IEICE TRANS. INF. & SYST., VOL.E95–D, NO.12 DECEMBER 2012

PAPER

A Forced Alignment Based Approach for English Passage Reading Assessment

Junbo ZHANG[†a)], **Fuping PAN**[†b)], **Bin DONG**[†c)], ***Nonmembers***, **Qingwei ZHAO**[†d)], ***Member***, ***and*** **Yonghong YAN**[†e)], ***Nonmember***

SUMMARY This paper presents our investigation into improving the performance of our previous automatic reading quality assessment system. The method of the baseline system is calculating the average value of the Phone Log-Posterior Probability (PLPP) of all phones in the voice to be assessed, and the average value is used as the reading quality assessment feature. In this paper, we presents three improvements. First, we cluster the triphones, and then calculate the average value of the normalized PLPP for each classification separately, and use this average values as the multi-dimensional assessment features instead of the original one-dimensional assessment feature. This method is simple but effective, which made the score difference of the machine scoring and manual scoring decrease by 30.2% relatively. Second, in order to assess the reading rhythm, we train Gaussian Mixture Models (GMM), which contain the information of each triphone's relative duration under standard pronunciation. Using the GMM, we can calculate the probability that the relative duration of each phone is conform to the standard pronunciation, and the average value of the probabilities is added to the assessment feature vector as a dimension of feature, which decreased the score difference between the machine scoring and manual scoring by 9.7% relatively. Third, we detect Filled Pauses (FP) by analyzing the formant curve, and then calculate the relative duration of FP, and add the relative duration of FP to the assessment feature vector as a dimension of feature. This method made the score difference between the machine scoring and manual scoring be further decreased by 10.2% relatively. Finally, when the feature vector extracted by the three methods are used together, the score difference between the machine scoring and manual scoring was decreased by 43.9% relatively compared to the baseline system.

key words: *CALL, automatic assessment, forced alignment, formant*

1. Introduction

At present, oral test has become an important part of English proficiency tests. With the increasing scale of oral test, a large number of oral test data of examinees shall be assessed, which needs a large scale of human resources. Due to the lack of scoring teachers and the defect of the subjective in manual scoring, there is demand for automatic scoring systems. A common question type in English oral tests is Reading Aloud, which required the examinees to read aloud a passage generally more than 100 words. The voice from the examinee and the reading text are input into the automatic reading quality assessment system, after the system has assessed the voice of the examinee, the system outputs a score of the examinee's reading quality. During the process of the automatic assessment, the scoring principals must be consistent with those of the manual scoring. Table 1 shows an example of a typical manual scoring principal.

There are many research achievements of pronunciation quality assessment in the phone or word level, in which the studies of Cambridge [1]–[4] and SRI [5]–[8] are the most representative. The study of pronunciation quality assessment in the phone or word level has become mature, the state-of-art method is to measure the pronunciation quality of phones by Phone Log-Posterior Probability (PLPP) [4], and to judge the pronunciation quality of words with weighted average of each phone's PLPP [8]. Compared to the pronunciation quality assessment in the phone or word level, the reading quality assessment studied in this paper is a much harder task, whose challenge is that although there are only a few grades for score (5 grades in the experiments of this paper), the reading quality is influenced by pronunciation, fluency and many other factors. So it is difficult to find an one-dimensional feature of high correlation with the manual score we have to extract multi-dimensional feature vector, and then map them to reading quality score in multi-dimensional space. Currently, there are few study of reading quality assessment. There is a study done by Zechner et al, whose system was based on large vocabulary continuous speech recognition (LVCSR) system, and phonetic sounds to be assessed were sent to LVCSR engine, and according to analyzing the recognition results, error ratio of recognition, numbers of silence, the multi-dimensional assessment features of reading quality is extracted, and then the features

Table 1 Manual scoring principal.

Score	Description
5	All words pronounced good. Fluency.
4	A few words pronounced wrong. Relatively fluency.
3	Many words pronounced wrong. Not very fluency.
2	Most words pronounced wrong. Not fluency.
1	Pronunciation is hardly to understand.
0	Noise, silence or something uncorrelated.

