistics researchers. We obtain the mapping between the Cantonese phonemes and English phonemes from "A Chinese Syllabry Pronounced According to The Dialect of Canton" by Huang, His-ling [7].

Examples of the phoneme mapping suggested by Huang are given in table 1. This shows that some phonemes such as /M/ and /T/ can be found in both languages. Therefore the speech segment correspond to Cantonese phoneme can be used to transform the parameters of the same phoneme in English. There are many phonemes in Cantonese that can be divided into smaller English phoneme segments. These speech segments are used to train the English phonemes according to the mapping rules. For those phonemes that are missing in Cantonese, we use native English phonemes as substitutions.

| Table 1: Examples of phoneme mapping between Cantonese and English |
|---|---|---|---|---|---|---|
| Cantonese | English | Cantonese | English | Cantonese | English |
| AA | AA | EOI | UW | Z | JH |
| AAI | AY | EON | UH | N | OE | ER |
| AAK | AA K | EOT | UH | T | OEI | UH |

4. Acoustics adaptation

The second problem we are facing after having obtained a phoneme mapping between Cantonese and English is to adapt English acoustic models to Cantonese data in order to obtain a Cantonese accented English acoustic model. In general, model adaptation is a good approach when only a small amount of data is available. Maximum likelihood linear regression or MLLR is a popular and successful method for model parameter adaptation. It finds a transformation that minimizes the mismatch between the models and the adaptation data. It transforms the mean and variance parameters of a Gaussian mixture HMM system such that each state in the HMM system is more likely to generate the adaptation data.

In MLLR, we assume a linear transformation from the original English model mean $\mu_i$ to the new estimated accented model mean $\hat{\mu}_i$ which is given by

$$\hat{\mu}_i = a_i \mu_i + b_i$$

where $b_i$ is the bias term and $a_i$ is the multiply term to be estimated.

In matrix notation, the estimated mean vector of the accent-adapted models can be re-written as

$$\hat{\mu} = W \xi$$

where $W$ is the $n \times (n+1)$ transformation matrix (where $n$ is the dimensionality of the data) and $\xi$ is the extended mean vector of the native accent English models,

$$\xi = [w, u_1, u_2, u_3, \ldots u_n]'$$

where $w$ represents the offset term for the regression ($w=1$ to include an offset in the regression, $w=0$ to ignore offsets). Hence $W$ can be decomposed into

$$W = [b \ A]$$

where $A$ represents an $n \times n$ transformation matrix and $b$ represents a bias vector. The transformation matrix $W$ is obtained by solving a maximization problem using the Expectation-Maximization (EM) technique. This technique is also used to compute the variance transformation matrix. Using EM results in the maximization of a standard auxiliary function $Q(\lambda, \hat{\lambda})$. 

| AAM | AA M | EP | EP A | OEK | ER K |
| AAN | AA N | EU | UW | OENG | ER NG |
| AANG | AA NG | F | F | OI | OY |
Let $\lambda$ be the current set of native accent model parameters and $\hat{\lambda}$ is the new estimated accented model parameter set. If all possible state sequences of length $T$ are denoted by the set $\Theta$, total likelihood of the model set generating the observation sequences is

$$F(O|\lambda) = \sum_{s \in S} F(O, \Theta = s | \lambda)$$

Define $r_s(t)$ as the a posteriori probability of occupying state $s$ at time $t$ given that the observation sequence $O$ is given by

$$r_s(t) = \frac{1}{F(O|\lambda)} \sum_{s' \in S} F(O, \Theta = s' | \lambda)$$

Therefore, the auxiliary function can be written as

$$Q(\lambda, \hat{\lambda}) = \text{constant} + F(O|\lambda) \sum_{ij} \sum_{s} r_j(t) \log b_j(o_i)$$

Expanding $\log b_j(o_i)$ in the auxiliary function. Complete the differentiation of the auxiliary function with respect to $W_s$ and equating to zero. Hence

$$\sum_{i=1}^{T} r_j(t) C_s^{-1} o_i \tilde{p}_{s_i} = \sum_{i=1}^{T} r_j(t) C_s^{-1} W_s \tilde{p}_{s_i}$$

This gives the general form for computing $W_s$ using the adaptation data.

MLLR is always found to give good results when there is only a small amount of adaptation data available [6]. This is due to the fact that MLLR performs a global transformation $W$ when few or zero observation for a particular model is available. When there is sufficient adaptation data, MLLR finds different transformations for different classes instead of one single global transformation. The classes are found by clustering. Since MLLR can work well for both cases, we choose MLLR for our acoustic model adaptation. In this case, by using only a small amount of data, a native English accent model set can be adapted to better fit the characteristics of another accent such as Cantonese accent.

