Evaluating Prosody of Mandarin Speech for Language Learning

Minghui Dong, Haizhou Li, Tin Lay Nwe

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

{mhdong,hli,tlnma}@i2r.a-star.edu.sg

Abstract

This paper proposes an approach to automatically evaluate the prosody of Chinese Mandarin speech for language learning. In this approach, we grade the appropriateness of prosody of speech units according to a model speech corpus from a teacher's voice. To this end, we build two models, which are the prosody model and the scoring model. The prosody model that is built from the teacher's speech predicts the reference prosody for the learning text. The scoring model compares the student's prosody with the reference prosody and gives a prosody rating score. Both the prosody model and the scoring model are built using regression tree. To make the two prosodies comparable, we transform the student's prosody into the teacher's prosody space. To build the scoring model, we derive from the corpus a reference data set, in which prosody rating is associated with prosody parameters. During speech evaluation, the student's prosody is first transformed into the teacher's prosody space and then evaluated by the scoring model. Experiments show that our model works well for speech of new speakers.

Index Terms: language learning, prosody evaluation

1. Introduction

Speech recognition technology has been applied in many ways to help understand human speech. Computer Assisted Language Learning (CALL) [1] [2] systems help language learners by identifying the errors in their speech, and thus improving their spoken language skills. One might find the language learning applications in two main areas: (1) identifying the pronunciation flaws in speech so as to help the speaker produce the correct sound. (2) identifying the prosody of speech so as to help the speaker speak more naturally.

Speech usually contains two types of information, segmental information and prosody information. Segmental information usually refers to the phonetic content as to what a speaker says, while prosody usually refers to how a speaker says. The basic sound of speech is determined by segmental information, while the naturalness of speech is usually determined by the prosody. Perceptually, prosody mainly refers to speech properties such as time length, pitch, loudness, intonation, breaks, rhythm, tones, etc. Acoustically, prosody exists in the form of duration, fundamental frequency contour and energy of speech units. In learning to speak like a native, a student needs not only to read each sound correctly but also to imitate the teacher's prosody.

There were some attempts [3-5] to teach intonation by comparing student productions to a target contour. However, the problem of evaluating prosody has not been fully solved. It is desirable to have a system capable of evaluating the prosody of students' speech and giving a score in computer-aided interactive language learning. Building such a system with statistical approach, one needs a reference database, which ideally consists of a standard speech corpus reflecting the desired prosody; and a substandard speech corpus that reflects possible erroneous prosody variations. It might be easy to collect a standard speech corpus from a teacher's voice. However, it is not so straightforward to collect a substandard speech corpus. One possible way is to collect speech samples from a large number of language learners [6]. However, even if it is possible to collect a large quantity of speech samples of language learners, the coverage of possible prosody variations may not be sufficient. What we expect from a substandard speech corpus is the statistics that reflect correlation between prosody rating and prosody parameters. To circumvent the need of such a data collection of substandard speech corpus, we generate the prosody rating reference data set from the standard speech corpus. This is done by relating difference between prosody parameters of two units to the difference between their linguistic features according to human perception experiences.

The paper is organized as follows. In section 2 we present the methodology, covering the system architecture, the prosody model, and the scoring model. In section 3, we report the experiments. Finally we conclude in section 4.

2. Methodology

2.1. Framework

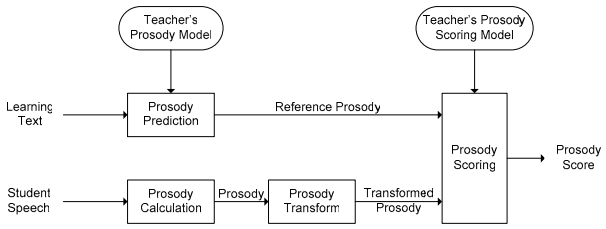

Figure 1 *System framework for prosody scoring*

We collect a standard speech corpus from a professional speaker, also referred to as the teacher's voice. The evaluation of the prosody is supported by a prosody model and a prosody scoring model built from this speech corpus.

Figure 1 is a diagram of the prosody scoring system. Suppose that the learner is expected to read a given paragraph. We process the speech uttered by the language learner and the corresponding text. From the student's speech, we derive the prosody parameters of each unit. From the text, we predict the expected prosody parameters using the prosody model. This prosody serves as the reference prosody, which indicates how

the teacher would read the text. Note that there exist intrinsic differences between speakers, such as pitch level. It is desirable to normalize the speaker's effect before a fair comparison can take place. Therefore, we use a transform to map the student's prosody into teacher's prosody space. The prosody scoring system reports the difference between the reference and the observed student's prosody parameters.

2.2. Reference Prosody

In Figure 1, the system produces reference prosody and observed prosody for comparison. We begin by describing the prediction of reference prosody that determines the desired prosody given the text script.

