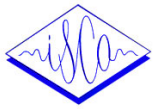

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

Automatic Assessment of Pronunciation Quality

Bin Dong, Qingwei Zhao, Jianping Zhang, Yonghong Yan

Institute of Acoustics, Chinese Academy of Sciences Beijing
{bdong, qzhao, jzhang, yyan}@hccl.ioa.ac.cn

ABSTRACT

Learning to speak a foreign language is not an easy task for many people. This paper describes approaches to automatic objective assessment of pronunciation quality. The approaches described here can be classified into two categories, text-dependent and text-independent, according to whether a teacher's voice presents. In the text-independent one, algorithms based on energy and pitch contour are introduced. Also, the average rate of variation in energy and pitch frequency, mean subtracted energy and pitch frequency are used as main features. Compared to previously reported approach using average phone segment posterior probabilities, the new approach achieves favorable performance on the same test set.

1. INTRODUCTION

Speech analysis plays an important role in computer-aided language instruction. It is usually required for this kind of language instruction to have reading, listening and text input functionalities. In addition, the feedback from the computer is very important for language learners, especially for beginners by which they can improve their pronunciation quality. So these algorithms about automatic assessment of pronunciation quality are necessary to satisfy this requirement.

An input utterance recorded from microphone can be converted into text by automatic speech recognition system. However such systems are not designed specially for evaluating a speaker's pronunciation in general. Therefore new algorithms are devised to grading speech quality according to the pronunciation and hearing characteristics of human beings.

The typical pronunciation scoring paradigm in [3] uses hidden Markov models (HMM) and searching algorithms to obtain phone segmentation information for input speech. HMM log-likelihood scores, segment classification scores, duration scores and timing scores are used in [1] to assess the pronunciation quality. In [2, 4], average phone segment posterior probabilities are used and a decent result is achieved. The method in [5] has combined individual machine scores which are introduced in [1-2] to predict the pronunciation quality.

In this paper, according to presence or not presence of the demonstrative pronunciation, an algorithm is defined as text-dependent algorithm or text-independent algorithm. The text-dependent algorithm is based on the acoustic similarities between input speech pronunciation and demonstrative pronunciation. In the text-independent algorithm, we use energy, pitch frequency and their variations as main features.

2. ESTABLISHMENT OF EXPERIMENT DATABASE

An experiment evaluation database is prepared for testing the performance of automatic assessment system. The database for an assessment system is different from that for automatic speech recognition system. It must include several different levels of data with perceptual discriminability. The database used here is comprised of four groups and each group includes 290 sentences which are frequently used in common life. The content of each group is the same and the difference is primarily in pronunciation quality.

The grouping for database was accomplished by acoustic experts in advance. Acoustic experts scored each sentence and put it into one of four groups according to its score. The detail of final database is shown in Table 1.

Category	Pronunciation quality
01	Excellent
02	Very good
03	Good
04	Bad

Table 1. Pronunciation qualities of 4 groups in database

Inevitably, subjectivity plays some roles in the final score given by acoustic experts. Since it is important for language learners to know whether progress is being made in pronunciation, we only asked acoustic experts to discriminate the utterances into four groups by score and the absolute value of score was not critical. We invited five experts to score all sentences for minimizing classification errors due to the inevitable subjectivity.

Then the scores are averaged and only the mean values are applied to group.

When the assessment system is tested with the database, the score of every utterance in each group will decrease as the group level increasing. The more the score result consists with the order, the better the performance of the evaluation system is. In addition, we will also compare the evaluation result between any two groups.

3. EVALUATION ALGORITHM FOR PRONUNCIATION QUALITY

Demonstrative pronunciation is usually required in traditional language learning. Beginners can not only learn language by simulating the demonstrative pronunciations, but also learn from the basic aspects of phonetic symbols and spellings. The evaluation algorithms for pronunciation quality can be designed either depending on the acoustic parameters of demonstrative pronunciations only or using the acoustic features of the input utterance without requiring presence of demonstrative pronunciations. We define the algorithm as text-dependent or text-independent algorithm by whether demonstrative pronunciations exist or not.

3.1. Text-dependent algorithm

In order to evaluate the quality of pronunciation, the acoustic parameters of input utterance must be matched to those of standard speech. The demonstrative pronunciations will be selected as standard speech in the text-dependent assessment algorithm. In other words, we score the speaker's pronunciation quality by measuring the distance of acoustic parameters between input speech and demonstrative speech.

Input and demonstrative utterance will be fed into the automatic speech recognition system (ASR). The match is on the basis of the output of ASR system. We select the posterior probability and duration of phone segment as assessment features. The method of matching is defined as:

$$Con = \frac{\sum_i \overline{St}_i \cdot \overline{Te}_i}{\sqrt{\sum_i \overline{St}_i^{\,2}} \sqrt{\sum_i \overline{Te}_i^{\,2}}} \tag{1}$$

Where i represents the total number of features, $\overline{St}$ is the feature vector of input utterance and $\overline{Te}$ is the feature vector of demonstrative pronunciation. In order to improve the accuracy of assessment, other features such as energy and pitch can be added in the feature vector.