Manuscript received September 30, 2011.
Manuscript revised May 17, 2012.

[†]The authors are with Key Laboratory of Speech Acoustics and Content Understanding, China.
a) E-mail: jbzhang@hccl.ioa.ac.cn
b) E-mail: fpan@hccl.ioa.ac.cn
c) E-mail: bdong@hccl.ioa.ac.cn
d) E-mail: qwzhao@hccl.ioa.ac.cn
e) E-mail: yyan@hccl.ioa.ac.cn

DOI: 10.1587/transinf.E95.D.3046

are mapped to reading quality score [9]–[12].

Our previous systems is also based on Automatic Speech Recognition (ASR) technique, but we use forced alignment instead of LVCSR [13], which is different from the Zechner's method. Both the forced alignment and LVCSR have their advantages. Methods based on forced alignment have the advantages in calculating phone pronunciation quality because there is mature forced alignment based quality assessment algorithms [4]. This study also focuses on the method based on forced alignment and uses the previous system as the baseline system. The baseline system calculate PLPP as the reading quality assessment feature, and use this feature and manual score to train the SVM model, then map the feature to the reading quality score. Whereas, there is some deficiency in this method, first of all, the method of extraction feature may rough. Second, the source of the feature is single and some of the useful information is not extracted. Therefore, this paper attempts to make the method of extracting feature based on PLPP more reasonable, and tries to extract more useful information which is helpful to assessment.

The rest of this paper is as follows: the data set of this study is introduced in Sect. 2; the baseline system is described and the definition of PLPP is introduced in Sect. 3; in Sect. 4, the PLPP based reading quality feature extracting method is improved, and a GMM based assessment algorithm of phone relative duration and a FP detection algorithm are introduced; in Sect. 5, we represent the performance improvement resulted from these method by experimental; and Sect. 6 concludes the paper.

2. Corpus

We recorded 7000 English passage reading data from a number of junior middle school students whose native language are Chinese, in which the gender ratio is 1:1. Each student was required to read a English passage about 100 words. The passage is from the English textbook of junior middle school. We employed English teachers to score these voice according to Table 1. To assure the objectiveness of the manual scoring, we employed three English teachers to score, and took the middle score of the three teachers as the manual score. Because the reading quality assessment system described in this paper needs data to train, we took the data from 4000 students to train, and the other data from 3000 students was used to test. We made the distribution of the manual score and the gender ratio of the speakers maintain consistency between the training set and testing set. In addition, we also chose two groups of native English speaker to record for about 100 hours for relative experiment in this paper. We refer to these two data sets as Native Speech (NS) and Alternative Native Speech (ANS).

3. Baseline System

The workflow of the baseline system [13] is shown in Fig. 1. The system is consist of two parts, which are the training part and the testing part. The training part is executed only once to generate SVM model, and then this model is used for testing. In the training part, the silence at the beginning and ending of the voice is cut, and then the waveform of the voice is transformed into Perceptual Linear Prediction (PLP) [14] vector, and then the PLP vector is sent to Viterbi decoder to perform forced alignment with the reading text. The structure of the decoding network of the Viterbi decoder is the series of the words in the reading text. After the forced alignment, the Phone Log-Posterior Probability (PLPP) [4] is calculated as the reading quality assessment feature, and the SVM model is trained with these features and the manual scores. The process of the testing part and training part are substantially the same, but the difference is that the output of the training part is the SVM model, while the output of the testing part is the reading quality score.

