interspeech 2008 incorporating SST 08

Forward Optimal Modeling of Acoustic Confusions in Mandarin CALL System

Fengpei Ge, Fuping Pan, Changliang Liu, Bin Dong, Yonghong Yan

ThinkIT laboratory, Institute of Acoustics, Chinese Academy of Sciences
Beijing 100190, P. R. China
{gefengpei, panfuping, liuchangliang,dongbin,yyan}@hccl.ioa.ac.cn

10.21437/Interspeech.2008-478

Abstract

Acoustic confusions degrade the accuracy of pronunciation assessment severely in Computer Assisted Language Learning (CALL) systems. This paper presents our recent study on optimal modeling of the acoustic confusions. We change the traditional mandarin syllable structure, which is composed of initial and final, to a novel phoneme structure. Several phoneme splitting strategies are investigated, and the question list used for building and merging decision tree is studied. The questions are special to each phoneme splitting strategy. Experiments show that the optimal phoneme splitting strategy outperforms the traditional initial-final structure in our CALL system, with relative 11.05% ASER improvement for nasal finals. This idea may be extended to improve the performance of automatic speech recognition (ASR).

Index Terms: CALL, Phoneme Splitting, Acoustic Confusion, Pronunciation Quality Assessment, Speech Recognition

1. Introduction

CALL systems using ASR technology have received increasing attention in recent years [1][2]. Many research efforts have been done for improvement of such systems, especially for second language learning [3][4]. Generally, they separate the utterance into segments firstly, and then assess the segmental pronunciation quality. The classical measurement of the segmental pronunciation quality is the phonetic posterior probability. Many previous works mainly investigate the 'Goodness of Pronunciation' (GOP) measurement [5][6][7][8][9].Witt and Young studied pronunciation quality assessment at the phone level, and suggested improving the evaluation performance by using explicit error modeling[10]. So far studies on CALL systems using ASR mainly concentrate on researching the measurement of segmental pronunciation qualities[11][12][13]. In fact, the confusion about acoustic model (AM) is also a critical problem in CALL systems, which may degrade the performance of pronunciation quality assessment. But few papers have focused on modifying AM about acoustic confusions in CALL systems.

Mandarin is a syllable language, which has two parts: the initial is a consonant and the final is a vowel. According to this strict initial-final structure, training finals needs to be paid more attention to. For example, if we use the same state number and GMM number as that for 'a' to model 'ueng', the reliability of AM is doubtable. Actually, the acoustic confusion caused by the inaccurate AM is popular in mandarin pronunciation quality assessment, especially for strong accented speech.

This paper presents our recent study in optimal modeling of acoustic confusions in mandarin CALL systems. We change mandarin syllables from the traditional initial-final structure to a novel phoneme structure. First of all, we analyze the acoustic confusions in terms of statistics and knowledge. Then, several phoneme splitting strategies are investigated. Meanwhile, for training AM, the question list used for building and merging decision trees is studied. For each phoneme splitting strategy, the question list is special. Lots of experiments show that the optimal phoneme splitting strategy outperforms the traditional initial-final structure in CALL systems. Additionally, the novel phoneme splitting idea may be applied to automatic speech recognition (ASR) in future.

The paper is organized as follows. In section 2, baseline system is introduced. Section 3 is dedicated to phoneme splitting strategies and question design. The experiment result is shown in section 4. Finally, the conclusion is drawn.

2. Baseline System

The phonetic pronunciation quality is traditionally evaluated by using speech recognition techniques based on hidden Markov model (HMM) and Viterbi decoding. A 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), which are fed into one-pass Viterbi decoder. 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, with two algorithms.

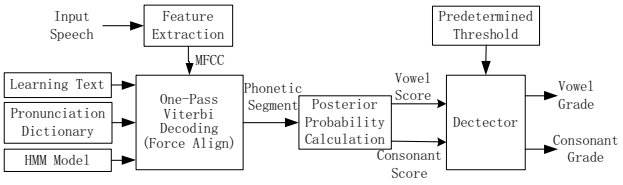

Figure 1: Architecture of our pronunciation evaluation system

The first one is the average of logarithm of the frame based posterior probabilities (AFBPP)[7][8][9].

$$\rho_{AFBPP}(PH|O) = \frac{1}{e-b+1}\sum_{t=b}^{e}\log P(s_t|o_t) \qquad (1)$$

where $O = [o_b,\ o_{b+1},\ \ldots\ o_e]$ is the force-aligned observation sequence of the phone PH (which is consonant or vowel), b is

Accepted after peer review of full paper

the begin frame of PH and e is the end frame of PH. $P(s_t|o_t)$ is the frame posterior probability.

