ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Automatic Tone Assessment of Non-Native Mandarin Speakers

Jian Cheng

Knowledge Technologies, Pearson
4040 Campbell Ave., Menlo Park, California 94025, USA
jian.cheng@pearson.com

Abstract

In this paper, we discuss the methods used to assess non-native Mandarin speakers' tone automatically. A context-dependent syllable-level tone modeling method for tone assessment is proposed. A direct comparison between a speaker's contours and ideal contours in energy and pitch using a syllable-level normalization technique provides a strong prediction of the speaker's tone as rated by humans. By combining features from energy and pitch with other features such as duration and spectral likelihoods at the phoneme level, we achieved a human-machine correlation coefficient of 0.77 at the response level and 0.85 at the participant level. As a comparison, the correlation coefficient between human raters was 0.66 at the response level. The results support both the new proposed method and also the use of Read Aloud as a task to assess non-native Mandarin speaker's tone automatically.

Index Terms: Chinese, Mandarin, tone, prosody, assessment, proficiency test

1. Introduction

Peking University and Pearson collaboratively developed a Spoken Chinese Test (SCT) designed to measure how well a person understands and speaks common standard Mandarin Chinese [1]. SCT is intended for adults and students over the age of 18. Computerized scoring [2, 3] allows for immediate, objective, reliable results that correspond well with traditional measures of spoken Chinese performance. Pearson's automated speech scoring technologies have been used in the Versant English, Spanish, French, Arabic, and Dutch tests [2, 3, 4]; however, none of these languages is tonal. Mandarin is the first deployed test to automatically score a tonal language. In order to assess non-native speakers' facility with tone in Mandarin, several sections in the SCT were designed to provide a tone subscore. One of these sections is Read Aloud. In this paper, we present research on the Read Aloud section in the SCT to assess a non-native Mandarin speaker's tone production automatically.

There is a considerable amount of previous work on Chinese tone processing, including both explicit and embedded tone modeling. In explicit tone modeling, individual segmented syllables are classified using various machine learning methods that take as input self-defined supra-segmental features [5, 6, 7, 8, 9]. Another approach is to incorporate the recognized tones from explicit tone modeling or tone related features (embedded tone modeling) to improve the automatic speech recognition (ASR) accuracy of Mandarin [7, 8, 10, 11]. In explicit tone modeling, the research either focuses on isolated syllable tone recognition or syllable tone recognition in continuous speech. The latter is close to the research presented here although our focus is on assessment rather than diagnosis. Some recent results from the literature are summarized in Table 1. Lattice refinement using phonological rules and language models may not apply to non-native speech. Although these results are not comparable because they used different test sets, they provide us with a rough idea what the state-of-the-art is in tone classification accuracy for native Mandarin continuous speech.

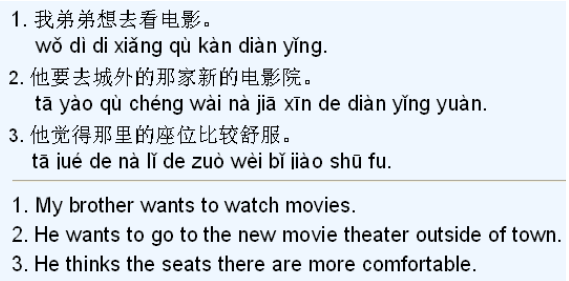
Table 1: Tone correct rate (TCR) reported by different groups using native Mandarin continuous speech.

Method	TCR
Decision tree	71.2% [5]
SVM + Normalization	83.1% [6]
Neural network	76.2% [8]
Tri-tone HMM + Unbroken pitch contour	74.4% [9]
Above + Lattice refinement	85.1% [9]
MSD-HMM without explicit segmentation	88.8% [12]

Almost all the previous work on Mandarin tones makes use of native speech. When applying known techniques to non-native speech, significantly lower accuracy is expected. At the same time, larger numbers of false positives are anticipated. Even for trained human raters, it is very challenging to make a binary decision about tones produced by non-native speakers. Thus, we mainly focused on automatically assessing tone production as a whole. We addressed tone assessment separately from pronunciation assessment.

2. Spoken Chinese Test

The Spoken Chinese Test (SCT) has eight task types [1]: Tone Phrases, Read Aloud, Repeats, Short Answer Questions, Tone Recognition (Word), Tone Recognition (Sentence), Sentence Builds, and Passage Retellings. The SCT score report provides an Overall score and five diagnostic subscores which describe the test-taker's facility in spoken Chinese. The Overall score is a weighted average of the five subscores: Grammar, Vocabulary, Fluency, Pronunciation, and Tone.