PLPP is a measure of the closeness between the actual

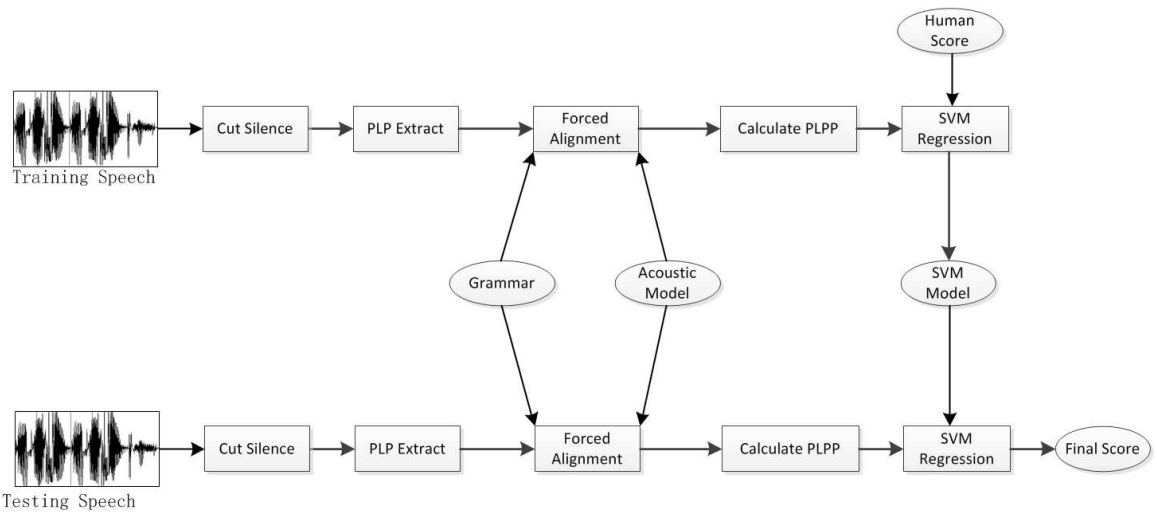


Fig. 1 The baseline system.

pronunciation and the standard pronunciation of a certain phone, which is the source of the reading quality assessment feature of the baseline system. The formula of PLPP is as Eq. (1):

$$PLPP(p) = \log P(p|O) = \log \frac{p(O^{(q)}|q)}{\sum_{p \in Q} p(O^{(q)}|p)} \quad (1)$$

where O is the forced-aligned observation sequence of the phone p; Q is the set of phones and q is a phone; $P(p|O)$ is the phone posterior probability.

4. The Improved System

4.1 Cluster PLPP

As described in Sect. 3, the baseline system calculates PLPP of all phones of the voice to be assessed and uses their average value as the reading quality assessment feature, which was appointed as an effective method in the paper of Ge [13]. However, there are improvement space for this method. To introduce the improvement, an experiment is shown: we forced aligned the data set NS with the corresponding reading text, and calculated PLPP of all phones in the data set, and then performed statistics of the average value of PLPP of each triphone. Because the data set NS was recorded from native speakers, the average PLPP value of each triphone here represents the ideal value of PLPP of the triphone in good condition of pronunciation. Figure 2 shows the statistic result for some triphones which were selected randomly. In this figure, we can see that there is big difference of the ideal value of PLPPs of these triphones. Thus, a big error might be brought if we simply calculate the average value of PLPP for triphones, which are with different ideal value of PLPP.

A possible method to resolve this problem is doing normalization for PLPP of all triphone, and we tested the z-score algorithm, shown as Eq. (2):

$$PLPP_{norm}(p) = \frac{PLPP(p) - IPLPP(p)}{SD(p)} \quad (2)$$

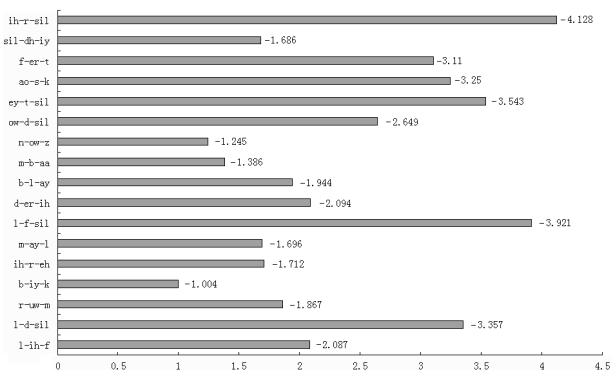


Fig. 2 Ideal values of PLPP of some triphones.

where p represents a certain triphone, $IPLPP(p)$ is the ideal value of PLPP of p, and $SD(p)$ is the standard deviation of PLPP of all triphone p in the voice.

