Mandarin Vowel Pronunciation Quality Evaluation by Using Formant Pattern Recognition

Fuping Pan, Qingwei Zhao, Yonghong Yan

ThinkIT laboratory, Institute of Acoustics, Chinese Academy of Sciences
{fpan, qzhao, yyan}@hccl.ioa.ac.cn

Abstract

In this paper we propose to apply formant pattern recognition to Mandarin vowel pronunciation assessment. We devise a novel pitch cycle detection method and suggest estimating formant frequencies from observations of the frequency domain by using pitch-synchronous analysis. Statistically based classifiers are trained to discriminate formant patterns for vowel pronunciation assessment. Five confusable Mandarin vowels are selected for experiments. Assessment results show an average human-machine score correlation improvement of 6.10% of the new method over ASR technique, and show an average improvement of 6.37% over traditional LPC analyzing method.

Index Terms: CALL, speech recognition, formant

10.21437/Interspeech.2007-86

1. Introduction

Conventional computer assisted language learning (CALL) system uses confidence measures, computed by automatic speech recognition (ASR) technologies, for automatic pronunciation quality evaluation [1][2]. These measures have the advantage that they can be calculated in similar ways for all speech units. However, ASR confidence measures also have the disadvantage that they are not very accurate: the discrimination ability of the acoustic model among confusable phones is relatively low; all speech sounds are computed in the same way without focusing on their specific char-ataractics; and the average human-machine correlations they yield are rather low. Given the disadvantage, many other alternative approaches have been proposed. [3] used the highest ROR (Rate of Rise) value, amplitude, and duration to discriminate /x/ from /k/ in Dutch. [4] tried to delete the non-linguistic information from speech and physically-implemented a phonologic structure to represent utterance for pronunciation assessment.

In this paper we report some progress of our CALL system for Mandarin pronunciation assessment. Like many other systems, we use ASR confidence measures as the primitive approach of evaluation. As an improvement, we try to make use of formant to assess the vowels' quality, because it has been commonly agreed that formants play a dominant role in vowel identification. Since traditional formant extraction method of fix-frame LPC analysis is not accurate, the results of its application in [5] seemed not very good. To be more successful, we propose to utilize the pitch-synchronous formant analysis technique: a novel pitch cycle detection algorithm is suggested and the formant is estimated by observations from the frequency domain. The formant patterns are discriminated by using statistically based classifiers, which are trained by standard Mandarin speech. And classification probabilities are calculated as measures of the pronunciation qualities.

The rest of this paper is organized as the follows: section 2 introduces the system structure; section 3 discusses the pitch-synchronous formant analysis; the formant information is applied to pronunciation assessment in section 4; some experiments and results are given in section 5; and at last the conclusion is drawn.

2. System Overview

Our CALL system evaluates the pronunciation quality of Mandarin speech, which includes syllables, phases, and sentences. In all these cases, Mandarin syllable is the fundamental assessment unit. All Mandarin syllables can be considered as a combination of initial and final parts. Phoneme structure of Mandarin syllable can be defined as shown in Figure 1. The initial part is consonant, the final part is vowel, and the two parts are articulated together to form the syllable's pronunciation. Mandarin is a kind of tonal language. The tone is mainly specified by the pattern of pitch contour of vowel portion of the syllable.

[Consonant] | [Medial] Syllabic vowel [Ending]

Initial | Final

Figure 1: *Structure of Mandarin syllable*

A syllable is evaluated from three aspects: pronunciation quality of consonant, pronunciation quality of vowel, and the accuracy of tone. The system initially evaluates the first two parts by using the ASR technologies of Hidden Markov Model (HMM) and Viterbi search, and then uses a pre-trained Gaussian Mixture Model (GMM) tone classifier to assess the tone accuracy. At last the Viterbi force-aligned vowel portion of the syllable is separated for re-assessment by using formant. The re-assessment procedure includes pitch-synchronous formant extraction, feature calculation and GMM classification. Its results can be compared with those of ASR confidence measures. The block diagram of the system is shown in Figure 2.

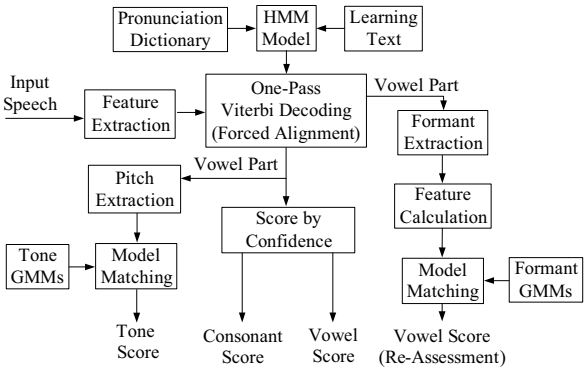

Figure 2: *System structure.*

3. Pitch-Synchronous Formant Extraction

Traditional fix-frame formant extraction method is performed at regular time intervals, typically every 25ms. It has many disadvantages. The spectral peaks may be smeared out by the convolution with the spectrum of the window function. And the assumption, that the speech during the analysis period is stationary, may cause a loss of detail in the spectral. To overcome these disadvantages, the pitch-synchronous analysis technique is investigated. The proposed pitch-synchronous formant extraction algorithm includes two parts: pitch cycle detection and formant estimation.

