# MANDARIN VOWEL PRONUNCIATION QUALITY EVALUATION BY A NOVEL FORMANT CLASSIFICATION METHOD AND ITS COMBINATION WITH TRADITIONAL ALGORITHMS

*Fuping Pan, Qingwei Zhao, Yonghong Yan*

ThinkIT Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

## ABSTRACT

This paper discusses the vowel pronunciation quality assessment of our computer assisted Mandarin Chinese learning system. Under the speech recognition framework, phonetic pronunciation assessment is usually based on the phonetic posterior probability score, which may be computed by normalizing the frame-based posterior probability or be calculated on the phone segment directly. By the first method, we can achieve a human-machine scoring correlation coefficient (CC) of 0.832 for vowel; and by the second, the CC can be up to 0.847. In order to improve the performance, we suggest employing the formant feature of vowel. This paper proposes a novel method to utilize formant: we plot formant candidates of each frame on the time-frequency plane to form a bitmap, and then extract its Gabor feature for pattern classification. When we use the classification probability score for pronunciation assessment, we get a CC of 0.842. Finally we combine the three scores with various linear or nonlinear methods; the best CC of 0.913 is gotten by using neural network.

***Index Terms***— Computer Assisted Language Learning, Speech Recognition, Formant, Gabor Feature, Neural Network.

## 1. INTRODUCTION

Over the last decades many research groups have investigated on automatic pronunciation quality assessment by speech recognition techniques [1-5]. Some works were focused on the assessment at speaker level and sentence level [1-4]; some others were focused on the assessment at phone level [5]. This paper is about some improvements of Chinese Mandarin vowel pronunciation evaluation. The phonetic evaluation is traditionally based on the phonetic posterior probability score under the speech recognition framework. There are primarily two algorithms to compute it: one is to calculate the average of the logarithm of the frame based posterior probability (AFBPP) [1-3]; the other is to calculate the phone log-posterior probability (PLPP) [5]. However, these two algorithms are not very accurate in some cases due to the limited discriminating ability of the acoustic model. In order to improve the vowel assessment accuracy, we suppose to employ the long-term information of speech, which is critical to vowel perception, by utilizing formant. Considering the difficulty of accurate formant tracking, a novel kind of formant feature is suggested. That is to convert the formant candidate plots on the time-frequency plane to a bitmap and then extract its Gabor feature to represent the formant trajectory. We use Gaussian Mixture Model (GMM) to classify the formant patterns and calculate the formant classification posterior probability (FCPP) score to assess the pronunciation quality. Such the formant classification score is complementary to AFBPP and PLPP, so we further investigate to combine them with various linear or nonlinear methods, and the best result is obtained by using neural network to combine the three scores.

The rest of this paper is organized as the follows: section 2 introduces traditional phonetic evaluation method; section 3 discusses the assessment method of formant classification; section 4 is dedicated to the combination of scores; some experiments and results are given in section 5; and finally the conclusion is drawn.

## 2. TRADITIONAL PHONETIC PRONUNCIATION ASSESSMENT METHOD

The phonetic pronunciation quality is traditionally evaluated by using speech recognition techniques of hidden Markov model (HMM) and Viterbi decoding. Block diagram of the system is shown in Fig. 1. The front-end feature extraction converts the speech waveform to a sequence of mel-frequency cepstral coefficients (MFCC) and these are fed into HMM model net to do one-pass Viterbi decoding. The HMM model net only consists of the models of the learning text, and the Viterbi decoding is only a force alignment between the speech frames and the HMM models in the net. With the frame index of each HMM state and the accumulated observation probability of the phone segment, the phonetic posterior probability score is computed as the measurement of the pronunciation quality of each phone $q$ . There are mainly two algorithms to calculate it.

