INTERSPEECH 2009 BRIGHTON

High Performance Automatic Mispronunciation Detection Method Based on Neural Network and TRAP Features

Hongyan Li [1], *Shijin Wang* [1], *Jiaen Liang* [1], *Shen Huang* [1], *Bo Xu* [1, 2]

[1] Digital Content Technology Research Center, Institute of Automation,
[2] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
{hyli, sjwang, jeliang, shenhuang, xubo}@hitic.ia.ac.cn

10.21437/Interspeech.2009-553

Abstract

In this paper, we propose a new approach to utilize temporal information and neural network (NN) to improve the performance of automatic mispronunciation detection (AMD). Firstly, the alignment results between speech signals and corresponding phoneme sequences are obtained within the classic GMM-HMM framework. Then, the long-time TempoRAl Patterns (TRAPs) [5] features are introduced to describe the pronunciation quality instead of the conventional spectral features (e.g. MFCC). Based on the phoneme boundaries and TRAPs features, we use Multi-layer Perceptron (MLP) to calculate the final posterior probability of each testing phoneme, and determine whether it is a mispronunciation or not by comparing with a phone dependent threshold. Moreover, we combine the TRAPs-MLP method with our existing methods to further improve the performance. Experiments show that the TRAPs-MLP method can give a significant relative improvement of 39.04% in EER (Equal Error Rate) reduction, and the fusion of TRAPs-MLP, GMM-UBM and GLDS-SVM [4] methods can yield 48.32% in EER reduction relatively, both compared with the baseline GMM-UBM method.

Index Terms: automatic mispronunciation detection (AMD), TempoRAl Patterns (TRAPs), Multi-layer Perceptron (MLP), fusion

1. Introduction

Automatic mispronunciation detection plays an important role in computer assisted language learning (CALL). In some certain interactive learning tasks especially pronunciation learning, the correctness of uttering words or phones is more important than a single proficiency score for language learners. Therefore, the aim of automatic mispronunciation detection is to automatically pinpoint the mispronunciations.

1.1. Problem definition in terms of hypothesis test

Automatic mispronunciation detection can be regarded as a hypothesis testing problem. We can define it as judging the question of "Testing pronunciation is incorrect?", and the corresponding Null hypothesis and Alterative hypothesis are as follows:

H_0: $\alpha \geq \alpha_T$ (i.e. Testing pronunciation is correct)

H_a: $\alpha < \alpha_T$ (i.e. Testing pronunciation is incorrect)

Here, the space of Rejection region is determined by a threshold α_T. If the test result falls in the Rejection region, H_0 is rejected, else H_0 is accepted. In this paper, our task is to explore testing variable α to effectively measure the pronunciation characteristics of speech samples, and choose an appropriate decision threshold. In many mispronunciation detection methods, the posterior probability (PP) is used as a common form of the testing variables. However, different kinds of PP derived from different methods result in different performance, so our objective is to develop some new kinds of PP for mispronunciation detection.

1.2. Motivations for the new approach

In fact, there are two key problems for mispronunciation detection: pronunciation feature and error detection method. The feature should be effective and robust to describe pronunciation quality, and the detection method should be able to separate the samples into correct classes. That is, the performance of mispronunciation detection can be improved by exploring new features and using powerful classifiers.

As one of the classic generative classifiers, Gaussian mixture model (GMM) is good at describing the distribution characteristics of speech, but with lower ability for pattern discrimination. In [1], the GMM output posterior probability derived from HMM (Hidden Markov Models) framework is used for phone-level pronunciation error detection. The formant and its varying trajectory are more useful for vowel discrimination. In [2], formant information was introduced into English vowel assessment for Chinese in Taiwan. Support vector machine (SVM), as a very efficient discriminative classifier, has been widely used in many tasks. Bolanos [3] utilized segmental features and SVM classifiers for the evaluation of children's speech. But the traditional RBF (Radial basis function) 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 practical applications. Instead of RBF, the GLDS (Generalized linear discriminant sequence) based SVM method with a novel model training strategy was introduced into mispronunciation detection in [4].

