PRONUNCIATION SPACE MODELS FOR PRONUNCIATION EVALUATION

Si Wei, Yi-Qian Pan, Guo-Ping Hu, Yu Hu and Ren-Hua Wang
{siwei, yqpan, gphu, yuhu, rhwang}@iflytek.com

iFLYTEK Research, Hefei

ABSTRACT
Posteriors probability is mostly used for pronunciation evaluation. This paper introduces pronunciation space models to calculate posterior probability replacing traditional phone-based acoustic models, which makes the calculated posterior probability more precise. Pronunciation space models are constructed using unsupervised clustering method guided by human scores and phone-level posterior probability. By using correlation between scores and human scores as the performance measurement, pronunciation space models based method shows its effectiveness for pronunciation evaluation in the experiments on a Chinese database spoken by Koreans with the correlation’s improvement from 0.390 to 0.415 comparing to the traditional method based on phone based acoustic models.

Index Terms— pronunciation evaluation, posterior probability, pronunciation space models, speech recognition

1. INTRODUCTION
With the development of computer science and artificial intelligence, much research has been done to assist language learning with computers, which is called Computer Assisted Language Learning (CALL). In the last decade, with the rapid improvement of speech technology, CALL systems have become much more intelligent than before. They can provide many potential benefits to both the language learner and teacher. Pronunciation evaluation is one of the essential functions of CALL system. With this function, CALL system can evaluate the pronunciations of speakers and give them informative feedback.

Many researchers have studied pronunciation evaluation. Members of the SRI speech group mainly focus on evaluating the overall pronunciation quality of learners [3, 8]. They take word posterior probability, timing and duration scores as the measurement of pronunciation and measure performance of evaluation algorithm by the correlation between machine scores and human scores. The joint research by the speech group of Cambridge University and the AI lab of MIT mainly focuses on pronunciation error detection and phone-level pronunciation evaluation [10]. They also investigate different ways to measure performance of mispronunciation detection. The VICK system developed by the University of Nijmegen investigates the reasonability of human scoring and the effect of prosody, fluency and segmental quality on human scoring [2]. The system from Kyoto University considers the importance of different phonemes in language learning and the effect of different types of errors on pronunciation proficiency [9]. Recently, structural representation has been used to assess pronunciation in order to capture the structure, or higher level aspects of the language, when spoken by non-native speakers [1, 7].

Almost all the methods are based on Automatic Speech Recognition (ASR). They share some common limitations. The most severe limitation is that all methods employ phone based acoustic models defined in ASR directly which contains only a little capability to distinguish mispronunciation of different serious levels such as partially changed mispronunciations. To break through this limitation, acoustic models need to be reconstructed so as to distinguish different serious level mispronunciations. This is very similar to pronunciation modeling of ASR. In pronunciation modeling, acoustic models are reconstructed to model partial changes of pronunciation. Liu and Fung [6] propose partial change phone models (PCPMs) to represent partial changes and then merge PCPMs into original acoustic models to increase the model resolution for pronunciation variations. All in all, in order to deal with pronunciation changes, acoustic models should be reconstructed to represent the pronunciation variations. In this paper, Pronunciation Space Models (PSM) are proposed to represent mispronunciation of different serious levels to enhance the discriminative capability of acoustic models for partially changed mispronunciation. These PSMs are built with unsupervised clustering method. Then posterior probability based on PSM is used as the measure for pronunciation evaluation. Experimental results indicate that the new method based on PSM outperforms the traditional method based on phone based acoustic models with the correlation’s improvement from 0.390 to 0.415.

This paper is organized as follows. Section 2 introduces the definition and the construction of PSM. Then pronunciation evaluation method using PSM is introduced in section 3. In section 4, database and experimental result is introduced. Then this paper is concluded in section 5.
2. PRONUNCIATION SPACE MODELS CONSTRUCTION

Phone-based acoustic models can only effectively deal with the mispronunciation from one phone to another. But for partially changed mispronunciation or mispronunciation not belonging to the phones in the model set, this kind of acoustic model will be less effective. At the same time, partially changed mispronunciation is one of the most frequent mispronunciations. So in order to efficiently evaluate the pronunciation, more effort should be put on partially changed mispronunciations. This paper builds acoustic models named PSM to represent the characteristics of each mispronunciation including partially changed mispronunciation.

2.1. Definition of PSM

Phone-based acoustic models are not suitable for mispronunciation detection because of its weakness in handling partially changed mispronunciations. There are two different approaches which have been investigated to deal with the partially changed mispronunciations with phone based acoustic models. One approach is multi-training method with mispronunciation and correct pronunciation samples. This kind of model is able to describe partially changed mispronunciation but sacrifice the capability of distinguishing mispronunciation from correct pronunciation. The other approach is to train acoustic models only with correct pronunciation. This kind of model retains the ability to distinguish correct pronunciation but its power to detect partially changed mispronunciation is limited due to the neglect of mispronunciation data at the acoustic training stage. In order to deal with all kinds of mispronunciations effectively, suitable acoustic models should be built reflecting all kinds of pronunciations. The native model trained from native database plus non-native model can be regarded as an example of this kind, but it can not model the frequently occurred intermediate pronunciation between native and non-native pronunciation precisely.