Our experiment shows that the normalization indeed improvement the performance of the baseline system, and the Sect. 5.2.1 in this paper provides the result of the improvement from the normalization. However, although normalization is conducive to calculate the average value of PLPP, a valuable information may be lost in normalization, that is, the ideal value of PLPP of each triphone. In order to explore whether the information of the ideal value of PLPP of each triphone is useful for reading quality assessment, we analysis Eq. (1). The numerator of Eq. (1) represents the likelihood of that the observation sequence O is in the triphone p, and the denominator represents the sum of the likelihood that the observation sequence O is in each triphone. It can be known that for a certain triphone p, if the ideal value of PLPP of p is higher, the likelihood that the voice of the triphone p will be closer to the sum of likelihood of that the voice of the phone p in all triphones, that is, that triphone will hard to confuse with other triphones during the process of Viterbi decoding. Therefore, we can deduce that the contributions of triphones with different ideal value of PLPP to the pronunciation quality of the whole passage are different, and separate treatment may improve the performance of the reading quality assessment. To validate our deduction, we cluster triphones to several classifications according to the value of the ideal average value of PLPP, and extract feature for each classification. In this way the information of the ideal value of PLPP is preserved, and triphones with different ideal value of PLPP has been treated separately. The detail of this method is to do K-means clustering for triphones, and the distance function of every two triphones is the absolute value of the difference of the ideal values of PLPP of the two triphones, shown in Eq. (3). The normalized average value of PLPP is obtained as a dimension of feature for each classification of triphone, in which normalization uses the method as shown in Eq. (2):

$$D(p_1, p_2) = |IPLPP(p_1) - IPLPP(p_2)| \quad (3)$$

where $D(p_1, p_2)$ is the distance function of triphone p_1 and p_2 in K-means clustering.

In this way, if the triphones are clustered as N classifications, then it will get N-dimensional feature to replace the one-dimensional feature of the baseline system. This method is referred as Cluster PLPP (CPLPP) in this paper. Regarding the influence of the classification numbers, the experimental result is shown in Sect. 5.2.1.

Another problem is whether the method to cluster by ideal value of PLPP is stable or not. Assume that another data set was replaced, if the clustering result changed a lot, thus this clustering was unstable and this method did not have the expandability. To validate the stability of this clustering method, we used another data set (data set ANS) with good pronunciation, whose recording source is different from that of data set NS. We conducted statistics of the average values of PLPP for the two data set in order to en-

sure the average value of PLPP for each triphone is stable when pronunciation is good. The results is shown in the Table 2.

The comparison results show that the difference of the average values of PLPP between the two data sets is small for most triphones, which demonstrates that it is feasible to cluster according to average value of PLPP.

4.2 Phone Relative Duration Model

The phone relative duration is shown in Eq. (4).

$$dur_r(p) = \frac{dur(p)}{\sum_k dur(p_k)} \tag{4}$$

where $dur_r(p)$ is the relative duration of a certain phone p, $dur(p)$ is the duration of p, and $\sum_k dur(p_k)$ is the total duration of the word which includes the phone p.

There is a certain rule of the of phone relative duration inside a certain word. For example, Fig. 3 shows the ratio of phone duration of a word "home" when its ratio of phone duration is reasonable or unreasonable. In the figure, sample 1 is the pronunciation with reasonable ratio of phone duration, and sample 2 is the pronunciation with unreasonable ratio of phone duration. The pronunciation of the phone "m" is too long in sample 2, which reflects the unconfident of the speaker's pronunciation.

Table 2 The comparison of the ideal value of PLPP between the date set NS and ANS.