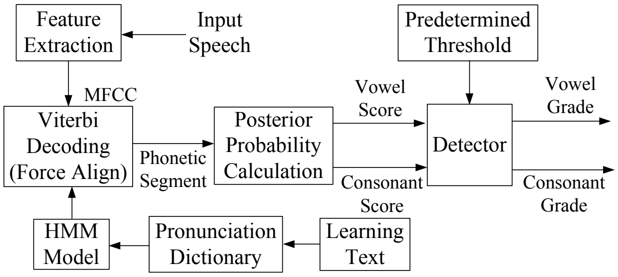

**Fig. 1.** Architecture of our pronunciation evaluation system

The first one is the average of logarithm of the frame based posterior probabilities (AFBPP) belonging to $q$ [1-3].

$$\rho = \frac{1}{t_e - t_b + 1}\sum_{t=t_b}^{t_e}\log P(s_t \mid x_t) \qquad (1)$$

Where $P(s_t \mid x_t)$ is the frame based posterior probability of the force-aligned state $s_t$ given the observation vector $x_t$ ; $t_b$ is the start frame and $t_e$ is the end frame of $q$ . The second one is phone log-posterior probability (PLPP) [5].

$$\rho = \frac{1}{\tau}\log[P(q \mid O^{(q)})] = \frac{1}{\tau}\log\frac{p(O^{(q)} \mid q)}{\sum_{p\in Q} p(O^{(q)} \mid p)} \qquad (2)$$

Where $\tau$ is the number of frames in the acoustic segment $O^{(q)}$ ; $Q$ is the set of Mandarin consonants when $q$ is consonant, otherwise is the set of Mandarin vowels when $q$ is vowel.

The final stage of evaluation uses predetermined thresholds to map the posterior probability scores to evaluation grades.

## 3. VOWEL PRONUNCIATION ASSESSMENT BY FORMANT CLASSIFICATION

Formants have long been regarded as the dominant parameters to describe vowels. Formant trajectories are supposed to convey the long-term information that is critical to vowel identification. We are to use the GMM to classify the pattern of the formant trajectory and use the classification posterior probability score for pronunciation quality assessment.

Numerous experiments have been carried out to classify a set of monophthong vowels of a specific language with formants [6-8]. It appeared that the most difficult problem is accurate automatic formant tracking. Errors tend to occur in highly transient phone boundaries [9]. We suggest a novel kind of formant feature to avoid the difficulty and can better represent dynamic properties of formant trajectory.

### 3.1. Feature Extraction

2-D Gabor function was pioneered by Daugman to model the spatial summation properties of simple cells in the visual cortex [10]. Local image features extracted by Gabor filter bank are widely used in face recognition, fingerprint identification, contour detection and many other image processing or computer vision applications. We propose to convert the formant candidate plots on the time-frequency plane to a bitmap and calculate its Gabor feature for formant pattern classification. As shown in Fig. 2, the calculation of the classification feature is described in detail as follows.

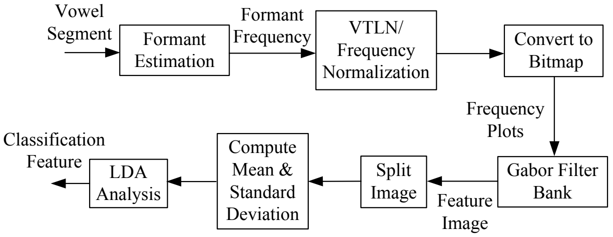

Fig. 2. Classification feature calculation process

After the vowel portion of speech is segmented by the force-alignment, its formant candidates are estimated at first. This is automatically done by using PRAAT. Differences of vocal tract length among different speakers result in an apparent expansion or compression of the frequency axis of formant trajectories, which will damage the performance of classification. So we seek vocal tract length normalization (VTLN) to compensate for the variation of formant location by a warp of frequency axis, as the second step shown in Fig. 2. With formant candidates of each frame plotted on a time-frequency plane, we only convert the plots in the region up to 4000 Hz to a bitmap, because such the region can safely contain all the first three formants. Then to be uniform, we resize the bitmaps to $200\times100$ pixels.

Typically, an input image $I(x,y),(x,y)\in\Omega$ ( $\Omega$ -the set of image points), is convolved with a 2-D Gabor function, $g(x,y),(x,y)\in\Omega$ , to obtain a Gabor feature image $r(x,y)$ as the following [11]:

$$r(x,y) = \iint_{\Omega} I(\xi,\eta)g(x-\xi,y-\eta)d\xi d\eta \qquad (3)$$

We use the following family of Gabor functions:

$$g_{\lambda,\Theta,\varphi}(x,y) = \exp(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2})\cos(2\pi\frac{x'}{\lambda}+\varphi) \qquad (4)$$

Where $x' = x\cdot\cos\Theta + y\cdot\sin\Theta,\quad y' = -x\cdot\sin\Theta + y\cdot\cos\Theta$ , $\sigma = 0.56\lambda$ and $\gamma = 0.5$ . In our experiments, the phase offset is set as $\varphi = 0$ ; the wavelength $\lambda$ and the orientations $\Theta$ are selected according to experiment results.

