INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Phoneme, Phone Boundary, and Tone in Automatic Scoring of Mandarin Proficiency

Jiahong Yuan and Mark Liberman

Linguistic Data Consortium, University of Pennsylvania
jiahong@ldc.upenn.edu, myl@ldc.upenn.edu

Abstract

Not every phone, word, or sentence is equally good for assessing language proficiency. We investigated three phonetic factors that may affect automatic scoring of Mandarin proficiency – phoneme, phone boundary, and tone. Results showed that phone boundaries performed the best, and within-syllable boundaries were better than cross-syllable boundaries. The retroflex consonants as well as the vowel following these consonants outperformed the other phonemes. Tone0 and Tone3 outperformed the other tones, and ditone models significantly improved the performance of Tone0. These results suggest that phone boundary models and phoneme- and tone- dependent scoring algorithms should be employed in automatic assessment of Mandarin proficiency. It may also be helpful to separate phoneme and tone scoring prior to the combination of individual scores, as we found that the worst phoneme and the best tone, with respect to automatic scoring of Mandarin proficiency, appeared in the same word.

Index Terms: spoken language proficiency, automatic scoring, goodness of pronunciation

1. Introduction

Automatic scoring of spoken language proficiency has been widely applied in language tests and computer assisted language learning (CALL) [1,2]. The common practice is to build HMM-based acoustic models using a large amount of "standard" speech data. To assess an utterance, pronunciation scores such as log likelihood scores and posterior probabilities are calculated by performing speech recognition (or forced alignment if the sentence is known) to the utterance based on the pre-trained acoustic models [3-6]. Prosody scores, e.g., duration, F_0, and pauses, have also been shown important [7, 8]. These individual scores are combined with statistical models such as linear regression, SVM, and neural network to produce an overall score for the test utterance [9].

In prior work [10], we demonstrated that not every sentence is equally good for assessing language proficiency, and the performance of an automatic scoring system could be significantly improved by excluding "bad" sentences from the scoring procedure. Much research is needed to understand the linguistic factors that determine the goodness of a sentence for automatic proficiency scoring. In this study, we investigate three phonetic factors that may affect automatic scoring of Mandarin proficiency – phoneme, phone boundary, and tone.

It is well known that some phonetic contrasts are more difficult in language learning. The retroflex consonants (/zh, ch, sh, r/) in Mandarin Chinese, for example, are difficult to learn for many speakers whose first language does not have retroflex sounds. The pronunciation of these consonants is a prominent cue for native speakers to perceive accent. Phone boundaries may also contain useful information about a speaker's language proficiency. The timing of voicing in stop consonants, which is measured by voice onset time (VOT), is a boundary-bound phonetic feature that has been extensively studied in linguistics [11, 12]. The VOT of stops varies across languages. Individuals who learn an L2 later in life are often failed to produce consonants with authentic VOT values in L2 [13]. Finally, the non-native production of tone is probably the most salient characteristic of foreign accent in Mandarin Chinese. Chinese speakers may find less difficult in acquiring Mandarin tones because there is usually a systematic mapping between Mandarin tones and the tones in their first language. Nonetheless, the production of tone is still problematic to less fluent speakers and, some tones may be more likely to bear accent than others.

In the following sections we first introduce the dataset in Section 2. Section 3 and 4 describe the experiments and report the results, on phoneme and phone boundary and on tone, respectively. Finally, Section 5 is conclusions.

2. Data

We used a dataset of *Putonghua Shuiping Ceshi* (PSC) from Beijing Normal University. PSC is the national standard Mandarin proficiency test in China, which is taken by several million people each year. The test consists of four parts: The first two parts are to read 100 monosyllabic and 50 disyllabic words; the third part is to read an article of 300 characters, randomly selected from a pool of 60 articles; and the last part is to speak freely on a given topic. The four parts are graded separately with a numeric score, and the total score (out of 100 points) is converted to a categorical proficiency level, out of seven levels.

