Decision Tree Based Tone Modeling with Corrective Feedbacks for Automatic Mandarin Tone Assessment

Hsien-Cheng Liao[1], *Jiang-Chun Chen*[1], *Sen-Chia Chang*[1], *Ying-Hua Guan*[2], *Chin-Hui Lee*[3]

[1] Information and Communications Research Laboratories, Industrial Technology Research Institute
[2] Department of Applied Chinese Language and Literature, National Taiwan Normal University
[3] School of Electrical and Computer Engineering, Georgia Institute of Technology

{hcliao, jtchen0901, chang}@itri.org.tw, yhguan@ntnu.edu.tw, chl@ece.gatech.edu

Abstract

We propose a novel decision tree based approach to Mandarin tone assessment. In most conventional computer assisted pronunciation training (CAPT) scenarios a tone production template is prepared as a reference with only numeric scores as feedbacks for tone learning. In contrast decision trees trained with an annotated tone-balanced corpus make use of a collection of questions related to important cues in categories of tone production. By traversing the corresponding paths and nodes associated with a test utterance a sequence of corrective comments can be generated to guide the learner for potential improvement. Therefore a detailed pronunciation indication or a comparison between two paths can be provided to learners which are usually unavailable in score-based CAPT systems.

Index Terms: tone assessment, computer aided language learning, computer assisted pronunciation training, feedback

1. Introduction

Mandarin tones are usually the most difficult part for learners with non-tonal native languages. Due to the limited time in the class, teachers can hardly give the students enough time for practicing tone production. As Computer Aided Language Learning (CALL) systems are becoming increasingly popular in foreign language learning, it is possible to use a CALL system to assist students to learn Mandarin tone during off-class hours. Moreover, such system can also provide a way for self-study of Mandarin pronunciation.

Computer Assisted Pronunciation Training (CAPT) is an important part of CALL systems. Therefore many efforts for developing automatic assessment system to support teaching and providing adaptive pronunciation exercises for students have been observed [2, 6, 4]. Although various products are available to date in the market place, most of them only provide waveforms, spectrograms or pronunciation scores. This set of information is usually not enough for learners to correct their pronunciation errors [8]. Some research has been attempted recently to improve pronunciation assessment effectiveness [14-15]. A common way is to train classifiers, which can categorize different kind of errors and then provide feedbacks which are relevant to associated illustrations or instructions. This approach can enrich the information that learners would perceived, however, such classifiers are usually a black-box model resulting in difficulties in inspecting error features and providing useful feedbacks to the learners.

In this research, a CAPT framework based on a decision tree for automatic Mandarin tone assessment is proposed. We use the decision tree, which is often intuitive and instructive for interpreting errors, to model pitch characteristics of bad and good tone productions. The decision tree [7] is a flexible model that allows experts and teachers to further inspect key attributes of a bad tone production and then add an associated feedback on the decision tree's nodes, leafs and paths to learners. Therefore, feedbacks corresponding to the traversed path when identifying a good or bad tone production can be provided to the learners for potential improvement.

2. Problem Characterization

The NTNU corpus used in this research was collected from 16 students of Mandarin Chinese with levels of study from under 1 year to approximately 4 years, with eight each with English or Japanese as native languages. A set of 28800 utterances containing one to four syllables were recorded. After bad-quality recordings were excluded, a total of 28,782 utterances, containing 59399 syllables, remained. We only consider four tones in Mandarin, including tone1 (high flat), tone2 (low rising), tone3 (high low rising), tone4 (high falling). Tone5 (neutral tone) is commonly regarded as coming up with unstable feature patterns in tone production and is therefore ignored in this study. The distribution of four tones is listed in Table 1. The process of determining good or bad tones using rating data provided by six experts is discussed next. Clearly quite a few syllables were labeled as bad tones in Table 1.

Table 1. *Distribution of four tones in the NTNU corpus.*

	tone1	tone2	tone3	tone4
Total number	14504	15486	11602	17807
No. of bad tones	496	1307	1141	542

In order to make our assessment be determined on the basis of what people deem acceptable. The corpus needs to be annotated on the quality of tone production. The annotation was performed by 6 native Chinese language experts from National Taiwan Normal University using a web-based annotation interface. Each annotator makes a score on a 1-5 scale [1]. A rating of 5 indicates native-like tone production, and a rating of 1 indicates that tone had obvious errors. Inter-rater correlation was used to evaluate consistency of the annotation across speakers. In Table 2 we summarize the inter-rater correlation on all syllables. The average correlation between raters was 0.68, showing that the agreement is not accidental and acceptable [9] for the following analysis.

