OBJECTIVE EVALUATION OF THE NEUTRAL TONE IN PUTONGHUA PROFICIENCY TESTING

Tang Lin[1, 2] **Yin Jun-xun**[1]

(1. College of Electronics & Information Engineering, South China Univ. of Tech., Guangzhou 510640, Guangdong, China)

(2. Department of Information Technology, Jiangmen Polytechnic, Jiangmen 529000, Guangdong, China)

china-tl@163.com

Abstract

Correct pronunciation of the neutral tone (T0) is an important factor in assessing Putonghua (PTH) proficiency levels. The T0 objective evaluation system is an important subsystem of the Putonghua Proficiency Test (PPT) objective evaluation system. After analyzing the T0 features, this paper presents a pitch normalizing algorithm which reflects auditory sensitivity. Seven features are selected as evaluation parameters. Using the Gaussian Mixture Models (GMM) or the Multi Layer Perceptron (MLP) Models, data drawn from 43 individuals' utterances is tested. The accuracy of the objective evaluations described above is 89% identical to that usually achieved in subjective evaluation.

1. INTRODUCTION

Chinese Putonghua (PTH, Mandarin or Standard Chinese) is a well-known tonal language in which pitch tones play important phonemic roles. Each syllable in Chinese is associated with a pitch tone. Normally there are 4 lexical tones in PTH, as well as many morphophonemic tonal sandhis. The T0, one of the morphophonemic tonal sandhis, plays an important role in determining one's fluency in PTH. So, the T0 objective evaluation system is an important subsystem of computer-aided PPT systems.

People who are engaged in the study of computer-aided PPT systems use the core of speech recognition as an evaluation algorithm of the whole word [1][2]. But the PPT must also test accuracy in pronunciation, in an analysis of such details as the initial, final, and lexical tone, as well as tone modification (including the T0). The goal of this study is to develop a system that can test the accuracy of the T0 pronunciation.

The T0 is not assigned to specific single words in order to distinguish them from their homonyms as other tones are, but is assigned, rather, to single characters (often the last word) in a phrase. Up to now, no standardized vocabulary has been assembled of words that are made up of characters that have the T0, and the number of phrases in which the T0 appears differs from dictionary to dictionary. We use a vocabulary of 548 phrases that are listed in [3], which should be mastered in the PPT.

In the subjective evaluation process there is no set criteria for assessing whether a character has been pronounced with a T0, and the evaluation rule is usually an arbitrary feeling of whether it has been pronounced with a light and short touch. The main challenge before us is how to introduce a more objective, precise and fair set of evaluation rules. This paper proposes a pitch normalizing algorithm, which is based on the speaker's musical scale proportion. The GMMs or the MLP Models serve as statistic models of evaluation rules. The results of our experiments demonstrate the feasibility of using a pitch normalizing algorithm based on the speaker's musical scale proportion and the MLP models for T0 evaluation.

The rest of this paper is organized as follows: Section 2 discusses the T0 characteristics; Section 3 identifies the evaluation rules; Section 4 introduces the GMMs and MLP Models as evaluation models; Section 5 sets out the results of the experiments, and our conclusions are presented Section 6.

2. T0 FEATURES

Fundamental frequency (F0) contours are the main acoustic manifestations of pitch tones, and they are also the distinguishing feature of the T0. The T0 is not characterized by sound intensity, but by sound length, F0 and F0 patterns [4]. The length of the T0 syllable is usually half of its preceding syllable. The starting F0 of the T0 is highly

dependent on the preceding tone. The T0, after a high-level tone, is sounded at the half low level (degree 2), and it is sounded at the middle level (degree 3) after a high-rising tone. After a low-dipping tone, the T0 is sounded at the half high level (degree 4), while after high-falling tone it is sounded at low level (degree 1). However, when it comes to actual pronunciation, this is not always the case. The F0 patterns of the T0 are variable depending on both the preceding tone and its following tone. Generally speaking, after the non-low-dipping tone, the T0 is pronounced with a low falling pitch. When T0 follows a high-falling tone, its starting F0 is very low; therefore its falling range is the least. The pattern of T0 after a low-dipping tone is a middle flat model or a lightly rising model.

Hence, the T0 features can be extracted from the F0 contour. The features are:

- F0 at the end frame of the preceding syllable.
- F0 contour of the T0 syllable.
- F0 at the beginning frame of the T0 (voiced segment).
- F0 at the end frame of the T0 (voiced segment).
- The ratio of the length of the preceding syllable and that of the T0 syllable.