2.2.1. Parameters for Prosody Scoring

Prosody can be broadly summarized as duration, pitch contour, and energy of speech. Based on the application needs, there may be different ways to define the prosody parameters. In this work, we define the prosody to address two important properties of Chinese speech, the tone and the rhythm.

Chinese is a tonal language, in which each character carries a tone. Tones exhibit as patterns of pitch contour from the acoustic point of view. Rhythm exists as prosodic unit groups. Researches have found the existence of prosodic word [7], which is a phenomenon that speech units are usually grouped to small prosodic units, normally consisting of 2-3 syllables. At acoustic level, prosodic word boundary is usually presented as duration, pitch change, and energy change. Experience in Chinese TTS system has found that tone and prosodic word groups affect the naturalness of Chinese speech very much. Therefore, the defined parameters should address the two important aspects of Chinese speech.

We define parameters to describe duration, pitch level, pitch range, pitch contour shape, energy distribution, etc. Totally, 40 prospective parameters are defined. However, there is redundancy among these parameters because many of them are highly correlated. To reduce redundancy, we cluster the parameters into groups [8]. The distance between parameters is calculated based on correlation value between two parameters in the corpus.

The clustering process built a tree structure of the parameters. At last, considering the similarity level and acoustic meanings of the parameters, we decided to keep 12 clusters, from each of which, a representative is selected. The selected prosody parameters for Chinese syllable unit are:

- Duration of the syllable (Duration)
- Durations of initial part and final part of the syllable (InitDuration, FinalDuration)
- Pitch mean and pitch range (PitchMean, PitchRange)
- Start, middle and end points from pitch contour (PStart, PMid, PEnd)
- Position that divides energy into half (EnergyCenter)
- RMS energy of the whole syllable (Energy)
- RMS energy (with frame of 50ms) of the start and end points of the syllable (StartEnergy, EndEnergy)

2.2.2. Prosody Model

The prosody model is built to predict the prosody parameters from the text input, which is characterized by linguistic parameters. In this work, prediction of each parameter is done using the CART decision tree [9] where a regression tree is built for each of the parameters. In this way, the prediction of the parameter is considered as a classification problem that assigns an input feature to one of the leaf nodes of the regression tree. For each leaf node, we can calculate the mean and standard deviation of this class. This is calculated by considering the samples of training data that fall into this class.

The prosody model $\lambda(\cdot)$ can be defined as follows:

$$G = \lambda(F) \tag{1}$$

where F are the linguistic features, $G = \{P, S\}$ are the prosodic parameters with P and S being the mean and standard deviation of the prosody parameters. The linguistic features are derived from the text script while the prosodic parameters are derived from the speech signal. F, P and S are in the form of vectors. The linguistic feature vector F consists of the following:

- Pronunciation of the current syllable (initial, final and tone)
- Pronunciation of previous syllable (initial, final and tone)
- Pronunciation of next syllable (initial, final and tone)
- Prosodic boundary type before the syllable (whether this syllable is a start syllable of a prosodic word)
- Prosodic boundary type after the syllable (whether this syllable is a end syllable of a prosodic word)

The prosody parameters are defined for each syllable unit. As a prosody event is defined in a context, the parameters also reflect prosody information beyond a syllable. For example, the pitch mean value describes the general pitch information in the whole utterance.

2.3. Prosody Scoring

We next derive prosody parameters from the student's speech, transform it to teacher's prosodic space and compare it with the predicted prosody parameters to arrive at a prosody rating.

2.3.1. Prosody Calculation

As our prosody rating will be calculated at unit level, before any other calculation, we need to identify the start and end of each unit in the speech utterance. The segmentation of speech unit is done by a forced-alignment between the input speech and the text script. An HMM-based Chinese Mandarin speech recognizer is used for this purpose. After the forced-alignment, we derive the parameters of each unit.

2.3.2. Prosody Transformation

We will compare the prosody of the student's speech with the reference prosody predicted from the teacher's prosody model. We expect that the student would follow the teacher's speech as close as possible. To establish sound comparison between the two sets of parameters, we first transform student's prosody parameters into the teacher's prosody space. The transformation is to normalize the speaker effect such as pitch, energy, etc.

Suppose the prosody parameter vector calculated from speech is $P^s = (p_1^s, p_2^s, ..., p_n^s)$, and the prosody parameter vector after transformation is $P^t = (p_1^t, p_2^t, ..., p_n^t)$. The transformation is done as follows:

$$p_i^t = a_i p_i^s + b_i \quad (2)$$

where p_i^s is the prosody parameter calculated from the student's speech, p_i^t is the prosody parameter in the teacher's prosody space, a_i and b_i are regression parameters for the *i*-th prosody parameter.

