AN EFFICIENT MISPRONOUNCIATION DETCTION METHOD USING GLDS-SVM AND FORMANT ENHANCED FEATURES

HongYan Li, JiaEn Liang, ShiJin Wang, Bo Xu

Digital Content Technology Research Center,
Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190

ABSTRACT

Mispronunciation detection is an important component in computer assisted language learning (CALL) system. In this work, we introduce an efficient GLDS-SVM based detection method, which is successfully used in language and speaker identification systems, and combine it with traditional methods. The main ideas include: extended MFCC features with normalized formant trajectory information, and then propose a novel multi-model strategy for model training to make full use of samples and solve the problem of data unbalance, finally combine GLDS-SVM method with UBM-GMM system to further improve the performance. Experiments show that GLDS-SVM is highly efficient than traditional RBF-SVM, and the fused system can achieve a significant relative improvement of 17.5% in EER reduction, compared with the baseline UBM-GMM system.

Index Terms—Computer Assisted Language Learning, Mispronunciation Detection, Support Vector Machine, Generalized Linear Discriminant Sequence, System Fusion

1. INTRODUCTION

Mispronunciation detection is one of the main issues of computer assisted language learning (CALL), and its aim is to automatically label each phone of testing speech by correct or incorrect. In the field of interactive language learning, to give the quality of pronunciation and provide corresponding feedbacks are more important than only one score.

In past few years, lots of studies have been investigated in this area, and most of them were based on pronunciation rules [1] or posterior probability [2] derived by speech recognition. The rule base methods can only detect some of common mispronunciations which are included in rule set, and the posterior probability based methods are deeply dependent on the precise of acoustic models. In fact, there are two key problems for mispronunciation detection: pronunciation feature and error detect method. The feature should be effective and robust to describe pronunciation quality, and the detect method should enable to separate the samples into correct classes. In terms of another view, the mispronunciation detection can be regarded as a kind of classification problem, and the mispronunciations can be checked out by some classifiers. In [3], Dong used formant feature and RBF-SVM to evaluate mandarin vowels quality. And Pan [4] utilized Garbor based formant feature and GMM for mandarin vowels evaluation.

As we know, support vector machine (SVM), as a very efficient discriminative classifier, has been widely used in many tasks. But the traditional RBF or other kernels based SVM has a relative higher computational and storage consumption, and the model size is often increasing remarkably when there are a large amount of training data, which is not fit for applications. Under these circumstances, in this paper, we introduce a GLDS based SVM method with a novel model training strategy to solve above problems and then investigate its fusion with other systems. Moreover, the formant trajectory information is directly incorporated into feature vectors to improve the depicting ability for pronunciation quality.

The rest of this paper is organized as follows. Section 2 presents the UBM-GMM system as baseline. Section 3 explores the feature fusion by adding formant information. In Section 4, we describe the GLDS based SVM methods in detail. Experiment results and analysis are given in Section 5. Finally, in Section 6, some conclusions are drawn.

2. THE UBM-GMM SYSTEM AS BASELINE

As a statistic model, Gaussian Mixture Model (GMM) is good at describing the distribution characteristics of speech, but with lower ability for pattern discrimination. In order to improve discrimination ability of GMM and utilize the concept of Goodness of Pronunciation (GOP) algorithm, we measure the pronunciation quality of a phone as:

$$score(phn_i) = \frac{1}{e_i - s_i + 1} \sum_{t=s_i}^{e_i} \log p(phn_i | x_t) \qquad (1)$$

Where s_i and e_i stand for the start and end frame of phn_i. By using Bayes rule, $p(phn_i|x_t)$ can be estimated as:

$$p(phn_i | x_t) = \frac{p(x_t | gmm_i) P(phn_i)}{\sum_{k \in V} p(x_t | gmm_k) P(phn_k)} = \frac{p(x_t | gmm_i)}{\sum_{k \in V} p(x_t | gmm_k)} \qquad (2)$$

Where gmm_i is the model corresponding to phn_i, V is either vowel set or consonant set, $p(x_t|gmm_k)$ is the output probability of the feature vector x_t upon model gmm_k. $P(phn_k)$ represents the prior probability of phn_k, and the value is set equally for each phone.

One GMM model is trained for each phone. In order to improve the training speed and the precise of models, we firstly train two universal background models (UBM) for vowels and consonants respectively, and then the final model of each phone is obtained based on its corresponding UBM by employing Bayesian adaptation algorithm [5].