Recently, the long-time TempoRAl Patterns (TRAPs) and Multi-layer Perceptron (MLP) hybrid method has been successfully used in many fields [6]. Since different features have different advantages in depicting the characteristics and qualities of speech, it is obvious that making full use of speech features will bring progress. Therefore, in this paper, TRAPs have been introduced as pronunciation features. Besides, introducing new detection methods and merging multiple methods are also encouraged, so we use the powerful classifier MLP to get the posterior probability of testing phone.

The rest of this paper is organized as follows. Section 2 presents the TRAPs-MLP system in detail. Section 3 explores the fusion of several detection methods. Experiment results

and analysis are given in Section 4. Finally, in Section 5, some conclusions are drawn.

2. TRAPs-MLP based method

In this part, we introduce TRAPs features and MLP system into mispronunciation detection task, and then recommend how to get the final posterior score for each testing phoneme with boundary information and MLP state posterior outputs.

2.1. Exploring temporal information with TRAPs

As we know, Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are the major features in speech recognition and pronunciation quality evaluation, while the pitch and formant are more important for vowel discrimination.

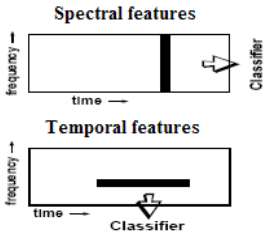

Figure 1: *Spectral features Vs. Temporal features.*

The TempoRAl Patterns (TRAPs) have gained some remarkable improvements in phoneme recognition recently by employing long-term time sequence of speech [5]. A drawback of spectral features is that they are quite sensitive to environment changes, and they are not more robust for different persons in pronunciation evaluation. On the contrary, the temporal features try to employ the information of the temporal domain, and find acoustic correlates of phonetic categories in speech spectrum. Figure 1 illustrates the difference between spectral features and temporal features (cited from reference [5]).

2.2. TRAPs-MLP system

While HMM posterior probability based detection method is a dominant approach in most state-of-the-art pronunciation error detection systems, neural networks (NN) are known as one of the most powerful nonlinear methods for pattern recognition, optimization, and forecasting. NN is an interconnected group of artificial neurons that uses a computational model for information processing based on a connectionist approach. NN is widely used in many fields of engineering, in particular for recognition, classification and prediction.

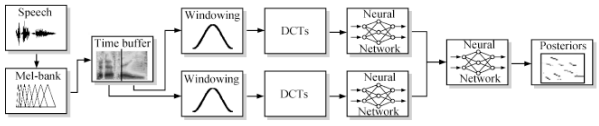

Figure 2: *The framework of TRAPs-MLP system.*

Although NN has been shown to be quite powerful in static pattern classification, their formalism is not very well suited to addressing pronunciation evaluation. In fact, for speech signals, there is a time dimension which is highly variable and difficult to handle directly in NN. In view of classification, the mispronunciation detection problem can be stated as follows: how can an input sequence (e.g. a sequence of spectral derived coefficients) be properly classified into corresponding groups (i.e. correct pronunciation and incorrect pronunciation) when different input sequences are not synchronous, since there are usually multiple inputs associated with each phone or word? Under these circumstances, TRAPs can provide synchronous input vectors for NN, so the combination of TRAPs and NN is a natural solution.

In past few years, several neural network architectures have been developed, here we choose feed-forward Multi-layer Perceptron (MLP) in our experiment.

In our work, in order to increase processing speed and reduce the huge complexity of MLP, a simplified version of the TRAPs-MLP system is adopted for state posterior outputs [6]. As shown in Figure 2 (cited from reference [6]), the temporal trajectory was split into two parts: left context part and right context part. This allows for more precise modeling of the whole trajectory while limiting the size of the model (i.e. reducing the number of weights in the MLP). Both parts are processed by DCT to de-correlate and reduce dimensionality. The feature vector which is feed to MLP is created by concatenation of vectors over all filter bank energies. Two MLPs are trained to produce the phoneme state's posterior probabilities for both context parts. As a merger, the third MLP produces final set of state posterior probabilities. Three state based HMM models are used, so each MLP produces 3 posteriors for the beginning, the center and the end of a phoneme.