1. 我弟弟想去看电影。
wǒ dì di xiǎng qù kàn diàn yǐng.
2. 他要去城外的那家新的电影院。
tā yào qù chéng wài nà jiā xīn de diàn yǐng yuàn.
3. 他觉得那里的座位比较舒服。
tā jué de nà lǐ de zuò wèi bǐ jiào shū fu.

1. My brother wants to watch movies.
2. He wants to go to the new movie theater outside of town.
3. He thinks the seats there are more comfortable.

Figure 1: Examples of the Read Aloud task.

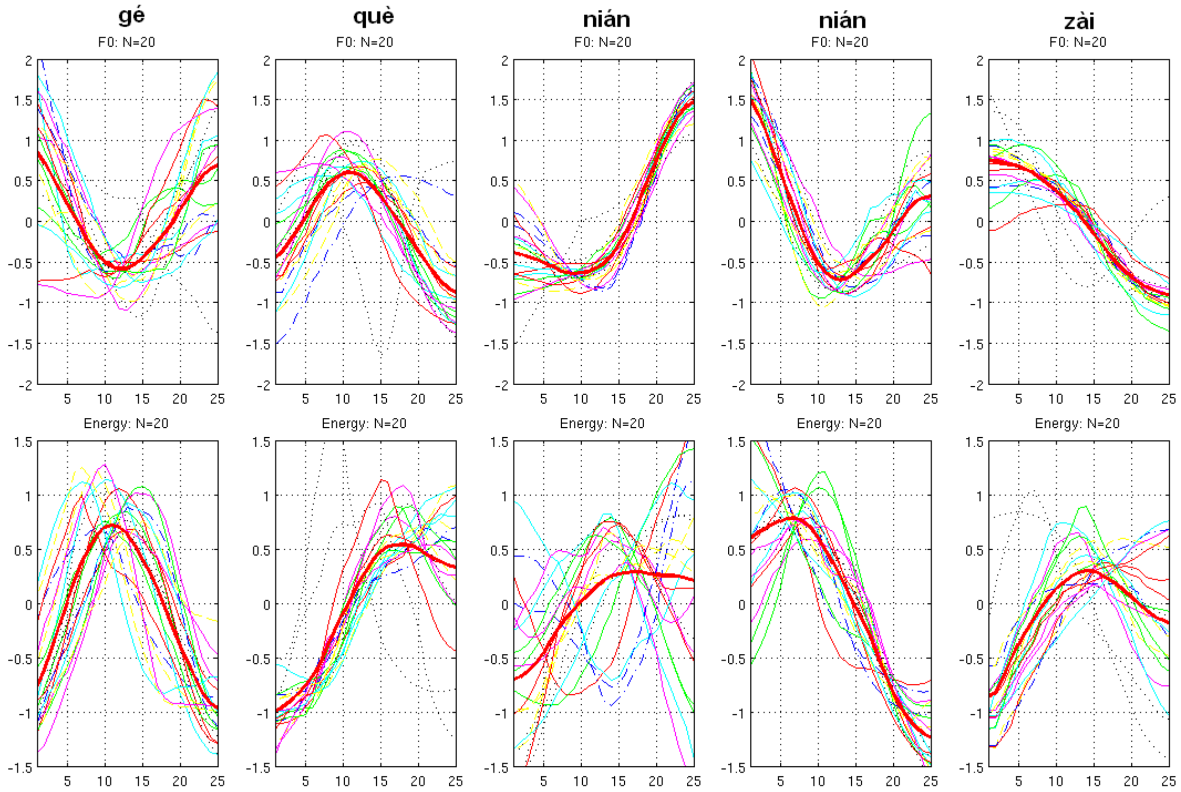


Figure 2: A typical example of models for individual syllables. The syllables presented here are five continuous syllables appearing in the middle of a sentence. The x-axis is 25 equivalent distance points. The y-axis is normalized z-scores. The rows show graphs of F0 (top) and Energy (bottom) for all the native samples. The heavy lines are the averages that are used for the final models.