Range of the difference	Percentage of triphones
[0, 0.5)	23.5%
[0.5, 1.0)	58.3%
[1.0, 1.5)	9.5%
[1.5, 2.0)	7.3%
[2.0, +∞)	1.2%

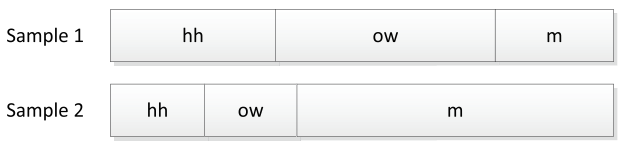


Fig. 3 Phone relative durations of the word "home".

Whether the phone relative duration is correct is an important indicator of the reading quality. However, from Eq. (1) we can see that the phone relative duration can not be reflected by PLPP. For this reason, it is necessary to design a feature to measure whether the relative duration is correct. We forced aligned the data set NS and the corresponding reading text, and calculated the relative duration values of all triphones following Eq. (4). These relative duration values are used to train Gaussian mixture models (GMM). We train models for all triphones. The models provide the probability of a certain triphone taking a certain relative duration value under good pronunciation. This GMM is referred as PRDM in this paper. Posterior probabilities of the relative duration of triphones are calculated with PRDM. The higher the posterior probability is, the closer the relative duration between the phone and the good pronunciation is. The average value of the posterior probabilities of the relative duration of each phone is used as a feature. The feature is added to the reading quality assessment feature sequence.

4.3 Filled Pause Detection

Some unskilled speaker pronounce unconsciously meaningless pronunciations such as "eh", "ah", which are called as Filled Pause (FP). We consider the duration of FP is an important indicator of the reading quality. Audhkhasi et al. studied the FP phenomenon, and pointed that the part of FP's first formant (F1) is more stable than normal voice [15]. Figure 4 shows the F1 curve of two voices, where the reading text of the two voices are same, but the pronunciation of Fig. 4 (a) is good, clear and aloud; while the pronunciation in Fig. 4 (b) is shuffling and there exists "eh", "ah" filling pauses. The F1 curve of the FP parts are marked in the figure. We can see the F1 curve in FP parts approximate to flat.

In this paper, the stability of a formant at a given frame is quantified by computing the Standard Deviation (SD) of the formant value over a window of W frames centered on

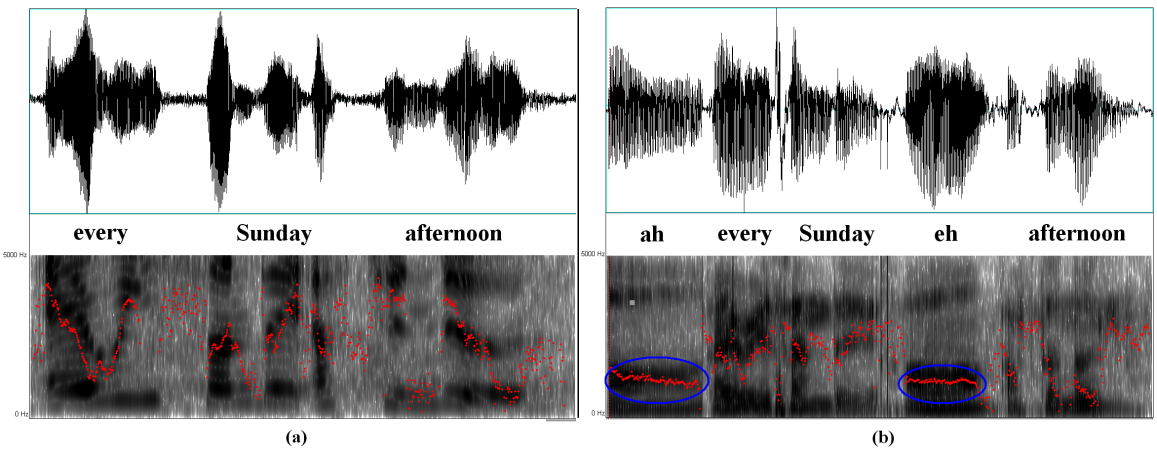


Fig. 4 F1 curves of two voices (a) Normal speech (b) Speech with filled pauses.

the frame. The SD is calculated as Eq. (5):

$$F1SD(i) = \sqrt{\frac{1}{W-1}\sum_{W}(F_i - \bar{F})} \quad (5)$$

where $F1SD(i)$ is the SD of the first formant of the i-th frame, W is the frame numbers selected around the i-th frame, F_i is the F1 value of the i-th frame, and $\bar{F}$ is the F1's average value inside this W frame. In this paper we set $W = 7$ as an experience value.