The other is phone log-posterior probability (PLPP) [10], which is calculated as equation 2.

$$\rho_{PLPP}(PH|O) = \frac{1}{\tau}\log P(q|O^{(q)}) = \frac{1}{\tau}\log\frac{p(O^{(q)})|q}{\sum_{p\in Q} p(O^{(q)}|p)} \quad (2)$$

where τ is the number of frames in the acoustic segment $O^{(q)}$; Q is the set of Mandarin consonants when q is a consonant, and 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. Phoneme Splitting Strategies and Question Design

3.1. Acoustic Confusion analysis

Acoustic confusions caused by inaccurate AM are very common in mandarin pronunciation quality assessment, especially for strong accented speech. Based on some statistics about the pronunciation quality assessment, it is observed that long finals are performed poorly in our CALL system. In Table 1, some examples about the acoustic confusion and their confusion degree with other finals are presented. It is shown that these long finals are highly confused with others, e.g. 23.62% of 'ing' are mistaken as 'in'. Additionally, we find that 'ing', 'van', 'ueng', 'ang' and 'ian' are all nasal finals. Different from other phonemes, nasal finals are composed of several phonemes, and they are articulated more complexly. So we focus on the sixteen highly-confused nasal finals of Table 2 in this work. The symbol '*' in the following tables stands for tone.

Table 1: *Some acoustic confusion examples for long finals in our CALL system*

Reference	Recognition Result	Acoustic Confusion Degree
ing	in	23.62%
van	ve	15.07%
ueng	eng	12.50%
ua	uang	12.62%
ang	a	11.96%
uo	o	10.26%
ian	ie	8.22%

With our training AM tool, all of the phonemes are modeled with the same state number and GMM number. So, when 'a' is modeled well, 'ueng' can not be modeled accurately. Because of the voice length difference, modeling phonemes in such way is not reasonable and the reliability of such AM is doubtable. So it is necessary to split long finals into sub-phonemes.

Table 2: the nasal finals

an_*	en_*	in_*	un_*	vn_*	ian_*
uan_*	van_*	ang_*	eng_*	ing_*	ong_*
iong_*	iang_*	uang_*	ueng_*		

3.2. Phonological Analysis

As a machine-readable phonetic alphabet, SAMPA-C has developed a labeling convention for standard Chinese. It is mainly for labeling mandarin articulations, which are consonants, vowels, tones, allophones, etc. We need to address the marking symbol for every phoneme or allophone firstly. According to Table 3 and Table 4 of SAMPA-C Labeling Convention for Standard Chinese [14], there are 23 allophones for consonants, including two back-nasal phonemes, '(a)n' and 'ng', and there are 25 allophones for vowels. All of the finals are composed of those allophones. Table 5 and Table 6 showed in [14] demonstrate the similarity between different phonemes. Therefore, we investigate the acoustic confusions of our CALL system in accordance with the SAMPA-C rule.

3.3. Phoneme Splitting Strategies

According to the phonological analysis and the characteristic of mandarin pronunciation quality assessment, we need to accurately model each phoneme. The traditional initial-final syllable structure can not fully describe long finals, which include several allophones, for example, in 'uang', there are three ones. We investigate the method of splitting such articulations to allophone sequences, in order to model every allophone respectively. After splitting, the new pronunciation unit is shorter and its short-time stability is better. So, AM with such units can describe the characteristics of pronunciations more accurately. This is important for the pronunciation quality assessment.

3.3.1. Splitting methods for all of the 145 long finals

In this section, we split all long finals using four methods according to SAMPA-C as follows.