2.1. Read Aloud

In the Read Aloud task, test-takers read printed, numbered sentences (or items), one at a time, in the order requested by the examiner voice. Read Aloud items are grouped into sets of three sequentially coherent sentences as in the example above. Two sets of three sentences are presented. There are a total of about 79 syllables in this task per test, or about 13 syllables per item.

3. Syllable level tone modeling

Mandarin has a contour tone system in which tones depend on the shapes of the pitch contours instead of the relative pitch levels as in a register tone system. As long as the shapes of the pitch contours produced are similar to those of ideal contours, the correct tones are produced. Because of tone sandhi, tone coarticulation, phrase boundaries, stress, emotional expression, etc., in continuous Mandarin speech, it is well-known that the pitch contour for a syllable does not correspond well with the canonical pattern of its lexical tone. Therefore, the tone modeling should be context dependent. All the previous work mentioned here used context dependent methods. The only issue is in what extend the context should be integrated into the models.

With tone assessment, we considered three types of tone problems:

- Tone Error: The character has an incorrect tone or has no identifiable tone.
- Tone Defect: The character's tone is identifiably correct, but does not sound fluent and native-like (includes non-native tone register, pitch range, or contour).
- Context Defect: The tone is not modified or neutralized as expected in context (includes tone-swaps, missing neutralization, and missing sandhi).

Context is an important factor in determining whether or not a tone has an error. Incidentally, our preliminary research showed that human raters had significant difficulty making consistent binary judgments. Thus, we developed a method of judging overall tone quality. We proposed a novel method of analyzing prosody patterns at the syllable level by considering the context of the syllable. Some ideas related to this new method were based on previous research results [13]. In the current research, some adjustments were made regarding the treatment of special Mandarin characters. The main idea is to determine how close non-native tones are to native tones.

Preprocessing. We ran each recording through ASR with an item specified bigram language model that was estimated on the item. The F0 and energy were calculated at 10 ms intervals using an ESPS pitch tracker get_f0. The recognition results were aligned with the expected answer at the syllable level. Any unmatched syllables were ignored. Each matched syllable within the recording was associated with a time boundary, F0 and energy values. Different from various normalization methods mentioned in the literature [7, 8, 13] such as normalization per speaker or per utterance, or moving window normalization (MWN), etc., we normalized both F0 and energy values by converting them to z-scores using syllable-level specific means and variances. Our reasoning for this approach was that Mandarin supports syllable level normalization (SLN) because of its contour tone system. MWN addresses F0 declination over an utterance; whereas SLN nullifies this issue. More detailed discussions about different normalization strategies are presented in Section 5.

Modeling. For unvoiced segments of speech, we used the linear interpolation method. The interpolation carried context-dependent information from previous and next voiced parts. Then, both F0 and energy were smoothed by a mean filter with a span of 13 frames at the syllable level. The number of F0 and energy frames for the same syllable spoken by the same or different speakers was often different. Instead of using dynamic time warping to align the syllables, all the F0 and energy frames for a syllable were sampled at 25 equivalent distance points that spanned the length of the syllable. Based on our observations, 25 sample points was enough to represent the different shapes of both F0 and energy at the syllable level. Using a syllable as a computational unit in Mandarin tone assessment was straightforward since every syllable had a specified tone and usually energy started and ended at the syllable level.

Every response in the native set was used to build F0 and energy canonical contour models. An underlying assumption was that native speakers would not have problems producing the correct tones. We built individual F0 and energy canonical contour models for every syllable in an item. The context was naturally embedded in such a model. The average feature vectors from the native data were used as our final models for every syllable. A typical example is displayed in Figure 2. From the figure, we can see that there are strong patterns for both F0 and energy contours. Outliers were always present. Some syllables had more stable shapes than others. The pitch contours did not correspond well with the canonical patterns of their isolated syllables.

Procedure. For each new response recording, we ran ASR and aligned the recognized results with the expected answer. For the matched syllables that had models, we computed the raw score as the Euclidean distance between the feature vector for that syllable and the model vector and then normalized the scores for each syllable. The final score for an utterance was the unweighted average. The normalization procedure in this step actually is an automatic weighting process, such that syllables with unstable shapes contribute less to the final score.

4. Some other features

Aside from F0 and energy, duration is an important component of prosody. Phoneme level duration statistics may be sensitive to the fact that one tone may be longer than others. Duration statistics from native speakers were used to compute the log likelihood for durations of phonemes produced by non-native speakers: log_seg_prob. More detail can be found in [13].