Our dataset consists of recordings of ~800 college students at Beijing Normal University who took the PSC test in 2011 and the grades they received on the test. We only used the part of article reading in this study. The students who read an article being selected for less than 9 other students (i.e., the total number of students reading that article is less than 10) were excluded. The students who had proficiency scores in the lowest two levels were also excluded. The final dataset contains 604 speakers reading 42 articles. Each speaker was graded by two examiners. The average of the two examiners' scores on the part of article reading was used as the speaker's proficiency score.

We selected 143 speakers who had the best proficiency scores in the dataset to train models of "standard" speakers. The rest 461 speakers were tested using the "standard" models. The goodness of pronunciation of phoneme, phone

http://dx.doi.org/10.21437/Interspeech.2016-510

boundary, and tone was calculated by an approximation of its posterior probability given the data and model. The correlation between the goodness of pronunciation scores and the examiners' scores on the 461 speakers was used to determine the usefulness of a phone, phoneme boundary, and tone in automatic scoring of Mandarin proficiency. A greater correlation is expected if a phone, phone boundary, or tone bears more information about language proficiency. Detailed methods and results are described below.

3. Phoneme and Phone Boundary

3.1. Training acoustic models

GMM-HMM acoustic models of phonemes (initials and finals) and phone boundaries were trained on the utterances of the 143 "standard" speakers. While researchers have disagreed on the vowel phonemes in Mandarin Chinese, the inventories of initials and finals in the language are largely straightforward. The initials are consonants. A final in Mandarin Chinese may consist of one or more vowels (or vowels and glides, depending on the adopted phonological analysis), with or without a nasal coda. The final /i/ has three pronunciation variants, often transcribed as [ɿ] (when following an alveolar fricative or affricate), [ʅ] (when following a retroflex fricative or affricate), and [i] (in all other contexts). The three variants were treated as different finals, /i/ for [i], /ii/ for [ɿ], and /iii/ for [ʅ]. In our acoustic models, initials, monophthong finals (/a, e, i, ii, iii, u, v/), and silence were 3-state HMMs, all other finals (including diphthongs, triphthongs, and nasal-coda finals) were 5-state HMMs.

The phone boundary models were a special 1-state HMM (as shown in Figure 1), in which the state cannot repeat itself. Therefore, a boundary can have one and only one state occurrence, i.e., aligned with only one frame. In prior work [14, 15], we demonstrated that employing explicit phone boundary models within the HMM framework could significantly improve forced alignment accuracy.

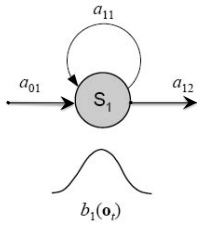

Figure 1: *Special 1-state HMM for phone boundaries with transition probabilities $a_{11} = 0$ and $a_{12} = 1$.*

The special 1-state phone boundary HMMs were combined with phone HMMs. Given a phonetic transcription, phone boundaries were inserted between phones. For example, "*sil i g e sil*" became "*sil sil_i i i_g g g_e e e_sil sil*". The boundary states were tied through decision-tree based clustering during the training procedure, similar to triphone state tying in speech recognition. The acoustic models were trained on the standard 39 PLP (Perceptual Linear Prediction) features extracted with 25ms Hamming window and 10ms frame rate, using the HTK Toolkit.

3.2. Computing goodness of pronunciation

Following the method in [5], we computed a goodness of pronunciation score for every phone and phone boundary in the utterances of the 461 test speakers. The idea is to find the posterior probability of a phone p given its acoustic segment $O^{(p)}$, $P(p|O^{(p)})$, which can be approximated by the likelihood of $O^{(p)}$ corresponding to phone p, divided by the maximum likelihood of $O^{(p)}$:

$$\mathrm{GOP}(p) = \log \frac{p(o^{(p)} \mid p)}{\max_{q \in Q} p(o^{(p)} \mid q)} \quad (1)$$

where Q is the set of all phone and boundary models. The acoustic segment boundaries of $O^{(p)}$ and the corresponding likelihood (the numerator) was determined by forced alignment. To compute the maximum likelihood of $O^{(p)}$ (the denominator), all test utterances were recognized using the acoustic models and an unconstrained phone and boundary loop. The likelihood of $O^{(p)}$ corresponding to the best hypothesis within its boundaries (it may contain more than one phones or boundaries) was used to approximate its maximum likelihood.