In order to emphasize the characteristics of syllables which most annotators marked as bad and excluded extraordinary subjective judgments, we simplified the annotation data as follow. A syllable would be taken as a bad tone production only if more than half of the annotators rate it as level 1 or 2. Otherwise, the syllable would be marked as good. These binary labels were then used in decision tree

training. Note that in Table 1, the number of bad tone productions is much smaller than the number of good ones. This is similar to the study by Peabody [10]. The reason could be that the annotators would only mark an error only when is a serious tonal error to be pointed out while student is learning Mandarin. Although such annotation policy would make the results more real, however it leads to a biased problem in model training. That might results in a high false rejection rate while detecting bad tone productions. In order to minimize the overall false rejection and false acceptance rates, an additional cost has to be applied on the biased class when training the decision tree. This is commonly seen in the optimization of classification problem [5], and will be discussed later.

Table 2. *Inter-rater correlation on syllable-level*

Rater ID	1	2	3	4	5	6
1	1	0.72	0.75	0.78	0.56	0.8
2	0.72	1	0.7	0.68	0.62	0.73
3	0.75	0.7	1	0.7	0.56	0.72
4	0.78	0.68	0.7	1	0.53	0.77
5	0.56	0.62	0.56	0.53	1	0.62
6	0.8	0.73	0.72	0.77	0.62	1

3. Tree-Based Tone Assessment

Comparing to other machine learning methods, decision tree can be easily interpreted by its attributes tested on each node with a simple Boolean logic. This advantage gives us a way to inspect resulted criteria of identifying bad tone productions after training with experts' annotations. A block diagram of the proposed approach is shown in Figure 1. In the training phase we calculate the tone features based on the acoustic cues and tone labeling. Then we can build our tone models by using decision tree with the C4.5 algorithm [11] according to the annotated tone corpus and extracted tone features. Considering the human knowledge about tone production, a set of detailed comments can be associated with the constructed decision tree. Finally to assess a test utterance, the decision tree then can be traversed and the corresponding feedbacks can then be retrieved for the learners according to the traversed result.

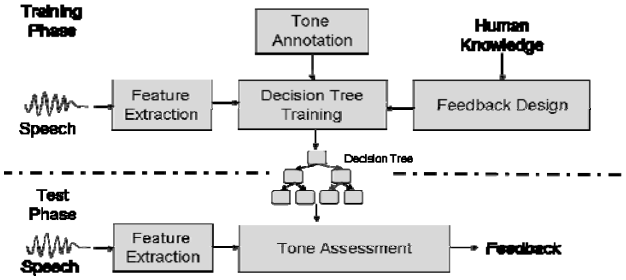

Figure 1: *A decision tree based tone assessment system block diagram with training (upper) and testing (lower) subsystems.*

3.1. Feature Extraction

Mandarin tones are usually distinguished by the shape its pitch contour. Other characteristics, such as amplitude and duration, can also be utilized. We therefore use pitch information as the primary feature to distinguish bad from good tone productions.

For extracting pitch information, we first use the RAPT algorithm [13] to extract F_0. Then we use a 5-point moving average filter to smooth the pitch contour. Due to differences in the mean F_0 of speakers, F_0 has to be further normalized across speakers to make meaning comparisons. The process is based on a method commonly used for Mandarin tone studies [16,12], i.e., with x being the observed raw pitch value, it is normalized according to the following formula:

$$p(x) = 4\frac{\log x - \log \mathrm{Min}}{\log \mathrm{Max} - \log \mathrm{Min}} + 1 \quad (1)$$

where Max and Min are the highest and lowest F_0 over all the syllables of each speaker after smoothing. This will put F_0 on a common 5-pt scale, which was originally proposed by [1].

After normalized into a 5-pt scale, Figure 2 shows the averaged pitch contours of the four tones labeled as good and bad. We can see that good and bad tone productions reveal significant different characteristics. For example, tone2 should be produced at a lower pitch register with a slope that becomes positive in the middle of the corresponding syllable. However, for bad tone productions we often observed a negative slope.