2.3. Posterior probability output of testing phoneme by dynamic programming

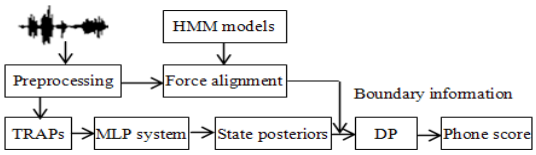

Figure 3: *Mispronunciation detection on phone level.*

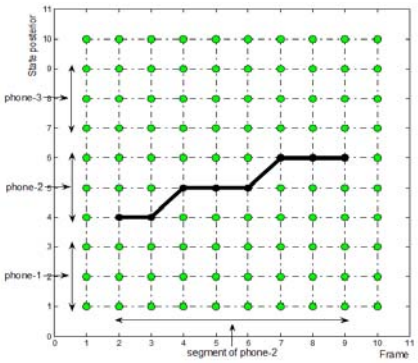

Figure 4: *PP output of testing phone by DP.*

TRAPs-MLP system includes the inherently discriminant nature of the training algorithm for probability estimation, but it has not been demonstrated to function well for pattern classification because of the need for accurate boundaries of speech signals. Therefore, in Figure 3, we get the boundary information by force alignment step in advance (using HMM's Viterbi decoder without any language model). Then, with given boundaries, the corresponding confidence score for each phone can be obtained by dynamic programming based on state posterior outputs from MLP system, which is fully illustrated in Figure 4. That is, the phone boundary is obtained by HMM, while the state boundary is gained by DP operation of MLP state posteriors.

3. System fusion strategy

For universal cases, multiple features are passed through multiple detection systems to produce multiple scores for measuring pronunciation quality. Such scores are complementary to each other, so we further investigate to merge them with various linear or nonlinear methods, and some best results can be gained. For linear fusion method, a final quality score is obtained by weighting different output scores, and then a decision is made based upon whether the score is above or below a predetermined threshold.

Data fusion can be generally divided into two kinds of input fusion and output fusion. Using different input features belongs to the kind of input fusion, and the output fusion is the utilization of the outputs of several systems to form a final result. Output fusion usually performs better than input fusion [7]. 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.

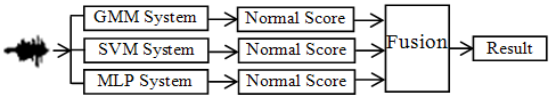

Figure 5: *Architecture of the fusion system.*

In this paper, the GMM-UBM and GLDS-SVM systems from reference [4] are used for combination with the TRAPs-MLP method, just as shown in Figure 5. The linear fusion of above GMM, SVM and MLP systems can be stated below,

$$score = w_{gmm} \cdot score_{gmm} + w_{svm} \cdot score_{svm} + w_{mlp} \cdot score_{mlp} \qquad (1)$$

where $score_{gmm}$, $score_{svm}$ and $score_{mlp}$ are output scores from GMM, SVM and MLP systems, and all the scores have been normalized into universal score domain (i.e. normal scores). Meanwhile, w_{gmm}, w_{svm} and w_{mlp} are corresponding weights for above three systems.

As GMM, SVM and MLP systems have different modeling structures, such a situation can yield significant complementary information and make these three systems good candidates for merging.

4. Experiments and results

4.1. Experimental setup

The experiments are carried out on a large English speech corpus collected from lots of Chinese students. We separate the corpus into training set and testing set, as shown in Table 1 in detail. 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 detection method, and the experiment results are obtained on the testing set. Due to the difficulties of collecting actual mispronunciations with manual labels, we generate mispronunciation samples for testing set with modified transcripts. The modification is carried out by substituting some standard phones with the other ones within the phone set. Though this course of action has a little difference towards the practical situation, it is a compromise way to a certain extent.

Table 1. *Speech corpus in the experiment.*

Data Set	# male	# female	# total	# hours
Training set	100	200	300	16.4
Testing set	28	50	78	4.8
All set	128	250	378	21.2

The speech signals are divided into 25ms long frames with 10ms shift. The basic feature includes 13 MFCC with logarithm energy, and their first and second order derivatives. The acoustic HMM models used for force alignment are trained by 120 hours with speech corpus.