The goodness of pronunciation scores computed from (1) are zero or negative. They are expected to have a positive correlation with human scores: A lower goodness of pronunciation score suggests that the phone or boundary fits the "standard" models less well hence should receive a lower proficiency score.

3.3. Results

For every speaker in the test set, we calculated his/her mean goodness of pronunciation score on every phoneme. The phone boundaries were grouped into two types: within-syllable (i.e., boundaries between an initial and a final) and cross-syllable (i.e., boundaries between a final and an initial), and a mean goodness of pronunciation score was calculated for each type. For each phone and boundary type, we then computed the correlation between the 461 speakers' mean goodness of pronunciation scores and their proficiency scores (461 pairs of scores). The results are listed in Table 1. Only the phones that appeared in every test speaker' utterances are included.

We can see from Table 1 that the correlation varies greatly across phonemes. The two boundary types have the highest correlations, suggesting that phone boundaries are more helpful than phonemes in automatic proficiency scoring. Within-syllable boundaries work better than cross-syllable boundaries. Among the phonemes, the retroflex consonants, /zh, ch, sh/, and the vowel following these consonants, /iii/, are better than the others. The vowel /e/ is the only phoneme that has a negative correlation, although the correlation is not significant. /e/ appears in the possessive particle "的" (de0) in Mandarin Chinese, which is the most frequent word in the language. In our dataset, there are 23,501 /e/ tokens, 15,919 (64.7%) of the tokens were from the word "的" (de0). "的" (de0) has a neutral tone, and its vowel is similar to the schwa in English.

Table 1. *Correlations between goodness of pronunciation and proficiency scores: on phoneme and phone boundary.*

Phone or boundary	Correlation (Pearson's r)	Phone or boundary	Correlation (Pearson's r)
within-syl	**0.472**	g	0.157
cross-syl	**0.445**	r	0.144
iii	**0.422**	b	0.141
sh	**0.383**	uan	0.126
zh	**0.327**	m	0.125
s	0.277	iao	0.120
a	0.271	iu	0.114
ch	**0.269**	ai	0.114
ian	0.256	ei	0.112
i	0.245	n	0.111
ing	0.238	eng	0.110
d	0.225	en	0.102
h	0.225	ie	0.100
an	0.224	k	0.060
l	0.214	ong	0.054
z	0.210	uo	0.052
q	0.202	ao	0.045
t	0.194	iang	0.041
j	0.192	u	0.036
f	0.190	ang	0.029
in	0.182	v	0.019
x	0.179	ii	0.007
ui	0.174	***e***	***-0.004***

*The correlations lower than 0.12 are not significant.

4. Tone

4.1. Training tone models

There are five tones in Mandarin Chinese, Tone 1 through Tone 4, and the neutral tone, Tone 0. GMM models were trained on each of the five tones using the 143 "standard" speakers' utterances. Tone boundaries, which are the same as syllable boundaries in our investigation, were obtained by forced alignment using the acoustic models from Section 3. Kaldi pitch features were extracted over the duration of tones [16], which include frame-wise normalized pitch, probability of voicing, and delta-log-pitch. We then applied DCT (Discrete Cosine Transform) to each of the feature contours within the tone boundaries. A fixed number of DCT coefficients were used for all tones to train GMMs, regardless of the duration of the tone. The duration of tone was used as a separate feature.

The phonetic realization of a tone is greatly affected by its context, especially the preceding tone. Besides monotone GMMs, we also trained GMMs for ditones, i.e., the combination of two tones, such as T1+T2. For every ditone in the dataset, its pitch features were extracted over the duration of two tones, whereas its duration feature was the duration of the second tone.

For monotone GMMs, we employed four DCT coefficients and 50 Gaussian mixtures; for ditone GMMs, we employed six DCT coefficients and 10 Gaussian mixtures.