Since the process of feature extraction from input and demonstrative pronunciations is similar, a high performance of ASR system will give agreeable assessment result in most cases. The accuracy of this assessment algorithm will be relatively high compared with the text-independent algorithm due to the existence of demonstrative pronunciations.

3.2. Text-independent algorithm

Text-independent assessment algorithm is designed in cases that demonstrative pronunciation is not easy to be found. In this circumstance, only the acoustic parameters of input speech are available. Some features such as duration of pronunciation, energy and pitch frequency are used to evaluate the quality of input utterance.

3.2.1. Segment duration

The duration of utterance is the span of utterance from the start to the end. Usually a speech decoder can output segmentation information for each word and phone. According to [1], a good result can be achieved by using phone's posterior probability and segment duration. Since the absolute rate of speech is not directly related to the quality of pronunciation, to eliminate its effect to some extent the posterior probability for each phone is normalized by its duration. The method using posterior probability and segment duration is defined as:

$$Con = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d_i}\sum_{t=t_i}^{t_i+d_i-1}\log P(q_i \mid y_t) \tag{2}$$

Where,

$$P(q_i \mid y_t) = \frac{p(y_t \mid q_i)\,p(q_i)}{\sum_j p(y_t \mid q_j)\,p(q_j)} \tag{3}$$

d_i is the duration of phone q_i. N is the number of phones included in the current word.

3.2.2. Speech energy

Energy is an important feature for speech signal in speech detection and perception. But experiments show that the absolute value of energy has no strong relations with the quality of pronunciation. In other words, the absolute value of energy can not represent the quality of pronunciation very well. The utterance energy of language learners will be fluctuant greatly and the wave shape may be not smooth. So we use the average rate of variation and mean subtracted energy as features to evaluate the quality of pronunciation. The average rate of variation is the mean of absolute value of energy difference for adjoined frames, which is defined as:

$$\rho_{en\_diff} = \frac{1}{M-1}\sum_{i=1}^{M-1}\left|En_{i+1} - En_i\right| \tag{4}$$

Where, M is the total frame number of input utterance, En_i is the energy of each frame.

The mean subtracted energy is the mean distance between energy and average energy for input frames, defined as:

$$\rho_{en\_meanSub} = \frac{1}{M}\sum_{i=1}^{M}\left|En_i - Avg_{En}\right| \quad (5)$$

Where, M is the total frame number of input utterance, En_i is the energy of each frame. Avg_{En} is the average energy of input utterance, defined as:

$$Avg_{En} = \frac{1}{M}\sum_{i=1}^{M} En_i \quad (6)$$

3.2.3. Pitch frequency of speech

The pitch frequency contour is related to rhythm of speech closely. Most beginners especially nonnative learners focus on the pronunciation of each single word while the rhythm of whole sentence is often ignored. Therefore they can not master the rhythm of pronunciation well. Pitch frequency is used as one of important features to evaluate the quality of pronunciation. We select the average rate of variation and mean subtracted pitch frequency as features.

The average rate of variation is the mean of absolute value of pitch frequency difference for adjoined frames; the mean subtracted pitch frequency is the mean distance between pitch frequency and average pitch frequency for input frames, defined as:

$$\rho_{pitch\_diff} = \frac{1}{M-1}\sum_{i=1}^{M-1}\left|Pitch_{i+1} - Pitch_i\right| \quad (7)$$

$$\rho_{pitch\_meanSub} = \frac{1}{M}\sum_{i}\left|Pitch_i - Avg_{pitch}\right| \quad (8)$$

Where M is the total frame number of input utterance, $Pitch_i$ is the pitch frequency of the ith frame utterance, Avg_{pitch} is the average value of pitch frequency.

The pitch frequency for each frame above is obtained by the sub-harmonics summation method. The frame length is 25ms and the frame shift is 10ms. The curve of pitch frequency will be below a threshold on the period of unvoiced and silent speech. We use the normalized autocorrelation method to eliminate the non-periodic data frames and regard the speech period as unvoiced or silent data when pitch frequency is below the threshold. These speech periods will not be calculated.

3.3. Comparison of two kinds of algorithms

The two kinds of algorithms introduced above have different advantages depending on whether demonstrative pronunciation is applied. Differences of the two are:

(1) Demonstrative speech corpus is needed in text-dependent algorithm and the corpus grows as new learning materials are added while text-independent algorithm does not require the corpus.
(2) In text-dependent algorithm, to achieve reasonable assessment score, some features such as the gender and age of learners need to be close to those of teacher.
(3) If conditions mentioned in (2) are satisfied, the reliability of text-dependent algorithm is usually better than that of text-independent algorithm because each utterance has a corresponding standard pronunciation to be matched with.

4. EXPERIMENT AND RESULT

We use the same database to test our text-independent algorithms for pronunciation quality. The data in the Table 2-3 below shows that the correct rate of comparing assessment result of two groups, three groups and four groups.