- Baseline: In our baseline system, 219 items are included in the monophone list, which are all initials or finals. 'sil' and 'sp' are labels for silence and pause respectively.
- Splitting Type 1-1: We split all of the finals and obtain 150 phonemes, which are all composed of allophones. Different from SAMPA-C, the allophones from different phonemes can not be tied together. The mapping relations between Splitting Type 1-1 and the baseline are illustrated with some examples in Table 3. The left side of '→' is the baseline phoneme, and the right side is the corresponding separated phoneme.
- Splitting Type 1-2: Based on Splitting Type 1-1, the six zero-initials, which are 'aa', 'ee', 'ii', 'oo', 'uu', 'vv', are added to the phoneme list, 156 items.
- Splitting Type 1-3: Based on Splitting Type 1-2, the last articulations of gingival nasal finals (e.g. an) and velar nasal finals (e.g. ang) are modified with tone, that is to say, the last articulations of nasal finals with different tones are modeled respectively. In this method, there are 164 items in the monophone list.
- Splitting Type 1-4: 166 monophones are trained in AM. Some detailed tunes about the phoneme 'o' are done as shown in Table 4.

3.3.2. Splitting methods for highly-confused nasal finals

For all of the sixteen nasal finals shown in Table 2, the acoustic confusions about them are investigated as the emphasis. On the baseline, we attempt to split the nasal finals with two methods.

- Splitting Type 2-1: Nasal phonemes of 'nn' and 'ng' are separated from the long finals. After mapping as Table

Table 3: The mapping relations between Splitting Type 1-1 and the baseline.

ai*→a1_* i2_*	iong*→i1_* o2_* ng
an*→a1_* nn	iu*→i1_* o4_* u2_*
ang*→a1_* ng	ong*→o2_* ng
ao*→a3_* o2_*	ou*→o4_* u2_*
ei*→e2_* i1_*	ua*→u2_* a2_*
en*→e4_* nn	uai*→u2_* a1_* i2_*
eng*→e4_* ng	uan_*→u2_* a1_* nn
ia*→i1_* a2_*	uang*→u2_* a3_* ng
ian*→i1_* a4_* nn	ueng*→u2_* e4_* ng
iang*→i1_* a3_* ng	ui*→u2_* e2_* i1_*
iao*→i1_* a3_* o2_*	un*→u2_* e4_* nn
ie*→i1_* e3_*	uo*→u2_* o1_*
in_*→i1_* nn	van*→v1_* a5_* nn
ing*→i1_* ng	ve*→v1_* e3_*
io*→i1_* a3_* o2_*	vn*→v1_* nn

Table 4: Tune about the phoneme 'o' in Splitting Type 1-4.

Baseline→before tune→after tune
uo_*→u2_* o1_*→o1_*
ong_*→o2_* ng_*→o3_* ng_*
io_*→i1_* a3_* o2_*→i1_* o5_*
iong_*→i1_* o2_* ng_*→i1_* o3_* ng_*

5, we get 221 monophones.

- Splitting Type 2-2: Based on Splitting Type 2-1, we further modify 'nn' and 'ng'. Like Splitting Type 1-3, the tone is appended at the back of them. That is to say, 'nn' and 'ng' in different tonal finals are distinct.

3.4. Question Design

Because of the top down training strategy requested by the decision tree theory, the question list needs to be designed in some way to support building and merging decision trees. Awarding to[14], we design questions for every phoneme splitting strategy. Generally, the question which involves more articulations is placed more forward the top. The question designed for each phoneme lies at the bottom of the question list. Taking Splitting Type 1-1 for example, a single list of questions is given in Table 6, in which '*' stands for contexts.

4. Experiment

To evaluate the effectiveness of different phoneme splitting strategies for acoustic confusion modeling, we employ two measures. One is in terms of ASR using the average recognition error rate (ARER), and the other is scoring pronunciation qualities using the average scoring error rate (ASER). They are both implemented on our mandarin CALL system.