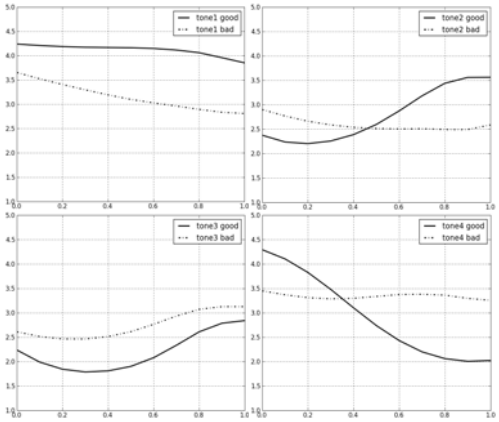

Figure 2: *Averaged pitch contours of good and bad tone productions*.

Features were extracted using an algorithm originally used in tone recognition [3], with a slight modification. Firstly all the utterances were segmented into syllables via forced alignment of Viterbi decoding based on a speaker-independent speech recognition engine. 707 syllables were removed from the recognition phase since their syllable durations are less than 40 msec or more than 600 msec, which are considered to be erroneous cases in forced alignment. Furthermore, the well-known tone sandhi rule applying to two consecutive tone3 syllables was taken care of by a manual transcription.

As shown in Figure 3, F_0 of each syllable was equally divided into three segments. We then adopt the mean value of F_0 of each segments and the differences between them as feature vectors of each syllable. This gives us a 6-dim feature vector for each segmented syllable.

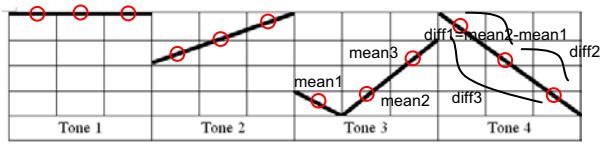

Figure 3: *Illustration of feature vector extraction.*

3.2. Decision Tree

Decision trees play a critical role in our proposed framework. In the classification phase the features representing pitch shape are tested at each node in a decision tree according to a pre-defined characteristic question about the tested segment. Then leaf nodes categorize each hypothesized syllable as good or bad tone production. On the other hand in the training and regression phase we select a sequence of questions to split the training samples into two parts at each node to effectively

minimize classification errors for training attributes. In this work, we used the C4.5 algorithm [11] for our decision tree training. It chooses an attribute with the highest normalized information gain to split its set of samples into subsets. It is a well-known algorithm in data mining and was proven to be effective on various classification problems.

However, since our annotations were extremely biased, C4.5 cannot minimize the false acceptance and false rejection rates at the same time. To overcome this problem, we need to assign different costs to different errors. A method called MetaCost [5] can be utilized to help us turn C4.5 into a cost-sensitive tool and the different cost setting would result in different error rates. For example, as shown in Figure 4, the amount of class B is increased after the tuned cost matrix is applied on the feature re-classification. The re-labeled features are then used to re-estimate the model parameter, therefore an improved model can be achieved for the biased data. We will discuss this issue in detail in the Experimental Results Section.

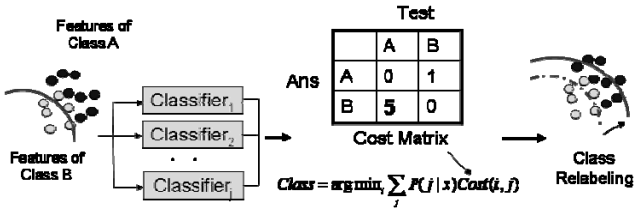

Figure 4: *The concept of the MetaCost approach.*

3.3. Feedback Labeling

After building tone models with decision tree training, we can inspect the attributes of each leaf node which classifies a syllable as a bad tone production. As shown in Figure 5, Path1, in solid arrow sequences, represents a bad tone2 production. According to the criteria on the traversed path, we can see samples classified to this leaf node have problems of an extraordinary high pitch level at the beginning and a negative slope during tone production. Therefore, we can potentially give corrective feedbacks, such as "you can lower your pitch in the beginning" or "you can raise your pitch gradually" on the associated nodes of Path1. By utilizing such advantage of the decision tree, we can further exam each node about the criterion used to split the node and then add some instructions for the learners about how to correct the error. Moreover, we can also summarize such errors in a systematic manner for each learner to make overall adjustments in tone production.