The smaller the F1SD value is, the less fluctuation of the curve is. We calculate the percentage of frames whose F1SD are smaller than a certain threshold, and the percentage value is used as a dimension of reading quality assessment feature, which is added to the reading quality assessment feature vector.

5. Experiment

5.1 Performance Metrics

This paper use three indicators to measure the performance of the reading quality assessment system. The indicators are: Average Scoring Absolute Difference (ASAD), Recall Rate (RR) and Correlation Coefficient (CC).

To describe ASAD, Scoring Absolute Difference (SAD) will be introduced firstly. SAD is the difference of the absolute value between the manual score and machine score as shown in Eq. (6).

$$SAD_i = |sc_i - sh_i| \quad (6)$$

where sc_i is the machine score of the i-th sample, and sh_i is the manual score.

ASAD is the average value of SAD as shown in Eq. (7).

$$ASAD = \frac{1}{N}\sum_{N} SAD_i \quad (7)$$

where N is the sample number of the testing data set.

RR is used to measure the percentage of the samples whose assessment error is larger than a certain threshold value. The calculation of RR is show in Eq. (8).

$$RR = \frac{1}{N}\sum_{N} sgn(SAD_i - T) \quad (8)$$

where the T is the recall threshold value. According to Table 1, we set $T = 2$. The function $sgn(x)$ is defined as Eq. (9):

$$sgn(x) = \begin{cases} 1, x > 0 \\ 0, x \le 0 \end{cases} \quad (9)$$

CC refers to the correlative coefficient between machine score and manual score. In the three indicators, we shall decrease ASAD and RR, and increase CC as much as possible.

5.2 Experiment Result

5.2.1 Performance of CPLPP

This section compares the assessment performance of the reading quality assessment features from the original PLPP, normalized PLPP and CPLPP.

To achieve the optimum property of CPLPP, the classification numbers of clustering shall be determined firstly. We tested the target classification numbers of clustering, and tested the performance for each classification number as shown in Fig. 5, which shows that the performance was best when the target classification numbers was 7.

Table 3 compares the performance among using the average value of the original PLPP as feature, using the average value of the normalized PLPP as feature, and using the multi-dimensional CPLPP as feature vector. The performance of the normalized PLPP feature is better than that of the original PLPP, while the performance of the CPLPP feature is significantly improved further. The normalized PLPP feature decreased ASAD of the baseline by 9.2% relatively, while the CPLPP features decreased ASAD of the baseline by 30.8% relatively. On RR and CC, the performance of the normalized PLPP and CPLPP have been increased correspondingly.

Normalized PLPP feature improves the performances of the reading quality assessment, that is because that normalization avoids the error brought from the averaging of PLPP of phones with different ideal value of PLPP. However, the process of normalization loses the information of the ideal value of PLPP, while the classification process of CPLPP supplements this information, moreover the classification process makes different triphone to be treated separately, thus it provides more information in the feature vector. The result shows that the performance increase brought

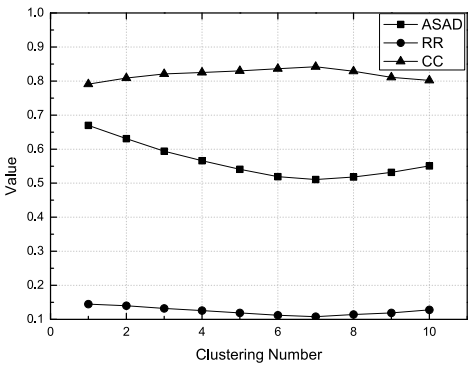


Fig. 5 Comparison of different target classification numbers.

Table 3 Comparison of original PLPP, normalized PLPP and CPLPP.