3. EXPLORING MULTIPLE FEATURES

Data fusion of different features and methods has been shown to be capable of increasing the performance in many tasks [6]. The Mel-frequency cepstral coefficients (MFCC) are commonly used in speech recognition. Meanwhile, the formant and its varying trajectory are more useful for vowel discrimination. In order to take advantages of above features, we combine traditional MFCC with formant and its first and second order derivatives as basic features, and the feature vector of each speech frame is constructed as:

$$x_t = [MFCC1, \cdots, MFCC39, F1, F2, F3, \Delta F1, \Delta F2, \Delta F3, \Delta^2 F1, \Delta^2 F2, \Delta^2 F3]^T \quad (3)$$

This concatenation of different features into a single vector can be treated as a special case of fusion on feature level. We use Praat tool [7] to extract the raw formant, and then the smoothing and normalization [8] are performed.

4. THE EFFICIENT GLDS BASED SVM METHOD

In this section, we briefly introduce the whole fusion system, and then a GLDS based SVM method with a novel model training strategy is presented in detail.

4.1. Overview of the whole system

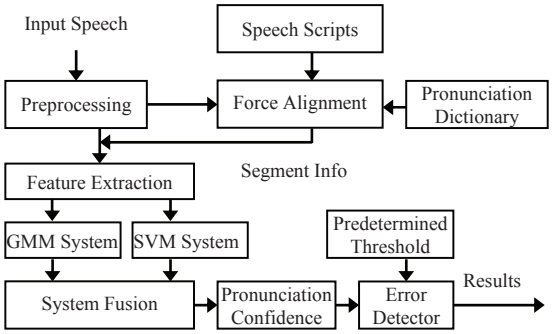

Fig.1. Architecture of the whole detect system

The overall improved system is proposed in Fig.1. The front-end feature extraction is to provide multiple pronunciation features for detect systems. Then the segment information can be obtained by force alignment. By incorporating pronunciation variations into dictionary, the precise of segmentation is improved significantly. With segment information, the features are passed through different detect systems to produce confidences. Such confidences are complementary to each other, so we further investigate to merge them. The final decision is made based upon whether the confidence is above or below a phone dependent threshold.

4.2. GLDS based SVM method

Recently, the generalized linear discriminant sequence (GLDS) based SVM method has shown dramatic performance gains in speaker recognition, and with obvious computational and storage advantages towards traditional SVM kernels [9]. In our work, we try to introduce GLDS based SVM method for mispronunciation detection and propose a practical model training strategy.

4.2.1. The GLDS based SVM with MSE criterion

Our GLDS based SVM system consists of several parts, as shown in Fig.2. Basic feature vector x_t is introduced into the system, and $p(x_t)$ is the vector of polynomial basis terms of the input feature vector. For a d-dimensional feature vector x_t and the polynomial degree q, the length of $p(x_t)$ is given by C_{d+q}^{q}.

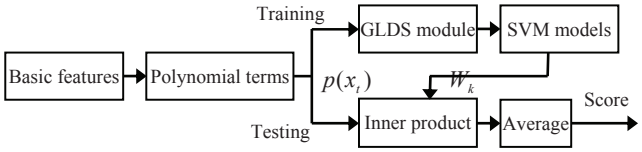

Fig.2. Block diagram of SVM detection system

In testing part, with the model W_i of phn_i, for each input feature vector x_t of phn_i, a score is produced by the inner product between W_i and $p(x_t)$. The score is then averaged over frames to produce the final output as:

$$score(phn_i) = \frac{1}{e_i - s_i + 1} \sum_{t=s_i}^{e_i} W_i^T p(x_t) \quad (4)$$

In training part, each phone's model W_k is obtained by using a discriminative training based on GLDS kernel instead of radial basic function (RBF) kernel with a mean-squared error (MSE) criterion. The training method can be simply approximated as:

$$w^* = \arg\min_w \left[\sum_{k=1}^{N_{pos}} \left| w^T p(x_k) - 1 \right|^2 + \sum_{k=1}^{N_{neg}} \left| w^T p(y_k) - 0 \right|^2 \right] \quad (5)$$

Where w^* is the optimum middleware of the model for one phone. Here, the positive training samples are denoted as $x_1, x_2, \cdots, x_{N_{pos}}$, and the negative samples are denoted as

$y_1, y_2, \cdots, y_{N_{neg}}$. Meanwhile, an output of one is desired for positive samples, and zero is desired for negative samples. With a series of w^*, the final model W for each phone can be obtained by the basis GLDS kernel, and the detail deduction is shown in [9].