The phoneme set used in our ASR carried tone information. We computed a few spectral likelihood features [3] although their individual predictive power was weaker than that of features from F0, energy and duration. For every response, the percentage of syllables read correctly ($percent\_correct$) from ASR was used as another feature.

5. Experiments and results

Experimental data. Recordings of Read Alouds from SCT's Phase A field test data were used. For this dataset, a total of 39 different languages were reported as the first language of the participants. The sample rate for the recordings was 8 kHz with 8 bits (telephone band). For context-dependent syllable-level tone modeling, 336 tests from 237 native participants were used. Ninety-nine participants took the test twice, but with different test items. On average, there were 17 responses per Read Aloud item. We believe these utterances cover the majority of the tone patterns natives typically produce for these given items. For development purposes, 151 randomly selected tests from 151 non-native participants were used to develop the models. For validation purposes, 120 randomly selected tests from 120 non-native participants were used. For every test, three randomly selected responses from six Read Aloud items were rated by two different human raters. All of the 10 trained human raters were educated native Mandarin speakers. Human raters rated these responses on a 6-point scale, with 6 representing the best tone rating. The rating criteria used was not listed here because of space constraints. The raters were asked to pay attention to tone error, tone defect and context defect. Rating 0 was used to identify silence or irrelevant or completely unintelligible material. There were 16 such ratings that were removed from this study.

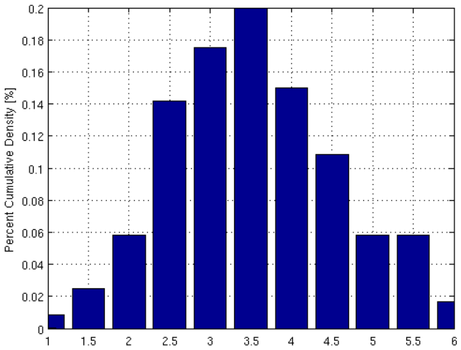


Figure 3: A histogram showing the average tone score distribution for individuals in the validation set as rated by humans.

Figure 3 shows the average tone score distribution for individual participants in the validation set. For the 11 pairs of raters whose number of ratings was more than 10, the average of the inter-rater correlations at the response level was 0.66. For each response, we randomly assigned one rating to Group 1 and the other to Group 2. Using this approach, the average correlation was 0.65. This level of correlation indicates that it is difficult for human raters to make reliable tone judgments.

In a real testing environment, test takers usually make mistakes such as insertions, deletions, and/or substitutions even in a Read Aloud task. In the validation set, when we compared the human transcriptions for these responses with the expected answers, the character error rate (CER) was 8.9%; for our ASR results, the CER was 6.5%.

Table 2: Correlations between human and machine derived by using different normalization methods in the development set.

Normalization method	Speaker	Utterance	Syllable
F0	0.383	0.281	0.556
Energy	0.401	0.432	0.505

Normalization. Both F0 and energy values were normalized. Both values were converted to z-scores using speaker-level, utterance-level, or syllable-level specific means and variances. Correlations between machine-generated scores and average human ratings are presented in Table 2. The results show that syllable level normalization (SLN) is more suitable for assessment purposes. SLN was, therefore, used in the final model.

Our explanation for these results is that SLN decreases the amplitude variances for both F0 and energy.

Ideal contours. Because we observed outlier contours for almost every syllable produced by natives as seen in Figure 2, we explored a method of disregarding a certain percentage of contours that were farthest from the average contour. Only very small performance changes were observed from this method. For example, when removing 10% of the outlier samples, the correlations became 0.547 for F0 and 0.499 for energy.

Table 3 lists the relationship between the number of native samples used to build ideal contours and the corresponding correlations. From the table we see that further increasing the number of native samples shows a trend of diminishing returns.

Table 3: Correlations between human ratings and machine scores derived by changing the number of native samples used to build ideal contours. E stands for Energy.

N	1	2	4	8	16	all
F0	0.461	0.479	0.531	0.544	0.554	0.556
E	0.452	0.478	0.483	0.490	0.504	0.505

Other methods. In our previous research [13], we proposed a method to assign the F0 of unvoiced segments a fixed low value to capture the difference between voiced and unvoiced segments (such as durations), then used the k-means clustering method to build canonical contour models at the word level for F0 and energy. This approach significantly improved the predictive power of F0 in English intonation assessment. Applying this technique here directly, the correlations became 0.492 for F0 and 0.475 for energy. These results were significantly better than the method proposed here with utterance-level normalization (Table 2, Column 3). We argue that the k-means separates amplitude variances to different clusters. We evaluated this by replacing the average method used by Table 2, Column 3 with the k-means. The correlations became 0.350 for F0 and 0.450 for energy. When changing the normalization method in [13] to SLN, the correlations became 0.524 for F0 and 0.487 for energy. These differences could suggest that after using SLN to decrease the amplitude variances, the performance gains from the k-means are no longer significant.