3.1. Pitch Cycle Detection

The pitch cycle detection method we proposed is inspired by the Empirical Mode Decomposition (EMD) algorithm of Hilbert-Huang transform (HHT) [6]. The main idea of EMD is to sift away the envelop information from the original signal gradually. We adapt this sifting process to a kind of time domain filtering process.

The first step is to divide the vowel waveform into small segments. The segments are short enough to observe little pitch change, 50 ms may be proper. Pitch cycle delimiters will be detected on each segment. For each segment, traditional pitch estimation technique is used to estimate the pitch period for future usage.

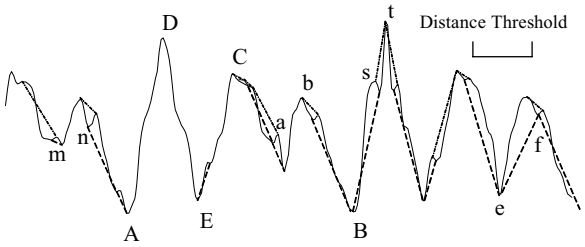

Figure 3: *Waveform filtering process.*

On each segment the filtering process is performed which is illustrated in Figure 3. From the point view of vision, we may say that a pitch cycle is normally delimited by two local minima, such as point A and point B in Figure 3, and it has a local maximum, such as point D. A and B have the smallest value, and D has the biggest value in its period. Our filtering process is to smooth the waveform to erase all the local minima and local maxima except for A, B and D in every period.

Like the EMD process of HHT, we connect waveform local minima to form its lower envelop and connect local maxima to form its upper envelop. But the connecting method is different. First of all, a small distance threshold is set. When two maxima are to be connected, two judgments should be made beforehand: (1) whether the horizontal distance between the two maxima is smaller than the preset threshold; (2) whether the local minimum between the two maxima is not the smallest among its most neighboring minima. If both judgments are satisfied, the two maxima are connected directly by beeline (for example, (s, t) in Figure 3); if not, the two maxima are connected by the waveform between them (for example in Figure 3, (D, C) violates the first judgment; (a, b) violates the second judgment). When two minima are to be connected, two similar judgments should also be made beforehand: (1) whether the horizontal distance between the two minima is smaller than the preset threshold; (2) whether the local maximum between the two minima is not the largest among its most neighboring maxima. If both judgments are satisfied, the two minima are connected directly by beeline (for example, (e, f) in Figure 3); if not, the two minima are connected by the waveform between them (for example in Figure 3, (A, E) violates the first judgment; (m, n) violates the second judgment).

After the lower and the upper envelops are formed, the mean of them is calculated and smoothed as the result. The same process is iterated on the newly produced waveform. After each iteration, the output waveform is compared with the input. If little change is made, then increase the distance threshold before go on with next iteration. The more times iteration is carried out, the simpler the waveform will be. When the distance threshold is about 0.8 times the pre-estimated pitch period, the iteration can be stopped. At this time the waveform ought turn to be a very simple oscillation with frequency of the pitch of the original waveform. It's then very easy to detect pitch cycle delimiters of the original waveform on this simplified oscillation.

3.2. Formant Estimation

Formant parameters are estimated on every pitch cycle. We assume that each pitch period is one of an infinite sequence of identical periods. Such an infinite sequence can be seen produced as show in Figure 4.

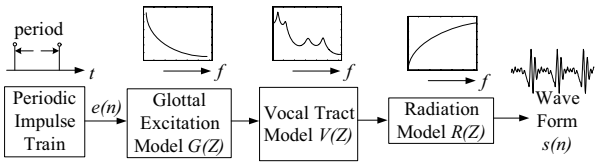

Figure 4: *Periodic speech production model.*

Of one pitch cycle, the process in Figure 4 can be written as:

$$S(Z) = G(Z) \cdot V(Z) \cdot R(Z) \tag{1}$$

As we all know, the influence of *G(Z)* can be alleviated by the pre-emphasis and *R(Z)*. Let $s'(n)$ denote the pitch cycle after pre-emphasis, we can get the following equation.