Suppose the filter bank has *I* wavelengths and *J* orientations, it will generate $I\times J$ Gabor feature images. We equidistantly split each image into *M* rows and *N* columns to produce $M\times N$ blocks, and calculate the mean ${\mu^{(l)}}_{mn}$ and standard deviation ${\sigma^{(l)}}_{mn}$ of the magnitude of each block as feature elements:

$$\begin{aligned} {\mu^{(l)}}_{mn} &= \iint \left|{r^{(l)}}_{mn}(x,y)\right| dxdy \\ {\sigma^{(l)}}_{mn} &= \sqrt{\iint (\left|{r^{(l)}}_{mn}(x,y)\right| - {\mu^{(l)}}_{mn})^2 dxdy} \end{aligned} \qquad (5)$$

According to [6-8], inclusion of vowel duration $D$ will result in a consistent classification improvement. So the feature vector $\bar{f}$ is constructed as:

$$\bar{f} = [{\mu_{11}}^{(1)}, {\sigma_{11}}^{(1)} \ldots {\mu_{mn}}^{(l)}, {\sigma_{mn}}^{(l)} \ldots {\mu_{MN}}^{(I\times J)}, {\sigma_{MN}}^{(I\times J)}, D] \qquad (6)$$

Where *l* is the index of Gabor feature image, and *m*, *n* are the block indexes.

$\bar{f}$ usually has a very high dimension. However, a low-dimensional representation of the vector is especially important for machine learning. We use linear discriminant analysis (LDA) to project the high dimensional feature to a lower dimensional space, and exploit the projected vector as the final classification feature.

### 3.2. Classification Posterior Probability Score

We use GMM to classify the Gabor feature for vowel pronunciation evaluation. One model is trained for each Mandarin vowel. For each coming testing feature $x$ , a formant classification posterior probability (FCPP) score is calculated as:

$$P(Vowel_{ref} \mid x) = \frac{p(x \mid Vowel_{ref})}{\sum_{k\in V} p(x \mid Vowel_k)} \qquad (7)$$

Where $Vowel_{ref}$ is model of the answer vowel; *V* is the Mandarin Chinese vowel set. $P(Vowel_{ref} \mid x)$ is mapped to evaluation grades by using predetermined thresholds.

## 4. SCORE COMBINATION

Scores of AFBPP, PLPP and FCPP contain information of different time scale and so are complementary to each other. We investigate to combine them to improve the assessment performance by using various linear or nonlinear methods. This is a problem of predicting the human subjective evaluations by the machine scores.

At first, we suggest the expected value of human grade $\tilde{h}$ is a linear combination of two or more machine scores $m_n$ plus a bias term $b$.

$$\tilde{h} = a_1m_1 + a_2m_2 + ... + a_nm_n + b \qquad (8)$$

The linear coefficients $a_1, a_2, ..., a_n$ and $b$ are estimated by minimizing the mean square error between the predicted and the actual human grades.

Probability distribution estimation is chosen as the second combination method. In this approach we compute the expected human grades by using estimates of the necessary conditional probabilities. The predicted human grade $\tilde{h}$ is computed as:

$$\tilde{h} = \arg\max_{h_i}[P(h_i \mid m_1, m_2, ...m_n)] \qquad (9)$$

Where $P(h_i \mid m_1, m_2, ...m_n)$ is the estimated conditional probability of the human grade $h_i$ given the machine scores $[m_1, m_2, ...m_n]$. Suppose $P(h_i) = P(h_j)\ (i \neq j)$, by using Bayes rule, the predicted human grade $\tilde{h}$ can be deduced as:

$$\tilde{h} = \arg\max_{h_i}[P(m_1, m_2, ...m_n \mid h_i)] \qquad (10)$$

In this work we model $P(m_1, m_2, ...m_n \mid h_i)$ by using Gaussian mixture model.