	ASAD	RR	CC
Baseline	0.738	0.181	0.749
Norm PLPP	0.670	0.145	0.791
CPLPP	0.511	0.108	0.842

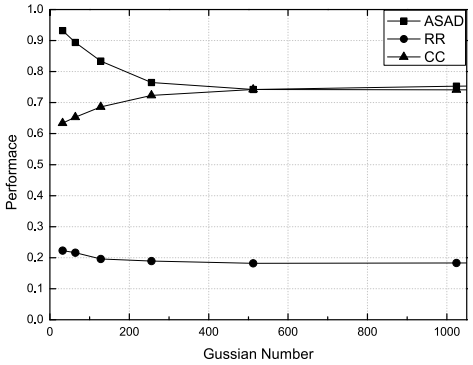


Fig. 6 PRDM performance with different gaussages.

Table 4 Performance of PRDM.

	ASAD	RR	CC
Baseline	0.738	0.181	0.749
PRDM (512 GMM)	0.742	0.182	0.741
Baseline + PRDM	0.637	0.131	0.810

Table 5 Performance of filled pause detection.

	ASAD	RR	CC
Baseline	0.738	0.181	0.749
FP	1.192	0.363	0.514
Baseline + FP	0.689	0.153	0.796

by CPLPP is much higher than that of normalization, and this result sufficiently illustrates the importance of the information of the ideal value of PLPP.

5.2.2 Performance of PRDM

This section shows the influence of the PRDM feature on performance. Figure 6 shows the assessment performance of selecting different gaussages of the GMM. It can be seen from Fig. 6 that the best performance was achieved when the gaussages was 512.

Table 4 shows the performance improvement from the PRDM feature. When the PRDM feature was used alone, it achieved the close assessment performance as the baseline system, but if the assessment feature of the baseline system and the PRDM feature are used together, then there was a relative 13.8% of performance improvement for ASAD, and there are corresponding performance improvement on RR and CC. This illustrates that the pronunciation and the relative duration of phones are both important for reading assessment, and the importance of the both are approximately equivalent, so there is approximately equivalent performance when each feature was used alone. When use the two features together, the performance improved obviously, which indicates that there are strong complementarily between these two features.

5.2.3 Performance of Filled Pause Detection

Table 5 shows the performance of the reading quality assessment feature extracted by FP detection. When we used the feature extracted by FP detection results alone, the performance was comparatively poor, indicating that the FP frequency can not fully reflect the reading quality. However, after integration with the baseline system, the performance was improved significantly, which indicated that the FP information is an effective supplement of reading quality assessment.

Table 6 Performance of the final system.

	ASAD	RR	CC
Baseline	0.738	0.181	0.749
CPLPP	0.511	0.108	0.842
CPLPP + PRDM	0.461	0.089	0.867
CPLPP + PRDM + FP	0.414	0.083	0.875

5.2.4 Performance of the Final System

This section shows that the performance was improved after adding the methods described in this paper to the baseline system. As shown in Table 6, after the PLPP feature of the baseline system was replaced by the CPLPP feature vector, ASAD was decreased by 30.8% relatively. After the PRDM feature was added, ASAD was decreased by 9.7% relatively. At last, after the FP detection feature was added, ASAD was decreased by 10.2% relatively. Finally, ASAD of the baseline system was relatively decreased by 43.9%, RR was relatively decreased by 54.1%, and CC was relatively decreased by 14.4%. The performance improvement was very obvious, especially the improvement from CPLPP was the most significant. The results shows that these methods are efficient.

6. Conclusion

As demonstrated by the experiments in this paper, the ideal value of PLPP, phone relative duration and duration of FP are very important information for reading quality assessment. It is demonstrated in these experiments that the triphones with the different ideal value of PLPP result in the different effects in the reading quality assessment, and clustering can preserve the information of the ideal value of PLPP and provides more information for reading quality assessment; It can be considered as a stable clustering method by clustering triphone according to the ideal value of PLPP. GMM can model the relative durations of good pronunciation phones. The F1 curve of the filled pauses is approximates flat, so analysing the F1 curve can be used to detect filled pauses.