$$S'(Z) \approx V(Z) = \frac{B(Z)}{A(Z)} = \frac{\sum_{h=0}^{N} b(h) \cdot z^{-h}}{\sum_{l=0}^{M} a(l) \cdot z^{-l}} \tag{2}$$

For accuracy, some zeros are assigned to *V(Z)*. In Equation 2, we preset *N*=2, and *M*=14. *a(l)* and *b(h)* are to be estimated. Change Equation 2 to the discrete Fourier transform:

$$S'(jk\omega_0) \approx V(jk\omega_0) = \frac{B(jk\omega_0)}{A(jk\omega_0)} \tag{3}$$

Where ω_0 is the fundamental frequency of *s(n)*; $S'(jk\omega_0)$ is the spectral sample of $V(j\omega)$ at harmonic frequencies. $S'(jk\omega_0)$ can be easily gotten by DFT of $s'(n)$. We use the minimum mean square error criteria of Equation 4 to find the proper *A* and *B* [7].

$$(b, a) = \min_{b,a} \sum_{k=1}^{N} W(k) \left| S'(jk\omega_0) \cdot A(jk\omega_0) - B(jk\omega_0) \right| \tag{4}$$

Where *W(k)* is the weighting factor for controlling the error function. It can help to preserve much more detail at specified frequency areas. *W(k)* is very helpful in some cases. For example, F1 and F2 of Mandarin vowel /u/ are very close to each other, they can hardly be distinguished by traditional LPC analysis. But by tuning *W(k)* of Equation 4 at candidate

frequencies (nominal formant frequencies are known in advance in CALL), those close-neighbored formants may be separated. When *A* and *B* are available, it is easy to calculate formant frequencies.

4. Apply Formant to Vowel Pronunciation Quality Evaluation

We suggest using statistically based classifiers to distinguish formant patterns for vowel pronunciation quality assessment. Formant frequencies of sample points on the first three formant trajectories are chosen as the classification feature. Beginning from 10% and ending at 90%, each formant trajectory is sampled at equal time intervals. The formant frequencies are not normalized but only converted to the Mel scale.

Pronunciation qualities of isolated Mandarin syllables spoken by Hong Kong native speakers will be assessed. These speakers have very strong accent, they often mispronounce some Mandarin vowels as others. Some easily confusable vowels, including two mono-phthong and three two-phthong, are statistically estimated and summarized as shown in Table 1. They are not discriminated by our ASR confidence measures very well. Vowels in the left column are those should be correctly pronounced. Vowels in the right column are those may easily be confused with those in the left column.

Table 1. *Some confusable vowels of Hong Kong speakers*

Answer Vowel	Confusable Vowel
e	ai, a
u	ao, e
ou	iu
ao	iao, iu
iu	ou, u

For each vowel in Table 1, we use formant features from standard Mandarin speech to train the GMM classification model. Then for each coming test utterance *O*, a posterior probability is calculated by using Equation 5.

$$P(Ref.Vowel \mid O) = \frac{p(O \mid Ref.Vowel)}{\sum_{k=1}^{N} p(O \mid Vowel_k)} \quad (5)$$

Where *Ref. Vowel* is the answer vowel, $Vowel_k$ is a confusable vowel of *Ref. Vowel*, and there are totally *N* confusable vowels, including the answer vowel. $P(Ref.Vowel \mid O)$ is a good measure of how well the formant pattern of utterance *O* is close to that of the answer vowel, so it can be directly used for pronunciation quality evaluation.

5. Experiments and Results

5.1. Corpus

The training corpus is chosen from our isolated Mandarin syllable speech database. It is spoken by 80 native Chinese (40 female and 40 male), who have no accent. Each speaker speaks about 1200 different tonal isolated Mandarin syllables. For each vowel in Table 1, about 1000 syllable speech samples that consist of it are selected to train the GMM model. The testing corpus is chosen from the Hong Kong Putonghua-Shuiping-Kaoshi (PSK) test speech samples, which are spoken by Hong Kong native speakers. There are totally 102 test participators (51 male and 51 female). For each answer vowel in Table 1, a syllable that consists of it is spoken by all the test participators, so each answer vowel has 102 test cases.

5.2. Correlation Coefficient

The popular way to evaluate the performance of the pronunciation assessment system is to calculate the correlation between machine grades and human expert's subjective grades. The correlation coefficient (CC) of a specific vowel v is defined as:

$$CC = \frac{\left\langle \bar{H}_v, \bar{M}_v \right\rangle}{\left\| \bar{H}_v \right\| \bullet \left\| \bar{M}_v \right\|} = \frac{\sum_i h_{vi} m_{vi}}{\sqrt{\sum_i h_{vi}^2} \sqrt{\sum_i m_{vi}^2}} \quad (6)$$

Where $\bar{M}_v$ is the machine grade vector of v constituted by grades of different speakers i, and $\bar{H}_v$ is the corresponding human rating vector.