4.1. Experiment Environment

The database for training AM is a standard mandarin speech data set, about 400 hours. AM used in our experiments are gender dependent, continuous mixture density, state tied, within word triphone HMMs. The HMM sets consist of about 5000 tied-states for female, and 4600 for male, each with 16 Gaussian components. The front end uses MFCC analysis to get

Table 5: Mapping relations between nasal finals and baseline phonemes in Spitting Type 2-1.

an_*→aa_* nn	en_*→ee_* nn
in_*→ii_* nn	un_*→uu_* nn
vn_*→vv_* nn	ian_*→iaa_* nn
uan_*→uaa_* nn	van_*→vaa_* nn
ang_*→ann_* ng	eng_*→enn_* ng
ing_*→inn_* ng	ong_*→onn_* ng
iong_*→ionn_* ng	iang_*→iaan_* ng
uang_*→uaan_* ng	ueng_*→ueen_* ng

Table 6: an example of question design.

Question	Content
NonBoundary	*+* ;
L_Silence	sil-* ;
L_Vowel	a1_1-* a1_2-* a1_3-* a1_4-* a1_5-* a2_1-* ...
L_a	a1_1-* a1_2-* a1_3-* a1_4-* a1_5-* a2_1-* ...
...	...
L_Stops	b-* d-* g-* ;
...	...
L_a1_1	a1_1-* ;
...	...
R_Silence	*+sil ;
R_Vowel	*+a1_1 *+a1_2 *+a1_3 *+a1_4 *+a1_5 ...
R_a	*+a1_1 *+a1_2 *+a1_3 *+a1_4 *+a1_5 ...
...	...
R_Stops	*+b *+d *+g ;
...	...
R_a1_1	*+a1_1 ;
...	...

39-dimensional feature, including 12-dimensional static cepstra and 1-dimensional energy with 1st and 2nd order derivatives.

The test corpus using ARER is about 10 hours standard mandarin speech from 60 speakers: 46 females and 14 males, which is all the single syllable utterance.

The test corpus using ASER is about 5 hours mandarin speech by 111 Hong Kong native undergraduates with a very strong southern China accent. Every speaker reads 75 utterances, of which the first 50 utterances are isolated syllables and the rest 25 utterances are two-syllable words. Machine score for every phoneme is computed as introduced in section 2. Meanwhile, five experts evaluate all of the speakers' pronunciations in the same way. And their evaluation scores are combined to form the final expert assessment result for every initial and final to estimate the accuracy of the CALL system. In order to test the new modeling method convincingly, we collect two groups of PSK (Putonghua Shuiping Kaoshi) test samples, PSK1 and PSK2. Speech content of the test samples within each individual group is the same, but the ones across groups are different. There are 60 speakers in PSK1, and 51 speakers in PSK2.

4.2. Experiment Results

4.2.1. Performance improvement in ASR

We compared those phoneme splitting strategies with the performances of ASR based on ARER, to demonstrate that AM becomes better. The experiment results are shown in Table 7.

Table 7: the performance comparison of different phoneme splitting strategies using ARER.

System	no-tone Syll-ARER	no-tone Pho-ARER	nasal finals ARER
Baseline	24.94%	14.43%	13.68%
Type1-1	23.58%	13.48%	12.87%
Type1-2	23.00%	13.08%	12.87%
Type1-3	22.66%	12.90%	12.01%
Type1-4	22.44%	12.72%	11.95%
Type2-1	24.30%	14.05%	12.86%
Type2-2	23.68%	13.79%	12.38%

From this Table, it is clear that the refined phoneme structure improves the performance as compared to the initial-final syllable structure in our CALL system. From Splitting Type 1-1 to 1-4, ARER is steadily going downward. But Splitting Type 2-2 is not better than Splitting Type 1-4. This may be attributed to the lack of elaborate AM. From the experiment results, we can see that:

1. The six zero-initials, aa/ee/ii/oo/uu/vv, should be included in the monophone list, as is known from the performance change from Splitting Type 1-1 to Splitting Type 1-2.

2. The nasal articulations of 'nn' and 'ng' should be treated as vowels, with five tones. This is supposed by the performance improvement from Splitting Type 1-3.

3. Some refined tune is necessary according to the characteristic of some special articulations.

4.2.2. Performance improvement in scoring pronunciation qualities

In CALL systems, the pronunciation quality assessment from the machine is supposed to be comparable with human assessment. The higher the consistency between them is, the better CALL systems are. ASER is the ratio of the utterance unit number of machine score unequal to human score and the total utterance unit number. We compared three of the phoneme splitting strategies based on ASER. The experiment results are given in Table 8. The initials and finals are context dependent.