4.2.2. The multi-model strategy for SVM training

As we know, training SVM models need both positive samples and negative samples. However, in most practical situations, it is extremely difficult to get enough error data as negative samples. Therefore, data unbalance is a crucial thing for SVM training. In order to solve this problem, we use current phone's samples as current phone's positive samples, other phones' samples as current phone's negative samples, and all the samples are obtained by force alignment of our collected English corpus. In addition, other phones related above denote the set of other vowels when current phone is a vowel, otherwise the set of other consonants when current phone is a consonant.

As a result, the negative samples are much more than positive samples for each phone, and it is not favor for SVM training. So we try to split all the negative samples into multiple sets in which samples are randomly selected, and then produce multiple SVM models for each phone. Since our SVM confidence is an output of inner product, the final SVM model for each phone can be calculated as:

$$W_k = \frac{1}{N}\sum_{i=1}^{N} W_k^i \qquad (6)$$

Where W_k^i is one of multiple models for phn_k, and N is the number of models. The above multi-model based training strategy is shown in Fig.3.

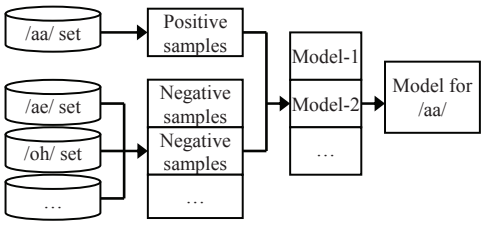

Fig.3. Multi-model based SVM training strategy

4.2.3. The advantages of GLDS based SVM system

The improved SVM detection method has several advantages over traditional RBF based SVM methods: 1) it is both computationally and memory efficient by using GLDS and inner product; 2) the discriminative ability of models is improved via MSE based training, and the method can collapse all the support vectors down into a single vector as the final model; 3) the multi-model strategy can make full use of samples and solve the problem of data unbalance, the averaged model is better than each of single random selected model.

4.3. System fusion

For mispronunciation detection, the fusion can be regarded as a problem of predicting the human subjective decisions by several machine confidences. Since a set of development data with manual labels is always needed for most of nonlinear fusion methods and it will increase system's complexity, out of the view of simple and practical application, only linear weighting method is performed in our work, and the best weights are obtained by step searching.

5. EXPERIMNETS

5.1. Setup

The experiments are carried out on a large English speech corpus, the details are shown in Table.1. Each speaker in the corpus pronounces 100 words and 100 sentences which are carefully designed. Training set is used to generate the gender dependent models for every detect method, and the experiment results are obtained on the testing set.

Corpus	# male	# female	# total
Training Set	100	200	300
Testing Set	28	50	78
All Set	128	250	378

Table.1. Construction of speech corpus in the experiment

The analysis frame length and frame shift are 25ms and 10ms respectively. The basic feature is a 48-dimension vector, including 13 MFCC with logarithm energy, F1, F2, F3 and their first and second order derivatives. The acoustic HMM models used for force alignment are trained by 120 hours of speech corpus. The number of Gaussian mixtures for GMM system is 16. For SVM system, the polynomial degree is 3, and the ratio by positive samples and negative samples is controlled as about 1:1. Note that in our work, we use the phone set and dictionary of BEEP.

In order to measure the performance of detection systems, we consider two measures of false acceptance rate (FAR) and false rejection rate (FRR). To fully reflect the changing performance of FAR and FRR with different thresholds, Detect Error Tradeoff (DET) curve and Equal Error Rate (EER) are also used in following experiments.

5.2. Experiment results

5.2.1. EER comparison of different methods

Fig.4. illustrates the DET curves of two improved methods, and the detail results of EER are listed in Table.2. In our experiment, the result by UBM-GMM system is treated as baseline. We can see that the SVM systems can improve the performance significantly, and RBF kernel is slightly better than GLDS kernel. However, under the similar EER

performance, GLDS has its special advantages towards RBF, which will be discussed in section 5.2.3.

Besides, by adding formant information into feature vectors, the performance is slightly improved. With the complementary between GMM and SVM systems, a best relative improvement of 17.5% in EER reduction is obtained by fusing the two methods. Meanwhile, we can see that the output fusion of systems has more obvious improvement than input feature fusion.