We estimate a_i and b_i using linear regression estimation from training samples of the intended student (p_i^t will be the predicted reference prosody parameter during linear regression estimation). This can be seen as a calibration step, which is achieved by using the first a few utterances from the student. The assumption here is that the student has pronounced phonetically as expected in the text script, and the prosody of most of the syllables is correct. Once we obtain the regression parameters a_i and b_i, the prosody parameters of the test utterances will be transformed to the teacher's space for prosody scoring.

2.3.3. Prosody Scoring Model

We attempt to evaluate the quality of prosody based on the difference between the observed prosody (after transformation) and the reference prosody. The scoring model is defined as:

$$q = \gamma(D) \quad (3)$$

where *q* is the prosody rating score, *D* is the normalized prosody difference vector:

$$D = (P^t - P^p)/S^p \quad (4)$$

where P^t, P^p and S^p are the observed prosody vector, the reference prosody vector, and the reference standard deviation vector, respectively. The scoring model $\gamma(\cdot)$ can be implemented using a regression tree.

2.3.4. Training Prosody Scoring Model

To train the scoring model $\gamma(\cdot)$, we first build a prosody rating reference data set, in which the prosody rating is associated with the quantifiable prosody difference.

Suppose unit *x* and unit *y* are two speech units from utterances *X* and *Y* from the standard speech corpus. They share the same sound (phonetically equivalent), but come from different contexts (with different prosody). When we replace *y* with *x* in utterance *Y*, the prosody naturalness of utterance *Y* is degraded. But how much the prosody quality is degraded? It is not straightforward to associate the prosody difference between *x* and *y* with the prosody rating *q*. However, note that we can easily associate prosody rating *q* with linguistic feature difference, $q \sim Q(F_x, F_y)$, according to human perception experience. As defined in Eq.(1), we know that there is a correspondence between the prosody parameters P and the linguistic feature F. Therefore, one can infer the association between the prosody rating and the prosody parameters, that is $q \sim Q'(P_x, P_y)$, from $q \sim Q(F_x, F_y)$.

Inspired by this idea, we create our training statistics from the standard speech corpus by permutation of linguistic feature of the speech units. For a given unit *s* in the standard speech corpus, with linguistic feature vector *F* (which is derived from text) and prosody parameter vector *P* (which is derived from speech), we change *F* to F'. The association $q \sim Q(F, F')$ of such a permutation can be formulated as:

$$q(s, F') = Q(F, F') \quad (5)$$

where $Q(\cdot,\cdot)$ defines some rules that convert the differences between linguistic feature *F* and *F'* into a score representing quality degradation of prosody according human perceptual experience.

At the same time, we calculate the normalized prosody difference between *P* and P':

$$D(s, F') = (P - P')/S' \quad (6)$$

where P' and S' are the reference prosody and the standard deviation. In this way, we associate the normalized prosody difference with a prosody rating. We record $< D(s, F'), q(s, F') >$ as a data item for each parameter permutation. By varying F' for each unit *s*, we are able to generate as many data items as needed in the prosody rating reference data set.

The parameter permutation F' is carried out to simulate possible prosody variations, which include tones of the syllables and prosodic boundary types as listed in Section 2.2. By altering the tone or prosodic word boundary type (binary value: yes/no), we are able to generate a new context for F' to produce a data item. The scoring model $\gamma(\cdot)$ can then be trained with the prosody rating reference data set.

3. Experiments

We used the following two corpora. Both of them were manually labeled with the syllable start and end points.

Corpus A: This corpus consists of about 155,000 syllables in 20,000 Chinese Mandarin utterances. Each utterance consists of 5-15 syllables. The speech was read by a professional female broadcast announcer. The script of the corpus was designed to cover Chinese syllable as many as possible with a greedy algorithm. This corpus is used as the teacher's corpus. It consists of two parts. The first 16,000 utterances are used as the training set to build the prosody model and the scoring model. The rest 4,000 utterances are used as the testing set.

Corpus B: This corpus consists of about 12,000 syllables in 800 Chinese Mandarin utterances read by 40 speakers. In each of the utterance, one to three of the units are not well pronounced. Totally, there are about 1,300 syllables with improper prosody (incorrect tone, unclear tone, improper prosodic break within and between words, etc). We labeled the units using a scale from 0 to 4 (0 for the best, 4 for the worst) by prosody appropriateness by human listening. This corpus is used for testing the prosody scoring model.

3.1. Prosody Model

In this experiment, we evaluate the performance of the prosody model with Corpus A. The prosody parameter vector and linguistic feature vector for unit are defined as described in section 2.2. First we derive the prosody parameters and linguistic features for all the units in the corpus. Then we train

prosody model (Eq. 1). A regression tree is trained for each parameter with the CART approach on training data. Finally we test the model using the testing set. The RMSE (root mean square error) and correlation values (between the predicted value and the actual value) of the parameters are as shown in Table 1. In Table 1, we find that the correlation coefficients are at the range of 0.61 to 0.83. This shows that the prosody model works reasonably well. The reason for large variations of the parameters is that prosody of speech is affected by many factors, some of which cannot be predicted from the text. This explains why we need to normalize the parameters in Eq.(4).