The T0 subjective evaluation rules usually use the values and patterns marked with a five-point scale, but the scale is a relative value. The F0 is an absolute value, and varies even though the same word is pronounced twice by the same person. It is, therefore, necessary to transform the above T0 features into relative features, as described in 4.1.

3. Evaluation Rules

The goal of this system is to produce objective evaluation results of the T0 in the PPT. Three marks are given for evaluation results: correct, wrong and faulty.

Statistics show that subjective evaluation performed by humans largely involves cases where the T0s are pronounced as their original tones, and the remaining errors are either wrong initials or wrong finals, which are addressed by other objective evaluation systems. All the errors described above are marked "Wrong". Faulty pronunciation of the T0 is mainly caused by an incorrect starting F0, although some are caused by dialect influences. Another aspect of faulty pronunciation is when the pronunciations still have the weakened original tone patterns. These are marked "Faulty"

Taking these features of the T0 into consideration, the evaluation rules are given as below:

Assume: **X** is the feature vector of the T0 syllable. λ is a T0 model. μ is the original tone model of the T0. $P(\mathbf{X} \mid \lambda)$ represents the probability that X belongs λ, and it will be discussed in section 4.The T0 evaluation rules can be described as:

If $P(\mathbf{X} \mid \lambda) > \alpha_1$, it is believed that this syllable is pronounced correctly.

Otherwise, if $P(\mathbf{X} \mid \mu) > \alpha_2$, it is marked "Wrong".

Otherwise, it is marked "Faulty".

Where α_1 and α_2 are thresholds which are determined by experiments.

4. EVALUATION METHODS

4.1. F0 normalization

(1) Speaker normalization

The dynamic pitch ranges of different speakers differ greatly; spanning a range of between 80Hz to 270Hz. Therefore the speaker independent objective evaluation system must normalize F0. This paper presents a method of speaker normalization which is based on a musical scale proportion of the F0.

Calculate the musical scale value of the F0:

$$P_i = 12\log_2(F_i / 320) + 50$$

Normalize the musical scale value by proportion:

$$P_i^{'} = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}} \times 5 + 1$$

where, P_{min}, P_{max} stands for the minimum and maximum musical scale value in one's whole pronunciation, and P_i stands for the ith frame's musical scale value.

(2) Normalization of the length

An individual's speed of speaking can be influenced from time to time by emotional, stylistic and environmental factors. As a result, the length may vary greatly even when pronouncing the same word. Therefore, a normalization of length is also needed.

In this system, the F0 contour of the T0 syllable is normalized into five points. Among those, the normalized values at the starting point and the end point are not changed:

$$P_0^{''} = P_B^{'} \qquad P_4^{''} = P_E^{'}$$

Where B and E are the beginning frame number and the end frame number of the voiced segment of the T0 syllable.

The normalized values of the remaining three points can be determined by the following rules:

- If a maximum or minimum value exists at j frame within the period:

  $$B + \frac{E-B}{4} \times (i-1) \quad to \quad B + \frac{E-B}{4} \times (i+1)$$

  and j fulfills the equation:

  $$| j - (B + \frac{E-B}{4} \times i) | \leq \frac{E-B}{8}$$

  then, $P_i^{''}$ is equal to the maximum value or to the minimum value.

- Else $P_i^{''} = (P^{'}_{B+\frac{E-B}{4}\times i-1} + P^{'}_{B+\frac{E-B}{4}\times i+1}) / 2$

Then the T0 feature vector consists of seven features: the normalized F0 at the end frame of preceding syllable, the ratio of the length of the preceding syllable and that of

the T0 syllable, and the normalized five points of the T0 contour.

4.2. T0 Models

The length of the T0 syllable is short, but its pattern is different when the tone of the preceding syllable is different. According to the rules of the tone sandhi, low-dipping continues with low-dipping, then the preceding tone changes to a high-rising tone. The T0 patterns can be divided into 15 classes. This system builds 15 models.