TRAPs-MLP models are trained on the same 35 hours subset of the training set as HMM models. TRAPs with 23 spectral bands have expanded 15 frames for both left and right context parts, and then are reduced to 11 coefficients for each band by Hamming window and DCT. That is, there are two 253-dimension vectors into the left or right context MLP. All MLPs output 133 posteriors, including 3 states of each phoneme and one state of silence. So, the two 133-dimension vectors are joined as a 266-dimension vector into the merger MLP at last. And the context and merge MLPs are assigned 1000 and 4000 hidden layer units, respectively. Besides, the Quicknet tool from the SPRACHcore package [8], employing three layer perceptron with the softmax nonlinearity at the output, was used in our experiment.

We use BEEP dictionary [9], which has 44 phonemes for the phone set, and is favorable for pronunciation diagnosis. Compared with CMU phone set, BEEP has eight diphthongs, i.e. /au/, /ai/, /ɛə/, /ei/, /iə/, /əu/, /ɔi/ and /uə/, it is more appropriate to use longer temporal features such as TRAPs.

4.2. Evaluation criterion

		STANDARD: Correct	STANDARD: Incorrect
DETECT	Correct	True Acceptance	False Rejection
	Incorrect	False Acceptance	True Rejection

Figure 6: *Confusion matrix for detection results.*

The detection results can be illustrated as a 2×2 confusion matrix in Figure 6, and there are two error types for any detection tasks. In order to measure the performance of our detection systems, we consider two measures of false acceptance rate (FAR) and false rejection rate (FRR). Any mispronunciations detected as correct is treated as false rejection (FR) or missing detection, and any correct pronunciations detected as error is treated as false acceptance (FA) or false alarm. With confidence score and different thresholds, each testing phoneme can be decided as mispronunciation or not. To fully show the changing performance of FAR and FRR with different thresholds, Detect Error Tradeoff (DET) curve and Equal Error Rate (EER) are used in following experiments.

4.3. Performance comparison of different methods

In our experiment, the result by GMM-UBM system is treated as baseline. Figure 7 illustrates the DET curves of three methods and their fusion systems, and the detail results of EER are listed in Table 2. We can see that TRAPs-MLP method can improve the performance significantly. So TRAPs-MLP is a good choice for mispronunciation detection.

Meanwhile, by combining TRAPs-MLP with GMM-UBM or GLDS-SVM system, the performance is further

improved. With the complementary of above three systems, a best relative improvement of 48.32% in EER reduction is obtained. However, we can also see that the fusion of GMM-UBM and TRAPs-MLP has little improvement compared to other fusion systems. This implies that the complementarity between GMM-UBM and TRAPs-MLP is relatively lower.

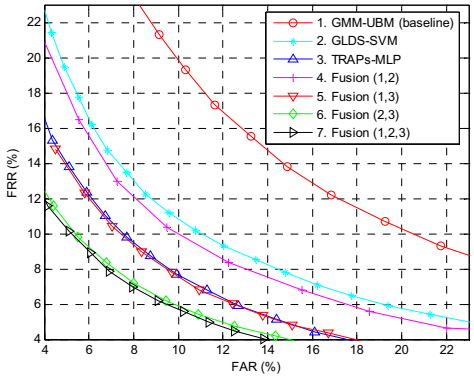

Figure 7: D*ET curve for different methods.*

Table 2. *EER comparison for different methods.*

	Method	EER (%)	Rel. imp. (%)
1	GMM-UBM (baseline)	14.32	—
2	GLDS-SVM	10.50	26.68
3	TRAPs-MLP	8.73	39.04
4	Avg fusion of 1, 2	9.92	30.73
5	Avg fusion of 1, 3	8.66	39.53
6	Avg fusion of 2, 3	7.60	46.93
7	Avg fusion of 1, 2, 3	7.40	48.32