Table 1 *Result of prosody parameter prediction*

Parameter	RMSE	Correlation
Duration	0.045 sec	0.701
InitDuration	0.019 sec	0.681
FinalDuration	0.040 sec	0.695
PitchMean	33.19 Hz	0.829
PitchRange	37.71 Hz	0.624
PStart	19.59 Hz	0.784
PMid	9.37 Hz	0.611
PEnd	18.76 Hz	0.819
EnergyCenter	0.090	0.740
Energy	697.5	0.681
StartEnergy	552.2	0.677
EndEnergy	550.5	0.681

3.2. Scoring Model

In this experiment, we evaluate the effectiveness of the scoring model on the standard speech. To this end, we first generate the prosody rating reference data set for scoring model training. Each data item is labeled with a level of prosody appropriateness on a scale from 0 to 4 based on the following two rules:

1) If the tone of the syllable is different from the reference syllable, the penalty is 2.
2) If the prosodic word boundary type before (or after) the syllable is different, the penalty is 1.

To arrive at a balanced data set, we generate roughly equal number of samples for each prosody rating level, resulting in two data sets: training set that consists of about 762,000 items; testing set that consists of about 198,700 items.

The prosody scoring model is trained with the training set. Please note that all the prosody levels are discrete values in the training data because they are assigned by rules. However, the trained scoring model outputs a continuous value.

We test our scoring model with the testing set. The RMSE of predicted prosody rating is 0.79. The correlation of predicted score with the original score is 0.75. This shows the model generated with the training speech works for the testing speech of the same speaker.

3.3. Scoring for New Speakers

To test whether the scoring model trained with teacher's voice works for speech utterances of language learners, we test our model on a different speech corpus. Corpus B consists of speech units labeled with prosody ratings. We chose 200 units from each level and totally 1,000 units from the corpus for our testing. Using our method to score the units, we achieved a correlation value of 0.71. This shows we are able to achieve similar result when scoring speech utterances from new speakers.

4. Conclusion

We proposed an approach to automatically evaluate prosody of Chinese Mandarin speech units for language learning. We also proposed to construct a prosody rating reference data set, which is labeled in terms of prosodic appropriateness according to human perception knowledge. We built a prosody model and a prosody scoring model on the basis of a teacher's corpus, and then used it to evaluate utterances of new speakers. Experiments have reconfirmed our ideas, and have found that, with one teacher's model, we are able to effectively evaluate new speakers' voices.

In the future, we will improve our method in several aspects: (1) We will try to include more linguistic factors that affect prosody and refine our scoring scheme when generating the reference data set; (2) We will try to include more prosodic events in our work. We will try to apply the method to identify different types of prosody flaws, thus giving language learner more specific instructions for improvement; (3) We will improve the mapping mechanism from student's prosody space to teacher's prosody space.

5. References

[1] A. Neri, C. Cucchiarini, H. Strik, and L. Boves "The pedagogy-technology interface in Computer Assisted Pronunciation Training", Computer Assisted Language Learning 15, 441-447, 2002.

[2] R. Hincks, "Speech recognition for language teaching and evaluating: A study of existing commercial products", Proceedings of ICSLP, 733-736, 2002

[3] Matthias Jilka, Testing the Contribution of Prosody to the Perception of Foreign Accent. Proceedings of New Sounds (4th International Symposium on the Acquisition of Second Language Speech), Amsterdam, pp. 199 – 207, 2000.

[4] Spaai, G. & Hermes, D. A visual display for teaching intonation. CALICO Journal 10(3), 19-30, 1993.

[5] Chun, D. "Teaching Tone and Intonation with Microcomputers." CALICO Journal 6, 3, 21-47, 1989.

[6] Carlos Teixeira, Horacio Franco, Elizabeth Shriberg, Kristin Precoda, Kemal Sönmez, "Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners", ICSLP 2000, Beijing, China.

[7] Minghui Dong, Kim-Teng Lua and Haizhou Li, "A Probabilistic Approach to Prosodic Word Prediction for Mandarin Chinese TTS", InterSpeech 2005, Lisbon, Portugal.

[8] Minghui Dong, Kim-Teng Lua, Jun Xu, "Selecting Prosody Parameters for Unit Selection-based Chinese TTS". IJCNLP 2004, Sanya, China.

[9] Breiman, L.; Friedman, J.; Olshen, R. and Stone, C. Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove, CA., 1984.