5.3. Methods

Each utterance is graded on a 0-2 scale. A rating of 2 indicates excellent pronunciation, and a rating of 0 indicates completely wrong pronunciation. The testing corpus has been graded by 5 human experts, whose average inter-rater correlation is 0.94. We use mean of the human experts' grades to calculate the human-machine grading correlation coefficients.

At first we use the ASR confidence measures to assess testing data. Correlations between the machine and human scores are shown in Table 2.

Table 2. *Correlations by ASR technique*

Answer Vowel	Correlation coefficient
e	0.804
u	0.880
ou	0.863
ao	0.868
iu	0.840

Then formants of the training data and testing data are extracted by using the pitch-synchronous analyzing method of section 3. Experiments with different number of sample points on the formant trajectories are carried out. Correlations between the machine and human scores are shown in Table 3.

Table 3. *Correlations by pitch-synchronous formant analyzing*

Answer vowel	Correlation coefficient by different number of sample points				Relatively improved according to ASR*
	3 points	4 points	5 points	6 points	
e	0.885	0.880	0.885	0.885	10.07%
u	0.923	0.923	0.923	0.923	4.89%
ou	0.885	0.868	0.863	0.851	2.55%
ao	0.928	0.912	0.907	0.896	6.91%
iu	0.846	0.868	0.885	0.891	6.07%
Aver.	0.893	0.890	0.893	0.889	6.10%

* This is calculated by using the best result among the left columns

In Table 3, the vowel formant discrimination ability responds differently to the changing of the number of the sample points. For vowels of /e/, /u/, /ou/ and /ao/, the formants are nearly steady through out the utterance. It seems 3 sample points are enough to convey the steady formant information, and more sample points may even cause some confusion. For the vowel of /iu/, the formants are dynamic, so more sample points are needed to convey the formant transient information. It's shown that 6 sample points bring the best results.

Compare Table 3 with Table 2, you can see that all the correlations have been greatly improved. This can prove the obvious predominance of formant in vowel pronunciation assessment.

At last, formants of the training data and the testing data are extracted by using the traditional LPC analyzing method. The sample points' numbers of the formant feature are selected according to the best results in Table 3. The score correlation results are shown in Table 4, and compared with those of pitch-synchronous analyzing method.

It can be seen that correlations of the LPC method are not as good as those of the pitch-synchronous analyzing method, and some are even worse than the results of ASR technique. So it is obvious that the proposed pitch-synchronous formant extraction method is more accurate than the traditional LPC method.

Table 4. *Correlations by traditional LPC analyzing*

Answer Vowel	Sample Points' Number	Correlation Coefficient		
		LPC	Pitch-Syn.	Relative Improve
e	3	0.874	0.885	1.26%
u	3	0.918	0.923	0.54%
ou	3	0.816	0.885	8.46%
ao	3	0.798	0.928	16.29%
iu	6	0.834	0.891	6.83%
Aver.	N/A	0.848	0.902	6.37%

6. Conclusion

Formant has long been regarded as dominant in vowel discrimination. But the difficulty in accurate formant extraction limits its application in speech recognition and pronunciation assessment. In this paper we propose to use the pitch-synchronous analyzing method to extract formant and apply the formant to the vowel pronunciation assessment. Our experiments show obvious predominance of formant in vowel pronunciation assessment over ASR technique and also prove that formants extracted by the new method are more accurate than results of the traditional LPC method.

7. Acknowledgement

This work is supported by Chinese 973 program (2004CB318106), National Natural Science Foundation of China (10574140, 60535030).

8. References

[1] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pro-nunciation Quality," Speech Communication, pp. 83-93, Volume 30, Issues 2-3, February 2000.

[2] S. Witt and S. Young, "Computer-assisted Pronunciation Teaching based on Automatic Speech Recognition." Proc Conf. Language Teaching and Language Technology, Univ Groningen, the Netherlands, 1997.

[3] K. Truong, A. Neri, C. Cucchiarini, H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," Proceedings of the InSTIL/ICALL Symposium, Venice, pp. 135-138, 2004.

[4] N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," Proc. Int. Conf. Spoken Language Processing, pp.1669-1672, 2004-10.

[5] K. Truong, "Automatic pronunciation error detection in Dutch as a second language: an acoustic-phonetic approach," MA Thesis, Utrecht University, The Netherlands, June 2004.

[6] N.E. Huang, S. Zheng, S.R. Long, et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear non-stationary time series analysis." [J].Proc R Soc London, A, 454: pp. 903-995, 1998.

[7] E.C. Levi, "Complex-Curve Fitting," IRE Trans. on Automatic Control, Vol. AC-4, pp. 37-44, 1959.