This paper uses the ideal value of PLPP to cluster for triphones, but there shall be some other effective clustering method, such as clustering based on linguistic knowledge. In addition, for the analysis of the formant, there shall be other methods such as linear fitting. We will test these in our further experiments.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275, 11161140319).

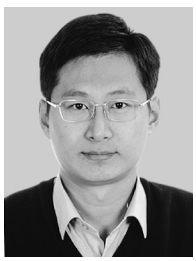
References

[1] S. Witt and S. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," Language Teaching and Language Technology Groningen, The Netherlands, 1997.
[2] S. Witt and S. Young, "Language learning based on non-native speech recognition," Fifth European Conference on Speech Communication and Technology, 1997.
[3] S. Witt and S. Young, "Performance measures for phone-level pronunciation teaching in call," Proc. Workshop on Speech Technology in Language Learning, pp.99–102, 1998.
[4] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech communication, vol.30, no.2-3, pp.95–108, 2000.
[5] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, pp.1457–1460, IEEE, 1996.
[6] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, pp.1471–1474, IEEE, 1997.
[7] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," Speech Commun., vol.30, no.2-3, pp.83–93, 2000.
[8] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The sri eduspeaktm system: Recognition and pronunciation scoring for language learning," Proc. InSTILL 2000, pp.123–128, 2000.
[9] K. Zechner and I. Bejar, "Towards automatic scoring of non-native spontaneous speech," Proc. Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp.216–223, Association for Computational Linguistics, 2006.
[10] K. Zechner, D. Higgins, and X. Xi, "Speechratertm: A construct-driven approach to scoring spontaneous non-native speech," Workshop on Speech and Language Technology in Education, 2007.
[11] K. Zechner, D. Higgins, X. Xi, and D. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," Speech Commun., vol.51, no.10, pp.883–895, 2009.
[12] K. Zechner, J. Sabatini, and L. Chen, "Automatic scoring of children's read-aloud text passages and word lists," Proc. Fourth Workshop on Innovative Use of NLP for Building Educational Applications, pp.10–18, Association for Computational Linguistics, 2009.
[13] F. Ge, F. Pan, C. Liu, B. Dong, S. Chan, X. Zhu, and Y. Yan, "An svm-based mandarin pronunciation quality assessment system," Sixth International Symposium on Neural Networks (ISNN 2009), pp.255–265, Springer, 2009.
[14] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," J. Acoustical Society of America, vol.87, no.4, pp.1738–1752, 1990.
[15] K. Audhkhasi, K. Kandhway, O. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken english," Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pp.4857–4860, IEEE, 2009.

Junbo Zhang received his BE from Beijing University of Chemical Technology. Now he is a Ph.D candidate of Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics (IOA), Chinese Academy of Science (CAS). His research interests include pronunciation quality assessing and speech recognition.

Fuping Pan received his Ph.D. in Information and Signal Processing from IOA, CAS, 2007. He is currently an Assistant Researcher in IOA. His research is focused on automatic pronunciation evaluation, speech signal processing and speech recognition.

Bin Dong received his Ph.D. in Information and Signal Processing from IOA, CAS, 2006. He is currently an Assistant Researcher in IOA. His research is focused on computer assistant language learning, automatic pronunciation evaluation, speech signal processing and speech recognition.



Qingwei Zhao received his Ph.D. in electronic engineering from Tsinghua University in 1999. Now he is an Associate Professor at ThinkIT Lab, IOA, CAS. Before joining CAS, he was with Intel as a Senior Researcher. His research interests include spontaneous speech recognition, keyword spotting and automatic pronunciation evaluation.

Yonghong Yan received his BE from Tsinghua University in 1990, and his PhD from Oregon Graduate Institute (OGI). He worked in OGI as Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998-2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. Currently he is a professor and director of ThinkIT Lab. His research interests include speech processing and recognition, language/speaker recognition, and human computer interface. He has published more than 100 papers and holds 40 patents.