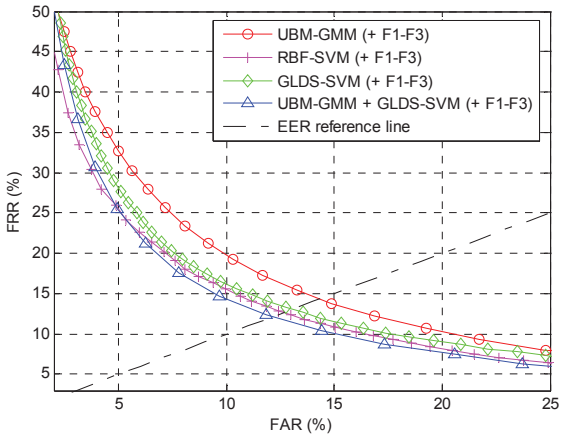

Fig.4. DET comparison for improved methods

	System	no F1-F3	+ F1-F3	Rel. Imp.
1	UBM-GMM	14.4%	14.3%	-
2	RBF-SVM	12.7%	12.6%	11.9%
3	GLDS-SVM	13.2%	13.0%	9.1%
4	Fusion 1 + 3	12.1%	11.8%	17.5%

Table.2. EER comparison for different methods

5.2.2. The effect of negative set for SVM training

Several configurations of negative set for SVM training are tested, just as shown in Fig.5. It can be seen that the multi-model strategy by using multiple negative sets is more useful for improving system's performance. When the number of negative set is over 10, the EER value is converged slowly, so the number of negative set is chosen as 10.

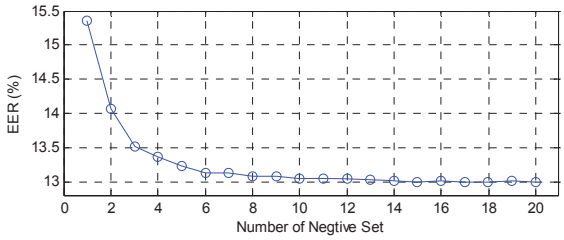

Fig.5. EER curve with different number of negative set

5.2.3. Time and storage consumption

Time and storage performance of each system is shown in Table.3. For time consumption, the average time of computing confidence for each phone of the input speech is recorded. For storage consumption, the size of used models is calculated. Results show that the GLDS based SVM system is much faster than others, and has a smaller model for each phone with the same size. And compared with traditional RBF kernel, the GLDS based SVM system is much superior on both time and storage consumption.

Method	millisec/phn	model size (MB)
UBM-GMM	2.74	1.44
RBF-SVM	1.99	138
GLDS-SVM	0.51	4.14

Table.3. Time and storage comparison of different methods

6. CONCLUSION

In this paper, we pay more attention to a GLDS based SVM detection method and propose a multi-model training strategy. The GLDS achieves an equivalent performance with higher speed and lower storage consumption than RBF. Moreover, exploring formant trajectory into features can bring progress. And the fusion of SVM and GMM systems can reduce the EER by 17.5% in relative.

7. REFERENCES

[1] Ito, A., Lim, Y., Suzuki, M., Makino, S., "Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree", *in Proc. EuroSpeech*, pp. 173-176, 2005.

[2] Franco, H., Neumeyer, L., Kim, Y., Ronen, O., Bratt, H., "Automatic Detection of phone-level mispronunciation for language learning", *in Proc. Eurospeech*, pp. 851-854, 1999.

[3] Dong, B., Zhao, Q., Yan, Y., "Objective evaluation of vowels of standard Chinese pronunciation based on formant pattern", *ACTA ACUSTICA*, Vol. 32(2) , pp. 122-128, 2007.

[4] Pan, F., Zhao, Q., Yan, Y., "Mandarin Vowel Pronunciation Quality Evaluation by A Novel Formant Classification Method and its Combination with Traditional Algorithm", *in Proc. ICASSP*, pp. 5061-5064, Las Vegas, Nevada, U.S.A, 2008.

[5] Reynolds, J. L., Quatieri, T. F., Dunn, R. B., "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing,* Vol.10, pp.19-41, 2000.

[6] Wong, E., Sridharan, S., "Fusion of Output Scores on Language Identification System", *in Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[7] Boersma, P., Weenink, D., "Praat: doing phonetics by computer (Version 4.5.16)", http://www.praat.org/.

[8] Chen, J.C., Lyu, R.Y., Chiang, Y.C., "Formant-Based English Vowel Assessment For Chinese in Taiwan", *in Proc. ICSLP*, Pittsburgh, Pennsylvania, U.S.A, 2006.

[9] Campbell, W. M., "Generalized linear discriminant sequence kernels for speaker recognition", *in Proc. ICASSP*, pp. 161-164, Orlando, Florida, U.S.A, 2002.