And the last combination method is neural network. A neural network can be capable of implementing arbitrary maps between input and output spaces. With this approach, the machine scores to be combined are the input to a neural network; the predicted grades are the output values of the network; the actual human grades provide the targets for the training of the network. Neural network parameters, the weights, are adjusted by the training algorithm to minimize the error criterion. After some preliminary experiments with different network architectures, we choose the two-layer back propagation network with a single linear output unit and a hidden layer of log-sigmoid units. We vary the number of hidden layer units; the best performance is obtained with 10 hidden units. The number of input units corresponds to the number of machine scores combined. The network is trained by using the mean square error criterion. A momentum term is used in the weight update rule to accelerate the training speed. To avoid over fitting to the training data and to obtain good generalization, we use a cross-validation set formed with 15% of the training data. Prediction performance is assessed after each training iteration on this set; the training is stopped when performance do not improve on the cross-validation set [3].

## 5. EXPERIMENTS AND RESULTS

### 5.1. Corpus

The following experiments are performed on the Hong Kong Putonghua-Shuiping-Kaoshi (PSK) pronunciation test samples. A PSK test set has 75 utterances, including 50 mono-syllable words and 25 double-syllable words. We only focus on vowel assessment in this paper. Each vowel in the test is graded on a 0-2 scale. A rating of 2 indicates excellent pronunciation, and a rating of 0 indicates completely wrong pronunciation. We collect 195 sets of test samples of the same content from 195 test attendees, among whom half are male and half are female. 80% of the collected samples are used as training set and the other 20% are used as testing set.

We use speech from a native Mandarin mono-syllable database to train the GMM model for formant classification. About 5000 utterances are collected for every Mandarin vowel. They are averagely spoken by 286 native Chinese speakers, among whom half are male and half are female.

### 5.2. Methods

The popular way to evaluate the performance of a pronunciation assessment system is to calculate the correlation coefficient (CC) between machine grades and human expert's grades. Both the testing and the training corpora have been graded by 5 human experts, whose average inter-rater CC is 0.94. We use mean of the human experts' grades to calculate the human-machine grading CC.

At first we use AFBPP and PLPP to grade the vowel pronunciation quality. Two thresholds are trained on the training data set to map those scores to grades of 0, 1 and 2, and then applied to the testing data set. The vowels' average CC is shown in Table 1. It looks like that the PLPP scores show better correlation with human grades.

**Table 1.** CC of traditional phonetic posterior probability score

| Machine score | Average CC |
|---|---|
| AFBPP (Baseline) | 0.832 |
| PLPP | 0.847 |

Then the same experiment procedure is followed to exam the FCPP score. We compare performances of four kinds of formant features. The first one is Gabor feature. Extensive experiments are done to determine the optimal Gabor feature parameters. We set $\lambda = 10$ and $\Theta = [0^\circ, 45^\circ, 90^\circ, 135^\circ]$ to generate the Gabor filter bank according to Equation (4) and equidistantly split each Gabor feature image into 10 rows and 4 columns, that is 40 blocks. Mean and standard deviation of each block together with the vowel segment duration form the feature $\overline{f}$, whose dimension is $1 \times 4 \times 10 \times 4 \times 2 + 1 = 321$. By LDA analysis, we reduce the vector size to 50. In order to demonstrate the profit of Gabor transformation, we compose the second kind of feature with means and standard deviations of the original untransformed image blocks (Non-Gabor M&SD). The original image is also split into 10 rows and 4 columns that lead to a vector size of 81 (including vowel duration). After LDA analysis, the dimension is reduced to 50, too. The third kind of feature is constituted by the vowel segment duration and direct formant measurements, which are sampled at 20%, 50% and 80% of the vowel continuance [7].

**Table 3.** CC of different mapping methods and combinations of machine scores

| Combination method | Machine scores | Average CC | Relative CC improve to baseline |
|---|---|---|---|
| Linear regression | AFBPP+PLPP | 0.852 | 2.4% |
| Linear regression | AFBPP+FCPP | 0.862 | 3.6% |
| Linear regression | PLPP+FCPP | 0.860 | 3.4% |
| Linear regression | AFBPP+PLPP+FCPP | 0.864 | 3.8% |
| Distribution estimation | AFBPP+PLPP | 0.859 | 3.2% |
| Distribution estimation | AFBPP+FCPP | 0.876 | 5.3% |
| Distribution estimation | PLPP+FCPP | 0.876 | 5.3% |
| Distribution estimation | AFBPP+PLPP+FCPP | 0.879 | 5.6% |
| Neural network | AFBPP+PLPP | 0.885 | 6.4% |
| Neural network | AFBPP+FCPP | 0.910 | 9.4% |
| Neural network | PLPP+FCPP | 0.909 | 9.3% |
| Neural network | AFBPP+PLPP+FCPP | 0.913 | 9.7% |