5. Experimental setup and result

In our experiment, native English phonemes models are adapted to Cantonese accented phonemes models using Cantonese speech. We also use Cantonese-accented English speech for adaptation in order to compare the result with using non-accented data. This experiment involves four kinds of speech data. There is no corpus that contains native, accented and non-accented data together. Therefore we use four different corpora and all the speech data are channel-normalized by Cepstral Mean Normalization. First, we use about 8000 native English utterances from WSJCAMJ0 [8] corpus to train the native English phonemes models. Second, we use 5000 Cantonese tokens from CUWORD [9] as the non-accented adaptation data. Third, 5000 Cantonese-accented tokens from HKTimit [1] are used as accented adaptation for comparison. Last, 200 Cantonese accented English utterances from FAE corpus are used for testing. We use various amount of adaptation data, ranging from 250 tokens to 5000 tokens, to carry MLLR adaptation. We use HTK toolkit with MLLR routines to train, adapt and test the models. The procedures of these experiments are as follows:

1. All the speech data of the four speech corpora is sampled, high frequency pre-emphasis is performed, Hamming windowed and followed by MFCC feature extraction on a frame by frame basis. Each utterance is converted into a series of observation vector contain 39 components, consists of 12 MFCCs and energy together with their first and second derivatives.

2. The 8000 utterances from the WSJCAMJ0 corpus are used to train a English phoneme recognizer with native accent. The phoneme recognizer consists 45 phonemes and one silence model. Each of them is modeled by a 3 states HMM with 16 mixtures and trained by Embedded Model Re-estimation.

3. 200 Cantonese-accented English utterances from FAE are used to test the baseline performance of the phoneme recognizer.

4. 250 to 5000 Cantonese tokens from CUWORD are used as adaptation data. MLLR adaptation is performed and an accent adapted phoneme recognizer is obtained by using this non-accented data.

5. 250 to 5000 Cantonese-accented English tokens from HKTimit are used as another set of adaptation data. Another accent adapted phoneme recognizer is obtained by using this accented-data.

6. The phoneme recognizers obtained by step 4) and step 5) are tested against the 200 Cantonese-accented English utterances from FAE.

The performance of using non-accented speech data are evaluated together with using accented data. The baseline phone accuracy is 35.25%. Table 2 shows the result on the phoneme accuracy using different amount of adaptation in both cases of using accented and non-accented data. After MLLR, all shown a increase in performance. On the average, using accented data for adaptation improves the phone accuracy by 69.98% while using source language data from accented speech improves the phone accuracy by 70.13%. This shows that both kinds of adaptation data give similar improvements.

Table 2: Phoneme accuracy in recognizing accented speech
<table>
<thead>
<tr>
<th>Number of tokens used for adaptation</th>
<th>Adaptation using Non-accented data</th>
<th>Adaptation using Accented data</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>46.52%</td>
<td>45.42%</td>
</tr>
<tr>
<td>1250</td>
<td>57.40%</td>
<td>56.32%</td>
</tr>
<tr>
<td>2500</td>
<td>67.04%</td>
<td>68.03%</td>
</tr>
<tr>
<td>5000</td>
<td>68.93%</td>
<td>69.90%</td>
</tr>
</tbody>
</table>

The result has shown that acoustics pattern of an accent group can also be found directly from the source language. Such information can also be used to build accent specific models by adaptation. The advantage of using source data is that we could save a lot of effort in collecting accent database.

### 6. Conclusion

Acoustic difference is one of the major mismatches between accents. We propose an easy way to build the accent-adapted acoustic models by just using the first or source language data as the training data. Since speakers carry their speaking style from the first language to the second language. We can extract the accent information directly from the first language. We use linguistic knowledge to find mapping between the phonemes of the first language and second language. We use MLLR to adapt the native English phoneme models to accented phoneme models with first language data only. This method can save our effort in collecting accent data and the accent-specific models can be rapidly built. The performance of using source language speech data is evaluated together with using accent data. The baseline phone accuracy is 35.25%. On the average, using accented data for adaptation improves the phone accuracy by 69.98% while using source language data from accented speech improves the phone accuracy by 70.13%. In all cases of using different amount of adaptation data, using source data give a similar result when using accented data. The result has shown that acoustic pattern of a accent group can also be found directly from the source language. Such information can also be used to build accent specific models by adaptation. As a result we can save a lot of effort in collecting accent database.

### 7. Reference


