# CHINESE INTONATION ASSESSMENT USING SEV FEATURES

*Dengfeng KE, Bo XU*

Institute of Automation, Chinese Academy of Sciences

## ABSTRACT

Intonation assessment is an important part of Chinese CALL system. Nowadays, most systems use the correlation and RMSE features to assess the quality of the intonation of a given speech. As correlation and RMSE assign unoptimized weights to different degrees of mismatching errors, they may lead to performance degradation. In this paper, we propose a new feature called Sorted Error Vector (SEV) for intonation assessment. The basic idea is to calculate mismatching quantities, sort them with ascending order, and then re-sample them to a K-points vector. This feature has four benefits: first, it is text-length independent; second, weights are let to train by classifiers; third, the relationship between the errors and the final results is not limited to any assumption; fourth, SEV is not sensitive to the performance of different pitch extracting algorithms. Experiments show that no matter in which case, SEV feature performs the best.

***Index Terms* —** Intonation Assessment, Intonation Evaluation, Intonation feature, Sorted Error Vector, SEV

## 1. INTRODUCTION

Generally, intonation refers to the variations in the pitch of a speaker's voice used to convey or alter meaning. But in its broader and more popular sense it is used to cover much the same field as 'prosody', where variations in such things as voice quality, tempo and loudness are included. In this paper, intonation refers to its narrow definition, which does not contain voice quality, tempo and loudness.

Looking through the history of intonation assessment, the assessment method can be separated into comparative methods[3][4][6][8][9] and recognition based methods[5][7]. In the beginning, intonation assessment is only used as an important part of assessing the performance of a synthesis system. In that period, intonation assessments were carried on by human beings [1][2]. Hermes [3] firstly used the mean distance, the root-mean-square (RMS) distance, and the correlation coefficient for objective intonation assessment. And after that, people started using root-mean-square-error (RMSE) and correlation features for intonation assessment until recently [4][6][8][9]. Although Jia [8][9] proposed new methods called "optimal similarity" and "weighted similarity", both of them are calculated from the correlation features of each prosodic components. In fact, they are correlation-based features.

In this paper, we proposed a new feature called the Sorted Error Vector (SEV) for comparative intonation assessment in computer assisted language learning (CALL) systems. In our system, the language learners are forced to imitate a given speech, the more similar, the better. The task is to assess the learner's level of intonation. In fact, intonation assessment technique developed in synthesis field can be used in this case. Since teacher's speech is also given, there is no need to use a recognition based assessment approach. Comparative method is the most suitable choice.

In this paper, the following sections are organized like this: In section 2, the intonation features will be introduced. After analyzing the shortcomings of traditional features for Chinese intonation assessment, a new feature for intonation assessment is proposed. In section 3, the database is detailed for its annotation and experimental results of our new feature are compared with other features and discussions are made on them. Finally, in section 4, draw to the conclusion that this new feature outperforms other features and further direction for intonation assessment study is pointed out.

## 2. INTONATION FEATURES

In CALL system, the task of intonation assessment is to assess how similar the learner's pitch is to the teacher's pitch. Figure 1 is the extraction flowchart of intonation features.

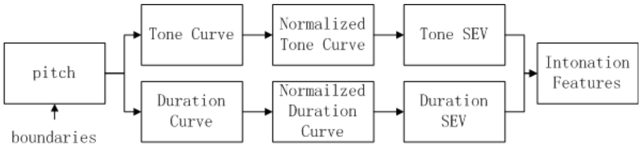

**Figure 1: flowchart of extracting intonation features**

Firstly, the length of the learner's pitch is quite different to the teacher's, not only within the whole sentence but also within all the syllables of the sentence. Boundary alignment technique is needed to align the phone boundaries of the learner's speech to the teacher's.

After force alignment to the speech, the original pitch curve is split into two things: the tone curve and the duration curve. The tone curve is the curve concatenated with all the sub-curves of each tone in the sentence. The

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

sub-curve of a tone is extracted from the final part of a syllable, because former studies show that the final part carries most tonal information. Each sub-curve is re-sampled from the original pitch of the finals into 10 points. Then we get a one-to-one mapping of the pitch from the learner to the teacher. We also get a one-to-one mapping of the durations of all syllables.

After this, the tone curve is normalized to deal with pitch range problem and the duration curve is normalized to deal with rate-of-speech problem.

Finally, SEV features are extracted from tone curve and duration curved respectively.

### 2.1. Intonation Similarity Features

As the lengths of the tone curve and the duration curve depend on the length of the text, text-length independent features are needed to assess the similarity of the learner's intonation. Traditionally, correlation and root mean square error (RMSE) are used for this purpose. The greater the correlation is, the more similar the learner's pitch is to the teacher's. Reversely, the less the RMSE is, the more similar the learner's pitch is to the teacher's.

Correlation is a good idea to compare the similarity of two curves. It is defined as formula (1). Where $N$ is the number of samples, $DX$ and $DY$ stand for the variances of X and Y respectively, $\overline{X}$ and $\overline{Y}$ stand for the means of X and Y respectively. But from this definition, we can learn that it may cause some problem if all the tones in the text are high tones of standard Chinese.

There are four normal tones in Chinese: the high tone, the rising tone, the low tone and the falling tone. They can be annotated in IPA as 55, 35, 21 (or 214) and 51 respectively. Although the high tone is 55, it is not strictly flat. It may be slightly downshifting, or slightly rising. Correlation feature may lead to negative value when comparing pitch slightly downshifting to the one slightly rising when the sentence consists of only high tones. The other problem is that if the pitch is not 100% right of the original speech, for example, halving or doubling problem exists, correlation may lead to great change.