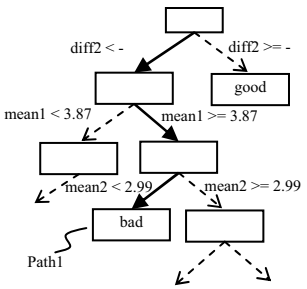

Figure 5: *Illustration of a path of a bad tone2 production.*

4. Experimental Results

To prepare for decision tree training, we divide the feature vectors into several right-context dependent (RCD) categories. Two RCD tones are different if their immediate right tones are different. For example, tone1 before tone2 in a word is defined as "tone1+tone2" in the RCD context, which is still a type of tone1. The number of RCD tone models is much larger than that of the context-independent tone models, thus a large corpus is required for reliable training of RCD tone models. For example, tone1 will have five models: model of tone1+tone1, tone1+tone2, tone1+tone3, tone1+tone4 and tone1+silence. Consequently, we have 20 RCD tone models in total. The whole training procedure was performed by using the NTNU corpus described in Section 2 and a well-known machine learning tool, Weka [17], to quickly build our tone models. A 10-fold cross validation is conducted to validate the performance. To target the effectiveness of tone assessment, the performance metric for model learning is defined as the summation of false rejection (FR) and false acceptance (FA) rates, which is consistent with the situation in pronunciation learning. To alleviate the data biased problem of tone modeling, a heuristic approach is experimented on tuning the cost weight of the biased class. Particularly in this experiment only the cost of the bad tone class is tuned, which is absolutely minor class compared with the good tone class.

As shown in Figure 6, considering all RCD tone models context-independently, equal weighting leads to poor performance when the cost equals 1, and in average 30% to 65% error reduction could be achieved for four tone models when the cost is increased for bad tone data.

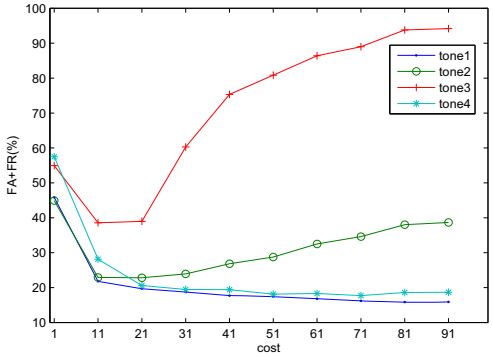

Figure 6: *The performance result when the cost parameter is heuristically tested for the minor data of bad tone.*

Moreover, to verify the reliability of tone modeling for good and bad tone productions, the detailed results of 20 RCD tone models are listed in Table 3. Instead of using the FA+FR, in Table 3 the good-bad binary classification rate is used, where the decision result is considered to be correct if it is consistent with the annotated result of human expert. In summary tone modeling of decision trees shows consistent results when considering the production difficulties of the four tones. In particular, tone3 shows a relatively low accuracy rate, which is taken as the result of highly variable pitch contour so that the overall characteristics could not be modeled correctly. The set of results in Table 3 demonstrates the reliability of RCD tone recognition, which is comparable when compared with related studies (e.g, [14]).

Table 3. *Assessment result of 20 RCD tone models in the NTNU corpus.*

	tone1	tone2	tone3	tone4
*+tone1	95.81	81.61	87.25	92.34
*+tone2	92.42	87.33	85.60	94.40
*+tone3	91.94	92.02	70.49	92.18
*+tone4	93.27	83.50	88.03	89.50
*+sil	95.83	93.78	85.78	91.37
Average	94.26	89.54	85.46	91.34

5. Discussion

In this study, we focus on the corrective feedback mechanism offered in the proposed decision tree based tone assessment framework. The traversed path and nodes of the decision tree are regarded to be an important cue to hint the category of tone production type, as described in Section 3.3. Moreover, a comparison between two paths can be provided to learners. For example, two traversed paths for tone2, identified in solid Path1 and bold Path2, are shown in Figure 7. In the first tone production, four kinds of error type can be found on Path1, and one or more feedbacks can be generated accordingly. For instance, "the tail of tone2 is not raised enough" (according to the feature "dif2<-0.07") or "the leading head of tone2 is a little too-high" (according to the question "mean1>=3.87") are both reasonable. Furthermore, after the second production of tone2, the comparison between Path1 and Path2 can easily be made. For instance, "you improved the mistake that the raising degree of tone2 is not enough" (according to "diff3<-0.61") or "you made a new mistake that the average pitch of tone2 is too low" or the combination of both comments are all possible.

In other words, even two tone productions are both classified into the 'bad' category, somehow the reason behind such mistakes are different. By using the decision tree, the difference of the two potential paths prompts some corrective suggestions, such as the unchanged errors, the new errors and the improved errors. Therefore our system can easily provide the feedback for one tone particular production or even for the comparison between any two productions.