Table 8: the performance comparison of different phoneme splitting strategies using ASER.

System	ASRE	Initial ASER	Final ASER	nasal finals ASRE
PSK1				
Baseline	12.64%	10.91%	13.83%	16.29%
Type2-1	12.46%	10.77%	13.63%	14.97%
Type2-2	12.38%	10.76%	13.59%	14.79%
Type1-4	12.05%	10.77%	13.17%	14.49%
PSK2				
Baseline	15.63%	14.75%	16.23%	25.49%
Type2-1	15.55%	14.70%	16.24%	24.95%
Type2-2	15.43%	14.48%	16.18%	24.48%
Type1-4	15.21%	14.52%	15.82%	23.58%

As shown in Table 8, the system using splitting Type 1-4 is the best, which outperforms the baseline system with relative decease of 11.05% and 7.49% ASER for nasal finals in the test set of PSK1 and PSK2 respectively. In other aspects, the phoneme splitting strategies reduce ASER more or less. In fact, the phonemes after splitting may contain more spectrum information, and can be described by AM more elaborately. Therefore, measuring the pronunciation quality with the refined AM can significantly improve system performance.

5. Conclusions

This paper presents our recent study toward optimal modeling of acoustic confusions in mandarin CALL system. We change the traditional mandarin syllable structure to a novel phoneme structure. The change alleviates the problem of acoustic confusions, which is mainly due to that some long finals are not modeled sufficiently. In addition, we systematically compared some phoneme splitting strategies, and shed some light on how to design question list for decision trees. Although the phoneme splitting strategy is investigated in the CALL system, the idea may be extended to apply to the field of ASR.

6. Acknowledgment

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102, 2006AA01Z195), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10574140, 60535030).

7. References

[1] Kazunori Imoto, Yasushi Tsubota, etc, "Modeling and automatic detection of english sentence stress for computer-assisted English prosody learning system", in ICSLP, 2002. IEEE Trans. Speech and Audio Proc., 7(6):697–708, 1999.

[2] Jared Bernstein, "Subrashii: Encounters in japanese spoken language education", CALICO, vol. 16, pp. 361-384, 1999.

[3] Goh Kawai, "A call system using speech recognition to train the pronunciation of japanese long vowels, the mora nasal and mora obstruent", in Eurospeech, 1997. "Interactive problem solving with speech", J. Acoust. Soc. Amer., Vol. 84, 1988, p S213(A).

[4] Sherif Mahdy Abdou, Salah Eldeen Hamid, etc, "Computer aided pronunciation learning system using speech recognition technology", in Interspeech, 2006.

[5] L. Neumeyer, H. Franco, etc, "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech", Proc. of ICSLP 96, pp.1457-1460, Philadelphia, Pennsylvania 1996.

[6] K. Tatsuya, D. Masatake, etc, "Practical use of English pronunciation system for Japanese students in the CALL classroom", INTERSPEECH-2004, pp. 1689-1692, 2004.

[7] 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.

[8] L. Neumeyer, H. Franco, etc, "Automatic Scoring of Pronunciation Quality", Speech Communication, Volume 30, Issues 2-3, February 2000, Pages 83-93.

[9] H. Franco, L. Neumeyer, etc, "Combination of machine scores for automatic grading of pronunciation quality", Speech Communication, volume 30, 2000.

[10] 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.

[11] J. Bernstein, M. Cohen, etc, "Automatic Evaluation and Training in English Pronunciation", ICSLP 1990, Kobe, Japan.

[12] Jiang-Chun Chen, etc, "Automatic Pronunciation Assessment for Mandarin Chinese", IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, June 2004.

[13] F. P. Pan, Q. W. Zhao, Y. H. Yan, "Improvements in Tone Pronunciation Scoring for Strongly Accented Mandarin Speech", Proceedings of ISCSLP 2006, pp. 592-602, 2006.

[14] http://www.d-ear.com/CCC/resources/SampaC.pdf