4.4. Weighting values for fusion systems

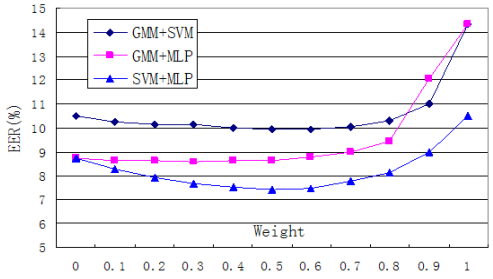

Figure 8: *EER curves with different fusion weights.*

In order to review the contribution of different methods for performance improvement, we expand the average fusion form by weighting fusion values. As show in Figure 8 (X-axis is the first method's weight value of each fusion system), several weighting values of the three fusion systems are tested. It can be seen that the performance varies with different fusion weights. GMM-UBM has lower complementary information for GLDS-SVM or TRAPs-MLP system, while GLDS-SVM and TRAPs-MLP are much more complementary for combination. Besides, we can see that the average fusion is always a better choice when there is no development corpus for parameter tuning.

4.5. Computation and storage efficiency

Time and storage performance of each system is shown in Table 3. For time consumption, the real-time ratio between computation time and input speech's time is calculated. For storage consumption, the size of used models is recorded. Results show that the GLDS-SVM system is slightly faster than others, and has a smaller model for each phone with the same size. The computation time of TRAPs-MLP system is relatively longer due to the complexity of neural networks, and it also has a bigger model. That is, compared with its higher detection performance, TRAPs-MLP system needs further improvements on both time and storage consumption.

Table 3. *Time and storage consumption.*

	Method	Computing time / speech time (X)	Model size (MB)
1	GMM-UBM	0.012	1.44
2	GLDS-SVM	0.010	4.14
3	TRAPs-MLP	0.053	13.40

5. Conclusions

In this paper, we focus on a TRAPs based MLP method for mispronunciation detection and explore its complementarity and fusion with other systems. Experiment results show that the TRAPs-MLP method can provide a respectable performance, but its computation speed and model storage consumption need further improvements. Consequently, we can draw some conclusions: firstly, TRAPs features can improve the ability of describing pronunciation quality. Secondly, MLP can provide discriminant-based learning, that is, models are trained to minimize the error rate while maximizing the distance between the correct model and its rivals. Thirdly, the system fusion strategy is more efficient for different modeling based methods due to their complementary information.

6. Acknowledgements

This work was supported by the National High Technology Research and Development Program of P. R. China (863 Program) under Grant 2006AA010103.

7. References

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

[2] Chen, J. C., Hsu, W. T., Lyu, R. Y., Chiang, Y. C., "Formant-based English Vowel Assessment For Chinese in Taiwan", in Proc. of ICSLP, pp. 1375-1378, Pittsburgh, Pennsylvania, USA, 2006.

[3] Bolanos, D., Ward, W., Wise, B., Vuuren, S. V., "Pronunciation Error Detection Techniques for Children's Speech", in Proc. of Interspeech, pp. 1725-1728, Brisbane, Australia, 2008.

[4] Li, H. Y., Liang, J. E., Wang, S. J., Xu, B., "An Efficient Mispronunciation Detection Method Using GLDS-SVM and Formant Enhanced Features", in Proc. of ICASSP, pp. 4845-4848, Taipei, Taiwan, 2009.

[5] Hermansky, H., Sharma, S., "TRAPs — Classifiers of Temporal Patterns", in Proc. of ICSLP, Sydney, Australia, 1998.

[6] Schwarz, P., Matejka, P., Cernocky, J., "Towards lower error rates in phoneme recognition", in Proc. of TSD 2004, number ISBN 87-90834-09-7, pp. 465-472, Brno, Czech Republic, 2004.

[7] Kittler, J., Hatef, M., Robert, P. W., Jiri, M., "On Combining Classifiers", IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(3): 226-239, 1998.

[8] ICSI Speech Group, "The SPRACHcore software packages", Online:http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html, accessed on 13 May 2008.

[9] Engineering Department of Cambridge University, "BEEP Dictionary Description and Availability", Online: http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html,accessed on 8 Aug 2008.