4.2. Training tone models

The goodness of pronunciation score on tone was calculated using formula 2:

$$\mathrm{GOP}(t) = \log \frac{p(o^{(t)} \mid t)}{\sum_{x \in T} p(o^{(t)} \mid x)} \quad (2)$$

where T is the set of all tone models, five in the monotone set, and 24 in the ditone set (excluding T0+T0, which is rare in the dataset). Different from formula (1), the denominator in formula (2) is the sum of the likelihoods of the acoustic segment $O^{(t)}$ corresponding to all models in the monotone or ditone set.

For monotone models, to calculate the mean goodness of pronunciation score on a tone is straightforward. For ditone models, the second tone in a ditone was used as the base to calculate the mean scores. For example, the mean goodness of pronunciation score on Tone 2 for ditone models was the average of the scores from five ditones, T0+T2, T1+T2, T2+T2, T3+T2, and T4+T2

4.3. Results

Figure 2 and 3 show the correlations between the test speakers' mean goodness of pronunciation scores and their proficiency scores, for tone models trained on DCT coefficients of Kaldi pitch features and trained on the duration of tone, respectively. The results of monotone and ditone models are compared.

We can see from Figure 2 that ditone models significantly improved the correlation on Tone0, and Tone0 and Tone3 have higher correlations than the other tones. From Figure 3 we can see that when tone models were trained on the duration of tone only Tone0 has a significant correlation. There are 20,795 Tone0s in the dataset, 15,919 (73.1%) of them are the word "的" (de0). Combining the result on the vowel /e/ discussed in Section 3, we can see that "的" (de0), the most frequent word in Mandarin Chinese, has an interesting position in automatic scoring of Mandarin proficiency: it is the worst when phone models are employed but the best when tone models are employed.

Table 2 lists the correlations for monotone and ditone models trained on the combination of DCT coefficients of Kaldi pitch features and the duration of tone. Again, Tone0 and Tone3 are better than the other tones. The best correlation is from Tone0 with ditone models.

Table 2. *Correlations between goodness of pronunciation and proficiency scores: on tone.*

	Tone0	Tone1	Tone2	Tone3	Tone4
monotone	0.238	0.210	0.165	0.279	0.085
ditone	**0.364**	0.212	0.182	0.269	0.184

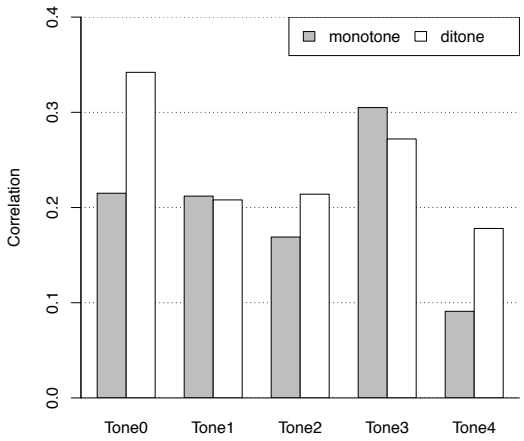

Figure 2: *Correlations between goodness of pronunciation and proficiency scores: tone models were trained using DCT coefficients of Kaldi pitch features*

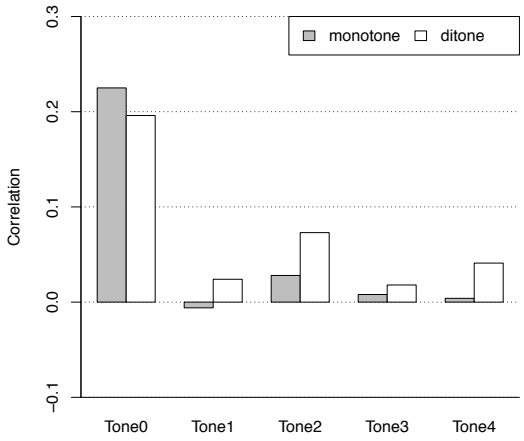

Figure 3: *Correlations between goodness of pronunciation and proficiency scores: tone models were trained using the duration of tone.*