To break through the above mentioned limitation, the authors try to build different acoustic models named PSM for mispronunciation of different serious levels, not just correct vs. mispronounced, or native vs. non-native. In PSM, traditional phone-based model is expanded to be {q₁, q₂, ..., qₖ}, representing K kinds of mispronunciation of different serious levels and K is a tunable parameter.

2.2. PSM construction based on unsupervised clustering method

The proposed PSM try to model all kinds of mispronunciations. If enough mispronunciation data labeled by humans could be obtained, these models could be built via maximum likelihood estimation or discriminative training. Unfortunately, it’s very hard to label enough mispronunciation data, and it is even harder to label enough data with acceptable consistency for model training. At the same time, some kinds of mispronunciation are very rare in the true pronunciation data. All these reasons make supervised pronunciation space modeling very hard to implement. This paper introduces an unsupervised pronunciation space modeling method with posterior probability calculated for each sample. This method is shown as follows:

Step 1, Collect all the samples belonging to one phone q from the acoustic model training database and mark it as oₗ = [oₗ,q₁, oₗ,q₂, ..., oₗ,qₖ], where R is the number of the samples of q. Then use the phone-based acoustic models to calculate the posterior probability for each sample with the boundaries determined by forced alignment.

Step 2, Cluster oₗ to be several groups with pre-calculated posterior probability and mark it as oₗ = [oₗ,c₁, oₗ,c₂, ..., oₗ,cₖ], where the number of groups K is determined by the definition of PSM. Here, the clustering is done by ordering posterior probability and then splitting oₗ into K groups with same size.

Step 3, Train the PSM via maximum likelihood estimation with the clustered acoustic training data. If the training data is limited, PSM can be trained by the adaptation method such as Maximum Likelihood Linear Regression (MLLR) [5] or Maximum a Posterior (MAP) [4].

The unsupervised clustering method only uses the posterior probability as the measurement for clustering. But posterior probability can not precisely measure the correctness of phone samples. This paper also uses human scores as an auxiliary measure for clustering. For example, for the most correct pronunciation space model, the clustered group should have best human score despite of the highest posterior probability. At the same time, for the worst pronunciation space model, the samples selected should have lowest human score.

Although the unsupervised method cannot perfectly split the data into correct pronunciation and mispronunciation, this pronunciation space modeling method is capable of giving us an elementary knowledge about the effectiveness of PSM.

3. PRONUNCIATION EVALUATION METHOD BASED ON PRONUNCIATION SPACE MODELS

This section introduces the pronunciation evaluation method used in this paper. As shown in the previous research, posterior probability, duration scores and rate of speech are all useful measures for pronunciation evaluation. But this paper focuses on posterior probability method and puts emphasis on how to improve the performance of posterior probability. Posterior probability using PSM for
pronunciation evaluation is introduced in the following sections.

3.1 Posterior Probability

Given an acoustic observation $o$ and a word sequence $W$, according to Bayesian rule, the posterior probability that $o$ is recognized as $W$ is defined as follows:

$$P(W | o) = \frac{p(o | W)P(W)}{P(o)}$$

(1)

where $P(W)$ is the probability of $W$, $p(o | W)$ is the probability of observation $o$ generated by word sequence $W$, and $P(o)$ is the probability of acoustic observation $o$. In Eq. (1), $P(W)$ is normally calculated from a language model, and $p(o | W)$ is calculated with acoustic HMMs. $P(o)$, which is difficult to be estimated directly, is often calculated according to the following formula:

$$P(o) = \sum_{H} p(o, H) = \sum_{H} p(o | H)P(H)$$

(2)

where $H$ denotes one of the possible hypotheses for $o$, and the summation is done over all possible hypotheses for $o$. It’s very hard if not impossible to predict all possible hypotheses so some approximations or constraints should be used to make the calculation feasible.

3.2 Pronunciation evaluation based on posterior probability using PSM

As shown in the previous section, $P(o)$ is approximated as Eq.(1). For pronunciation evaluation, the summation is done by a phone-loop network with phone based acoustic models, which is shown as follows:

$$P(o) = \sum_{H} p(o, H) = \sum_{h \in Q} p(o | h)p(q)$$

(3)

where $q$ is phone acoustic model and $Q$ is the phone set.