(1) GMM [5]

The probability density function of GMM is the sum of the M weighted Gaussian probability density function, and its equation is as below:

$$F(\mathbf{X}/\lambda)=\sum_{i=1}^{M}C_i N_i(\mathbf{X})$$

where, $F(\mathbf{X}/\lambda)$ is the probability density function of feature vector **X** under the T0 model λ . C_i is the Gaussian mixture weight, and fulfils: $\sum_{i=1}^{M}C_i=1$. N_i（**X**）, the ith Gaussian probability density function, is defined as:

$$N_i(\mathbf{X})=\frac{1}{(2\pi)^{\frac{d}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\mathbf{X}-\mu_i)'\Sigma_i(\mathbf{X}-\mu_i)\}$$

GMM is determined by three parameters; the centroid μ_i, the covariance matrix Σ_i and the prior (mixing fraction in the GMM) C_i. Take these parameters together: $\lambda=\{C_i, \mu_i, \Sigma_i\}, i=1, 2, \ldots, M.$ The mixture number M is determined by experiments.

Introduce a discrete stochastic variable q, its value range is $1, 2, \ldots, M$, and $P(q=i)=C_i$. Meanwhile, when q is a valid value, **X** is a Gaussian distribution:

$$F(\mathrm{X}\mid q=i)=N(\mathrm{X}\mid\mu_i,\Sigma_i)=N_i(\mathrm{X})$$

Then the margin probability distribution can be given:

$$F(\mathbf{X})=\sum_{i=1}^{M}P(q=i)F(\mathbf{X}\mid q=i)=\sum_{i=1}^{M}C_i N_i(\mathbf{X})$$

Using the expectation-maximization (EM) algorithm [6], the parameters of GMM can be obtained.

EM Algorithm:

- Initializing C_i, μ_i and Σ_i;

（In this study, the Euclidian distance and K-means clustering algorithm are employed to cluster to M classes. Then the centroid and the covariance of every class can be calculated. Meanwhile, the ratio of the number in one class divided by the total number of data is set as the Gaussian mixture weight.）

- E-Step:

Calculate the posterior probability of q by using the feature vector p^n of n'th training sample. The posterior probability of the hidden variable q is written:

$$C_{ni}=F(q=i\mid p^n)=C_i N(p^n\mid\mu_i,\Sigma_i)/[\sum_{i=1}^{M}C_i N(p^n\mid\mu_i,\Sigma_i)]$$

- M-Step:

Calculate：

$$C_i=\frac{1}{N}\sum_{n=1}^{N}C_{ni}\quad,\qquad\mu_i=\sum_{n=1}^{N}C_{ni}p^n/\sum_{n=1}^{N}C_{ni}$$

$$\Sigma_i=[\sum_{n=1}^{N}C_{ni}(p^n-\mu_i)(p^n-\mu_i)^T]/\sum_{n=1}^{N}C_{ni}$$

N is the total number of the training data.

- Repeat E-step and M-step until fulfilling the condition.

Using EM algorithm, the parameters of GMM are properly estimated. In fact, a GMM is a single state Hidden Markov Model.

(2) MLP Model

Artificial neural networks (ANN) have been applied successfully in speech processing systems. The MLP model, based on the back propagation algorithm in ANN, is simple and useful and widely applied in various applications.

This system uses 3-layers MLP models. The input layer consists of seven nodes, each representing a component of the extracted feature vector. To represent the three evaluation results: correct, wrong and faulty, the output layer is composed of three nodes. The size of the hidden layer is task-dependent and is determined empirically.

5. EXPERIMENTAL RESULTS

5.1. Preparations

The data corpus consists of 8749 utterances from 90 speakers(45 males and 45 females). Two individuals have an A+ level of the PPT, and other two have an A- level(they all come from north China). The four individuals read three times the 548 phrases in the vocabulary of [3] and 150 stress phrases corresponding to the T0 syllables' original tone in the vocabulary for objective tone evaluation, and the resulting data is used for training. Data from the remaining 86 individuals was collected in the locale of PPT, and we chose the T0 phrases among them for our use. The test results were marked on the locale (only marking the wrong or faulty syllables), and verified after the test. All the participants were college students who have grown up in the Cantonese dialect area except seven of them who have grown up in north China. Statistics according to gender are set out below in Table 1, while a breakdown by T0 Syllables and Stress Phrases is shown in Table 2.

Table 1. Data Corpus by Gender

Speaker's PTH level	A+	A-	B+	B-	C+	C-	Total
Male	1	1	3	8	20	12	45
Female	1	1	4	12	20	7	45

Table 2. Data Corpus: T0 Syllables/Stress Phrases

Speaker's PTH level	A+	A-	B+	B-	C+	C-	Total
T0 Syllables	2*3*548	2*3*548	29	86	176	82	6949
Stress Phrases	2*3*150	2*3*150					1800
Wrong by S.E	0	0	2	4	32	14	52
Faulty by S.E.	0	0	5	14	56	68	143

(S.E. — Subjective Evaluation)

5.2. Experiment Results

All data obtained from the A level participants was used to train the models. The data of the other 86 persons is divided into two groups: Group 1 was also used to train the models, and the Group 2 was used to test the objective evaluation system. Statistics according to gender and T0 Syllables of the two groups are shown in Table 3 and Table 4.