Conventionally, the rule-based descriptions about tone modeling, such like [1], is too primitive when compared with the modern statistical approaches. However for tone assessment, simple rules seem to be more practically useful when specific feedbacks are required to correct wrong production of tones. The constructed decision tree organizes the rules considering the statistical C4.5 algorithm and the annotated tone corpus. It is therefore beneficial to the qualitative and quantitative characterization of detecting problematic tone productions.

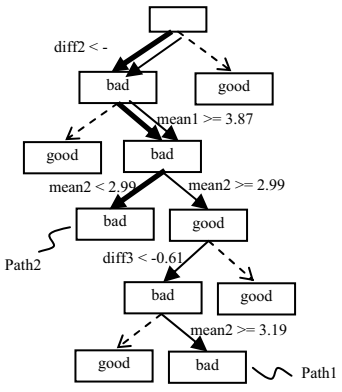

Figure 7: *Illustration of relative comparison of two bad tone2 productions.*

6. Conclusions

In this paper, a decision tree based tone assessment framework is proposed. It also provides corrective feedbacks to the user. By using the path traversed in the decision tree, our system recommends descriptive correction for tone production. Furthermore, based on the decision tree, the assessment system can reduce the large amount of human labor when compared with the preparation of course material in the traditional template-based CAPT systems. Finally, the proposed approach is highly flexible, which is not only works well for tone assessment, but also applicable to other kinds of assessment like pronunciation, prosody, duration and energy contours when learning many different kind of languages. In the immediately future work, the pedagogical and practical evaluation would be conducted with the education researcher in order to prove the effectiveness of the approach.

7. Acknowledgements

This study is partially supported by Project 9352MD3100 and conducted at ITRI under the sponsorship of the Ministry of Economic Affairs, Taiwan.

8. References

[1] Chao, Y. R., *Mandarin Primer*. Cambridge: Harvard University Press, 1948.

[2] Chen, J. C., Jang, J. S. and Tsai, T. L., "Automatic pronunciation assessment for Mandarin Chinese: approaches and system overview", *Computational Linguistics and Chinese Language Processing*, 12(4):443-458, 2007.

[3] Chen, S. H. and Wang, Y.R., "Tone Recognition of Continuous Mandarin Speech Based on Neural Networks", *IEEE Transactions on Speech and Audio Processing*, 3(2):146-150, 1995.

[4] Cincarek, T., Gruhn, R., Hacker, C., Nöth, E. and Nakamura, S., "Automatic pronunciation scoring of words and sentences independent from the non-native's first language", *Computer Speech & Language*, 23(1), 65-88, 2009.

[5] Domingos, P., MetaCost: A General Method for Making Classifiers Cost-Sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining* (KDD-99). 1999.

[6] Franco, H., Neumeyer, L. and Kim, Y., "Automatic Pronunciation Scoring for Language Instruction", *Proc. ICASSP*, pp.1471-1474, 1997.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth, Inc., Elmont, California, 1984.

[8] Neri, A., Cucchiarini, C. and StrikFeedback, H., "Computer Assisted Pronunciation Training: technology push or demand pull?" *Proceedings of ICSLP* 2002, Denver, USA, pp. 1209-1212.

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

[10] Peabody, M. and Seneff, S., "Annotation and Features of Non-native Mandarin Tone Quality", *Proc. Interspeech*, Brighton, UK, September 2009.

[11] Quinlan, J. R., C4.5: *Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[12] Rose, P., "Considerations in the normalisation of the fundamental frequency of linguistic tone", *Speech Communication*, 6(4):343–352, 1987.

[13] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", in W. B. Kleijn and K. K. Paliwal [Ed], *Speech Coding and Synthesis*, 495-518, Elsevier Science, 1995.

[14] Tang, L. and Yin, J. X., "Objective Evaluation of Putonghua Tones", *Journal of Chinese Information Processing*, 21(6), 116-124, 2007.

[15] Tsubota, Y., Kawahara, T. and Dantsuji, M., "Computer-assisted English vowel learning system for Japanese speakers using cross language formant structures", *Proc. ICSLP* 2000.

[16] Wang, Y., Jongman, A., and Sereno, J. A., "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training", *Journal of the Acoustical Society of America*, 113(2):1033-1043, 2003.

[17] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)*, Morgan Kaufmann Publishers, 2005.