While in this paper, $P(o)$ is calculated with the pre-obtained PSMs as follows:

$$P(o) = \sum_{H} p(o, H) = \sum_{i=1}^{K} p(o | q_i)p(q_i)$$

(4)

where $K$ is the parameter of PSM and the summation is done by a pronunciation space phone-loop network with PSMs. Then the posterior probability based on PSM is implemented as follows:

$$P(W | o) = \frac{\prod_{i=1}^{K} p(o | q_i)p(q_i)}{\sum_{q \in Q} \prod_{i=1}^{K} p(o | q_i)p(q_i)}$$

(5)

This paper uses standard pronunciation space model and accented pronunciation space model for specific phone and the posterior probability is calculated as follows:

$$P(W | o) = \frac{p(o | W)P(W)}{\sum_{q \in Q} p(o | q)P(q) + p(o | q_{accent})P(q_{accent})}$$

(6)

where $q_{std}$ and $q_{accent}$ represent the standard model and accented model for phone $q$ respectively. Here the standard and the accented pronunciation space models are trained as introduced in section 2.2 with posterior probability and human scores. Because in this paper, the non-native database is quite small, MLLR is chosen to build the PSMs. The five most native-like speakers are selected and the adaptation samples are also selected via posterior probability to adapt the original native phone based models to standard pronunciation space models and another most non-native like five speakers are chosen to build accented pronunciation space models.

4. EXPERIMENTS AND RESULTS

4.1 Database instruction

There are two databases used in this paper. The first one is the native Chinese database which is used to train the acoustic model. The second one is the Chinese database spoken by Koreans. Some of the second database is used to obtain the PSMs and the other is used as the pronunciation evaluation testing database. The first database contains 56 people and each people have about 3 hours’ continuous speech. The second database contains 48 Koreans with 100 Chinese sentences. Ten of them are selected as the PSM training database. Five of the ten speakers are quite native-like and the other five have heavily non-native accent. The first five speakers’ data is used to train the standard pronunciation space models and the other five are used to train the accented pronunciation space models. The training process is the same as shown in section 2.2. Then the pronunciation evaluation is done using posterior probability with PSMs as Eq.(6). The rest 38 speakers’ database is used as the testing database and each of the sentences is labeled by human evaluator. The sentences are scored in a 10-point scale varying from quite native-like to heavily non-native.

4.2 Performance measurement

The performance of the pronunciation evaluation algorithm is measured with correlation between the machine scores and the human score as follows:

$$Corr(A,B) = \frac{\sum_{i=1}^{N} ((S_{Ai} - \overline{S}_{A}) \times (S_{Bi} - \overline{S}_{B}))}{\sqrt{\sum_{i=1}^{N} (S_{Ai} - \overline{S}_{A})^2 \times \sum_{i=1}^{N} (S_{Bi} - \overline{S}_{B})^2}}$$

(7)

Where $A$ and $B$ are two evaluators, $N$ is the number of all sentences, $S_{Ai}$ and $S_{Bi}$ are the scores give by evaluator $A$ and
By using correlation, the performance of human evaluators can be measured by the correlation between them, which is shown in table 1.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Evaluator 1</th>
<th>Evaluator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.390</td>
<td></td>
</tr>
<tr>
<td>PSM</td>
<td>0.416</td>
<td></td>
</tr>
</tbody>
</table>

Table 1, Correlation between different human evaluators.

We can see from table 1 that because the scoring is done at the sentence level, even human evaluator’s correlation is rather low.

4.3 Experiment result

This section will give the performance of the method based on PSM and the baseline system with phone based acoustic models.

The experimental results are shown in table 2.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Correlation (AVE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.390</td>
</tr>
<tr>
<td>PSM</td>
<td>0.416</td>
</tr>
</tbody>
</table>

Table 2, performance of different pronunciation evaluation algorithm

Here AVE means that the correlation is the average correlation between machine scores and the two human scores.

Table 2 indicates that the evaluation method based on PSM is better than phone based acoustic model method.

5. CONCLUSIONS

This paper introduces the pronunciation evaluation method based on posterior probability using PSM. PSM are proposed to address the limitation of previous posterior probability based research. The limitation is that partially changed mispronunciations are ignored or mixed up with correct pronunciation during the acoustic model training stage in most of the ASR-based pronunciation evaluation method, which weakens the capability of pronunciation evaluation.

To deal with the limitation, PSM based method is proposed in this paper. Specific mispronunciation acoustic models are used to represent the mispronunciation of different serious levels. First pronunciation data is collected from a range of people to construct PSM with an unsupervised clustering method guided by posterior probabilities and human scores. Then posterior probability is used to evaluate pronunciation based on the obtained PSM.

Experimental results on Chinese database spoken by 38 Koreans indicate that the PSM based posterior probability method outperforms the phone based acoustic model method. When using PSM for pronunciation evaluation, the correlation between human and machine increases from 0.391 to 0.416.

Several aspects of the proposed method can be further investigated in future. First, only posterior probability based on PSM is used for pronunciation evaluation. Many other evaluation methods could benefit from PSM. Second, the PSM is constructed by unsupervised clustering with posterior probability and partially with human scores. How to utilize more information from human labeling for PSM construction can be further investigated. At the same time, considering pool evaluation performance comparing to human, more evaluation methods should be used for evaluation. In summary, the pronunciation evaluation problem is still a work in progress, and more effort needs to be carried out to solve this problem in future.

6. REFERENCES