Validation results. Table 4 lists the correlations. Although most sentences were read correctly (with only 8.45% of the responses having $percent\_correct$ values less than 1), $percent\_correct$ was still a powerful predictor of the speaker's final score. Using the development set, a backpropagation neural network model was built using features from F0, energy, phoneme duration likelihoods and a few spectral likelihoods. We applied the neural network model to the validation set and computed the correlation between the machine-generated score and the average human tone rating, which was 0.77.

Table 4: Correlations at the response level using different features in the validation set.

Features	Correlation
F0	0.545
Energy	0.500
F0 + Energy	0.584
log_seg_prob	0.433
$percent\_correct$	0.434
Neural network	0.765

If we use the average of all human ratings as the participant's final human score and the average of all machine scores as the participant's final machine score, at the participant level, the final correlation was 0.85. This result was achieved by using only six Read Aloud items. When combining this score with the tone subscores from other tasks, we are able to provide a reliable assessment of the speaker's tone in Mandarin.

6. Conclusions

We presented research on automatic tone assessment of non-native Mandarin speakers using the Read Aloud section in a Spoken Chinese Test. A new context-dependent syllable-level tone modeling method that is suitable for assessment purposes was proposed and validated. Direct syllable-level normalization improved the proposed method significantly. Combing features from energy and pitch with other features such as duration and spectral likelihoods at the phoneme level, our machine scores correlated with human ratings at 0.77 at the response level and 0.85 at the participant level. Considering only six Read Aloud items were used, these results are very promising. The findings support both the new proposed method and also the use of Read Aloud as a task to assess non-native Mandarin speaker's tone automatically.

7. References

[1] X. Xu, M. Suzuki, and J. Cheng, "An automated assessment of spoken Chinese: Technical definition of Hanyu standards for content and scoring development," in *The 7th International Conference & Workshops on Technology & Chinese Language Teaching: Conference Proceedings 2012*, pp. 468–473.

[2] J. Bernstein and J. Cheng, "Logic and validation of a fully automatic spoken English test," in *The Path of Speech Technologies in Computer Assisted Language Learning*, V. M. Holland and F. P. Fisher, Eds., pp. 174–194. Routledge, New York, 2007.

[3] J. Bernstein, A. Van Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[4] J. Cheng, J. Bernstein, U. Pado, and M. Suzuki, "Automated assessment of spoken modern standard Arabic," in *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, 2009, pp. 1–9.

[5] P.F. Wong and M.H. Siu, "Decision tree based tone modeling for Chinese speech recognitions," in *ICASSP 2004*, pp. 905–908.

[6] G. Peng, H. Zheng, and W.S.-Y. Wang, "Tone recognition for Chinese speech: a comparative study of Mandarin and Cantonese," in *ISCSLP 2004*, pp. 233–236.

[7] G. Peng and W.S.-Y. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines," *Speech Communication*, vol. 45, pp. 49–62, 2005.

[8] X. Lei, M. Siu, M.Y. Hwang, Ostendorf M., and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Interspeech 2006*, pp. 1237–1240.

[9] J.-C. Chen and J.-S. R. Jang, "TRUES: Tone recognition using extended segments," *ACM Transactions on Asian Language Information Processing*, vol. 7, no. 5, pp. 10:1–10:23, 2008.

[10] T. Ng, B. Zhang, K. Nguyen, and L. Nguyen, "Progress in the BBN 2007 Mandarin speech to text system," in *ICASSP 2008*, pp. 1537–1540.

[11] C. Ni, W. Liu, and B. Xu, "Using prosody to improve Mandarin automatic speech recognition," in *Interspeech 2010*, pp. 2690–2693.

[12] C. Liu, F. Ge, F. Pan, B. Dong, and Y. Yan, "A one-step tone recognition approach using MSD-HMM for continuous speech," in *Interspeech 2009*, pp. 3015–3018.

[13] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Interspeech 2011*, pp. 1589–1592.