5. Conclusions

We investigated three phonetic factors that may affect automatic scoring of Mandarin proficiency – phoneme, phone boundary, and tone. Acoustic models of "standard" speakers were trained and used to compute goodness of pronunciation scores for test speakers. The correlations between the test speakers' goodness of pronunciation scores and their proficiency scores on individual phonemes, phone boundaries, and tones were computed. Phone boundaries had the best correlations, of which within-syllable boundaries were better than cross-syllable boundaries. The retroflex consonants, /zh, ch, sh/, and the vowel following these consonants, /iii/, outperformed the other phonemes. With regard to the use of tones, ditone models significantly improved the correlation on Tone0, and Tone0 and Tone3 performed better than the other tones. These results suggest that phone boundary models and phoneme- and tone- dependent scoring algorithms should be employed in automatic assessment of Mandarin proficiency. We also found that the worst phoneme with respect to automatic scoring of Mandarin proficiency was /e/, which is the vowel in the most frequent word "的" (de0) in the language. However, the word "的" (de0) was the best when scored on tone models. This result suggests that it may be helpful to separate phoneme and tone scoring prior to the combination of individual scores to produce an overall proficiency score.

6. Acknowledgements

We thank Xiaoying Xu, We Lai, Weiping Ye, and Xinru Zhao for providing the dataset used in this research. This work is supported in part by Penn China Research & Engagement Fund.

7. References

[1] Wang, R. Liu, Q. and Wei S, "Putonghua Proficiency test and evaluation," In: *Advances in Chinese Spoken Language Processing*, pp. 407-429, 2006.

[2] Zechner, K., Higgins, D., Xi, X. and Williamson, D., "Automatic Scoring of Non-Native Spontaneous Speech in Tests of Spoken English," *Speech Communication*, 51, pp. 883–895, 2009.

[3] Franco, H., Neumeyer, L., Kim, Y. and Ronen, O., "Automatic pronunciation scoring for language instruction," *Proceedings of ICASSP 1997*, pp. 1471-1474, 1997.

[4] Neumeyer, L., Franco, H., Digalakis, V. and Weintraub, M., "Automatic scoring of pronunciation quality," *Speech Communication*, 30, pp. 83-93, 2000.

[5] Witt, S. and Young, S., "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, pp. 95-108, 2000.

[6] Hu, W., Qian, Y., Soong, F. and Wang, Y., "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, 67, pp. 154-166, 2015.

[7] Cucchiarini, C., Strik, H. and Boves, L., "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *Journal of the Acoustical Society of America*, 107, pp. 989-999, 2000.

[8] Nava, E., Tepperman, J., Goldstein, L., Zubizarreta, M. and Narayanan, S., "Connecting rhythm and prominence in automatic ESL pronunciation scoring," *Proceedings of Interspeech 2009*, pp. 684-687, 2009.

[9] Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O., "Combination of Machine Scores for Automatic Grading of Pronunciation Quality," *Speech Communication*, 30, pp.121-130, 2000.

[10] Yuan, J., Xu, X., Lai, W., Ye, W., Zhao, X. and Liberman, M., "Sentence selection for automatic scoring of Mandarin proficiency," to appear in *Proceedings of SIGHAN-8*, 2015.

[11] Lisker, L. and Abramson, A., "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," *Word* , 20, pp. 384-422, 1964.

[12] Cho, T. and Ladefoged, P., "Variation and Universals in VOT: Evidence from 18 Languages," *Journal of Phonetics*, 27, pp. 207-229, 1999.

[13] Flege, J., "Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language," *Journal of the Acoustical Society of America*, 89, pp. 395-411, 1991.

[14] Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V. and Wang, W., "Automatic phonetic segmentation using boundary models", *Proceedings of Interspeech 2013*, pp. 2306-2310, 2013.

[15] Yuan, J., Ryant N., and Liberman, M., "Automatic phonetic segmentation in Mandarin Chinese: boundary models, glottal features and tone," Proceedings of *ICASSP 2014*, pp. 2539-2543, 2014.

[16] Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal J. and Khudanpur, S., "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition", *Proceedings of ICASSP 2014*, pp. 2494-2498, 2014.