And the last one utilizes third-order Legendre polynomials of the formant trajectories [8] together with duration of the vowel segment. Correlation coefficients of the four features are compared in Table 2. It can be seen that the Gabor feature results in the best CC of 0.842, which is better than that of AFBPP score.

**Table 2.** CC of formant classification posterior probability score

| Formant feature | Feature dimension | Average CC | Relative CC improvement to baseline |
|---|---|---|---|
| Gabor feature | 50 | 0.842 | 1.2% |
| Non-Gabor M&SD | 50 | 0.818 | -1.7% |
| Formant samples | 10 | 0.819 | -1.6% |
| Legendre polynomials | 13 | 0.821 | -1.3% |

At last we evaluate the three different types of predictors, including linear regression, probability distribution estimation, and neural network, in mapping and combining different types of machine scores to increase the correlations. The parameters of the regression and estimation models are trained on the training set and evaluated on the testing set.

The evaluation results are shown in Table 3. It indicates that all combinations of scores can lead to an increase of the average CC. The inclusion of FCPP score is especially useful because it contains long-term information which is critical to vowel identification. The nonlinear combination methods are better than the linear one. The best case is combining the three scores with neural network, which increase the correlation by 9.7% with respect to baseline.

## 6. CONCLUSION

Accurate formant tracking is a very difficult problem, which limits its application in speech recognition and pronunciation assessment. This paper bypasses the difficulty by using Gabor feature to represent the formant trajectory and gets good results. The formant trajectory contains long-term information of speech, which is critical to vowel identification. After the formant classification score is combined with the traditional two kinds of phonetic posterior probability scores, the correlation between machine and human grades is greatly improved. It should be noticed that the mapping from machine scores to evaluation grades is more likely to be a nonlinear function than a linear one. The combination method of neural network gets the best CC of 0.913, which is very close to the CC of inter-human rating of 0.94.

## REFERENCES

[1] H. Franco, L. Neumeyer, etc, “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, etc, “Automatic Scoring of Pronunciation Quality”, Speech Communication, Vol. 30, Issues 2-3, February 2000, pp. 83-93.

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

[4] K. Tatsuya, D. Masatake, T. Yasushi, “Practical use of English pronunciation system for Japanese students in the CALL classroom”, INTERSPEECH-2004, pp. 1689-1692, 2004.

[5] SM WITT, SJ YOUNG, “Phone-level pronunciation scoring and assessment for interactive language learning,” Speech communication, 30:2-32-3, pp. 95-108, Elsevier, 2000.

[6] Nearey, T.M, Assmann, P.F., “Modeling the role of inherent spectral change in vowel identification", Jorunal of the Acoustical Society of America, Vol. 80, pp. 1297-1308, 1986.

[7] Hillenbrand, J.M., Getty, etc, “Acoustic characteristics of American English vowels”, Journal of the Acoustical Society of America, Vol. 97, pp. 3099-3111, 1995.

[8] P. Schmid and E. Barnard, “Explicit, n-best formant features for vowel classification,” in Proceedings of ICASSP 97,vol. 2, pp. 21-24, 1997.

[9] Lee, M. VanSanten, J. Mobius, B. Olive, J., "Formant Tracking Using Context-Dependent Phonemic Information," IEEE Transactions on Speech and Audio Processing, Volume 13, Issue 5, Part 2, pp. 741- 750, 2005.

[10] N. Petkov, “Biologically motivated computationally intensive approaches to image pattern recognition,” Future Generation Computer Systems, Vol. 11 (4-5), pp. 451-465, 1995.

[11] Grigorescu, S.E., Petkov, N., Kruizinga, P., "Comparison of texture features based on Gabor filters," IEEE Transactions on Image Processing, Volume 11, Issue 10, pp. 1160- 1167, Oct 2002.