Firstly, we compare the assessment result of any two groups, shown in Table 2. The correct assessment rate of two compared groups means that the assessment scores order obtained from machine is consistent with the pronunciation quality order given by acoustic expert. That is to say, if the pronunciation quality of the first group is better than the second one, i.e., 01>02. It is correct only when the assessment score of sentence in the first group is larger than that of corresponding sentence in the second one.

Category	**Posterior+dur** (%)	**En+diff** (%)	**En+meanSub** (%)	**Pitch+diff** (%)	**Pitch+meanSub** (%)
01>02	28.6	32.8	48.6	17.5	24.1
01>03	46.9	33.1	50.6	44.5	62.1
01>04	52.4	36.6	51.4	71.4	81.7
02>03	65.2	31	33.1	85.5	89.7
02>04	62.8	53.4	50.3	95.5	96.7
03>04	49.3	69.7	62.1	75.2	69.3
Mean	51.8	42.7	49.3	64.9	70.6

Table 2. correct rate of comparing assessment result of any two groups

Table 2 shows that correct rates of all methods reach more than 50% except those using energy as feature. The best result was achieved when using pitch frequency as feature. The mean value of correct rate comparing any two groups is equal to 64.9% and 70.6% when average rate of variation and mean subtracted pitch frequency are used respectively. The results are all better than that of using posterior probability and segment duration methods.

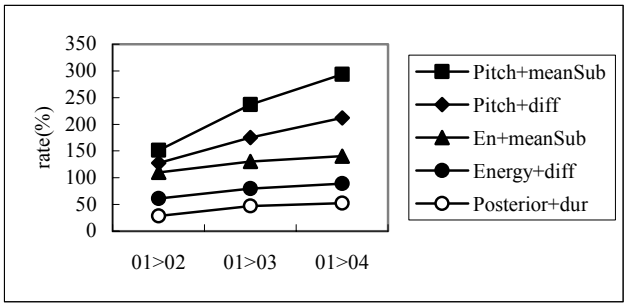

Fig. 1. Comparison of pronunciation quality between two groups using different algorithms

Secondly, we make a further analysis of assessment results with the increasing difference in pronunciation quality between two compared groups, which is depicted in Fig. 1. The correct rate of assessment results is increased with the increment of difference in pronunciation quality between two groups. The result agrees with the pronunciation quality distribution of the data set. That means the larger difference between two groups in pronunciation quality, the higher the correct rate of assessment results is.

Category	01>02	01>02>03	01>02>03>04
Posterior+ duration (%)	28.6	15.5	4.8
Energy+ diff (%)	32.8	6.9	5.9
Energy+ meanSub (%)	48.6	9.0	10.7
Pitch+ diff (%)	17.9	11.0	9.7
Pitch+ meanSub (%)	24.1	19.7	12.4

Table 3. Comparison pronunciation quality for two groups, three groups and four groups using different algorithms

Finally, we make this assessment task more difficult by calculating the correct rate of assessment results for three and four groups. The results are displayed in Table 3. The correct rate of assessment results of three groups refers to the assessment score order satisfies the pronunciation quality order. The correct rate of assessment results of four groups is obtained using the same method.

The correct rate is decreased with the increasing number of groups being compared because of the increasing of the assessment difficulty. But the result using energy and pitch frequency is still better than those using posterior probability and segment duration. After all, the best assessment result is obtained by using mean subtracted pitch frequency as feature.

5. SUMMARY

We present the text-dependent and text-independent algorithm to evaluate the quality of pronunciation. In the text-dependent algorithm, we bring forward the methods using energy and pitch frequency. The average rate of variation and mean subtracted energy and pitch frequency are selected as main features. We have designed experimental method and database and compared it with other researches in assessment of pronunciation quality. In our study, we have carefully prepared experimental data totaling 1160 sentences. At last, the assessment results using energy and pitch frequency are better than those using average phone segment posterior probabilities.

It can be found that energy and pitch frequency are good features for predicting of pronunciation quality. The energy based method can also reach a close result, but this method still needs other more effective parameters to get better result. Moreover, our future work is finding of more effective features and combining them to evaluate pronunciation quality.

6. REFERENCES

[1] H. Franco, L. Neumeyer, Y. Kim and O. Ronen, "Automatic Pronunciation Scoring For Language Instruction", *Proc. Int'l. Conf*. on Acoust, Speech and Signal Processing, pp.1471-1474, Munich, 1997.

[2] L. Neumeyer, H. Franco, M. Weintraub, and P. Price. "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech" , *Proc. of ICSLP 96*, pp.1457-1460, Philadelphia, Pennsylvania 1996.

[3] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "automatic Evaluation and Training in English Pronunciation", *ICSLP 1990*, Kobe, Japan.

[4] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality", *Speech Communication*, Volume 30, Issues 2-3, February 2000, Pages 83-93.

[5] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality", *Speech Communication*, volume 30, 2000.