Table 3.Group 1 for Training: by Gender and T0 Syllables

Speaker's PTH level	B+	B-	C+	C-	Total
Male	2	4	9	6	21
Female	2	6	11	3	22
T0 syllables	17	43	88	40	188
Wrong syllables by S.E.	1	2	16	7	26
Faulty syllables by S.E.	3	7	28	33	71

Table 4. Group 2 for Testing: by Gender and T0 Syllables

Speaker's PTH level	B+	B-	C+	C-	Total
Male	1	4	11	6	22
Female	2	6	9	4	21
T0 syllables	12	43	88	42	185
Wrong syllables by S.E.	1	2	16	7	26
Faulty syllables by S.E.	2	7	28	35	72

When training GMMs, 6576 correct T0 syllables are used to estimate the parameters λ_i, i=1,2,…15, corresponding to the 15 T0 models. In addition, 1800 syllables corresponding to the T0 syllables' original tone in stress phrases and 26 wrong syllables shown in Table 3 are used to estimate the parameters μ_i, i=1,2,…15, corresponding to the tone models which are the T0 syllables' original tone.

All data in Table 3 is used to determine the 30 thresholds, α_{1i}, α_{2i}, i=1,2,…15. Some combinations may not have training data. In this case, the threshold sets 0.6.

Training MLP models is relatively simple. All the training data was used to train 15 models.

The results of the experiment are shown in Table 5. From the table, it is obvious that MLP models perform better than GMM models. The MPL models have stronger nonlinear distinctive ability and have used the information among the classes. In Table 5 below, we see that amongst the GMM models, more wrong syllables were mistakenly assessed as correct syllables than vice versa. This is due to the fact that there are significantly more correct syllables than wrong syllables in the training data, so that the classifiers are in favor of the correct classes.

Table 5. Confusion Matrix

Models		GMM Models			MLP Models		
S.E. Results		C	W	F	C	W	F
O.E. Results	C	82	1	8	83	0	7
	W	1	19	5	0	21	4
	F	4	6	59	4	5	61
Coincidental Rate		86.49%			89.19%		

(S.E.— Subjective Evaluation O.E.— Objective Evaluation)
(C — Correct W — Wrong F — Faulty)

The fact that there is no such confusion with the MLP models is good evidence that the system proposed in this paper is significantly more effective in distinguishing between the correct and incorrect pronunciation of T0 syllables.

6. CONCLUSION

In this paper we present a musical scale speaker normalization method and use MLP models as assessment functions of the T0 objective evaluation system. The results show that the coincidental rate of the objective evaluation and the subjective evaluation has reached 89.19%. It is, however, true that the data corpus consists mainly of the utterances from Cantonese speakers. Future work that can extend this study is to collect more data from a wide range of different dialect areas.

7. REFERENCES

[1] GuoQiao, LuJilia, "The Evaluation System for the Learner's Pronunciation in a Computer Aided Chinese Teaching Program", *China Journal of Chinese Information Processing*, vol. 13(3), pp. 48-53, 1998 (in Chinese)

[2] YUE Dong-jian, CHAI Pei-qi, "Features of Chinese Computer-Aided Language Learning System", *China Mini-Micro System*, vol. 22(7), pp.848-850, 2001 (in Chinese)

[3] The Putonghua training and testing center of the national language committee, *Implementation outline for Putonghua Proficiency test*, The commercial Press, Beijing China, pp. 238-246,2005: (in Chinese)

[4] WANG Yunjia, "The effects of pitch and duration on perception of the neutral tone in standard Chinese" ,*China Acta Acustica*, vol. 29(5), pp. 453-461, 2004 (in Chinese)

[5] Roberts S. J., Husmeier D., Rezek I., et al, "Bayesian Approaches to Gaussian Mixture Modeling", *IEEE Trans. on Pattern Analysis and Machines Intelligence*, vol. 20 (11) pp. 1133-1142,1998,

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood From Incomplete Data via the EM Algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1-38, 1977.