RMSE is another good idea to compare the similarity of two curves. It is defined as formula (2). It can avoid the shortcomings of correlation when comparing two curves of all high tones. But it assigns different degrees of mismatch with equal weights. Also, it assumes that the final hearing result is a square root of a quadratic representation of all mismatches. But maybe linear, cubic or other representation would be better.

$$COR(X,Y) = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{DX \bullet DY}} \tag{1}$$

$$RMSE(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2} \tag{2}$$

We think that different degrees of mismatch must have different weights. Formula (3) shows the basic idea of the final mismatching result of hearing. Where $D(X_i, Y_i)$ is the mismatch error quantity from $X_i$ to $Y_i$. It may be the Euclidean distance or other suitable quantities. $f(\bullet)$ is a transformation function to map mathematical distances to hearing results. It may be linear or non-linear. $w_i$ is the weight for the i$^{th}$ mismatching error quantity.

$$Hear(X,Y) = \sum_{i=1}^{N} w_i f(D(X_i, Y_i)) \quad where \quad \sum_{i=1}^{N} w_i = 1 \tag{3}$$

By denoting $w_i = \frac{1}{N}$, $D(X,Y) = \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{DX \bullet DY}}$ and $f(X) = X$ formula (3) becomes COR(X,Y). By denoting $w_i = \frac{1}{N}$, $D(X,Y) = | X_i - Y_i |$ and $f(X) = X^2$, formula (3) becomes RMSE(X,Y)$^2$.

### 2.2. Sorted Error Vector

It is difficult to solve formula (3). Let us denote the set of mismatching error quantities as $\{D(X_i, Y_i)\}$. By arranging them in ascending order, it becomes $\{z_i\}$ where $z_i \le z_j$ *iff* $i < j$. Then formula (3) becomes formula (4). Re-sampling z from N-points to K-points, it becomes formula (5). If K is large enough, formula (5) will produce results that are accurate enough to approach formula (4). Noting that N depends on the length of the text but K is a constant, it is much easier to optimize $w_k$ and $f(z_k^{'})$ with certain algorithm.

$$Hear(X,Y) = \sum_{i=1}^{N} w_i f(z_i) \quad \text{where} \quad \sum_{i=1}^{N} w_i = 1 \tag{4}$$

$$Hear(X,Y) = \sum_{k=1}^{K} w_k f(z_k^{'}) \quad \text{where} \quad \sum_{k=1}^{K} w_k = 1 \tag{5}$$

The vector $(z_1^{'}, z_2^{'}, ..., z_K^{'})$ is extracted from the sorted error quantities, so it is named as Sorted Error Vector (SEV) in this paper. The similarity of two curves can be implied by this vector.

Take tone curve for examples, figure 2 illustrate how to extract SEV feature of the tone curve. At first the learner's pitch and teacher's pitch are converted into semi-tone scale, so that it is linear to human hearings. $D(X,Y)$ can be defined as $| X_i - Y_i |$. The pitch is re-sampled to 10-points for each final. Then, the same as RMSE and correlation method, both the learner's and the teacher's tone curves are further normalized to zero-mean. By calculating the absolute differences at each point, sorting them in ascending order, and re-sampled into K-points, the SEV of the tone curve is finally fetched as shown in (f) of figure 2.

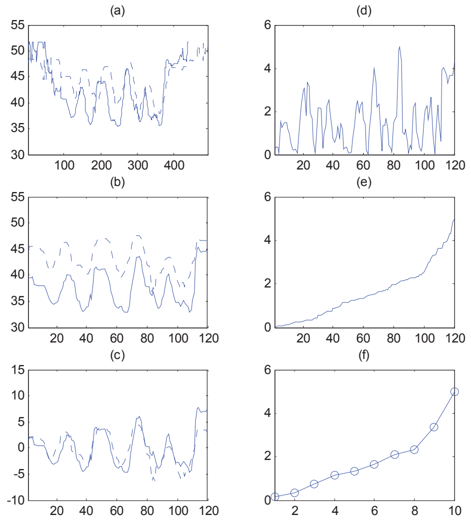

**Figure 2: illustration of SEV extraction**

(a) original pitches of the teacher(solid) and the learner(dot)
(b) pitches normalized by initial/final boundaries from (a)
(c) further normalized to zero-mean
(d) distances at each point of two curves from (c)
(e) distance sorted from (d)
(f) re-sampled to K=10 points from (e)

The SEV of the duration curve is extracted quite similar to that of the tone curve. The only difference is the duration normalization procedure. All the durations of each syllable are normalized by dividing the mean of all durations.

## 3. EXPERIMENTAL EVALUATION

### 3.1. Database Setup

The intonation database consists of one male teacher, one female teacher, 10 male learners and 10 female learners. Each person consists of 136 sentences with different length, from 4 characters to 25 characters, covering 16 kinds of intonations. Each learner's intonation is leveled by comparing to the corresponding teacher's intonation.

Intonation is very hard to annotate. It is intangible to define strictly the detail description for each level. So we use a 3-levels scoring method. The levels of the learners' intonation are defined as "good", "normal" and "bad" by comparing to their teacher's speech. Two native speakers of standard Chinese are ask to annotate the database again and again. In each round, the order the sentences are permuted randomly to avoid scoring memory.

Table 1 shows the consistency rates of two annotators at each round. It is clear that they go to further consensus at each round. In the first round, the annotating results are very poor, only 54.2% of the sentences are annotated with the same level by both annotators. But in the fourth round, their agreement comes to 81.1%. That is because they are getting more and more aware of the overall database.

**Table 1: consistency rates at each round**

| Round | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Rate | 54.2% | 60.9% | 74.8% | 81.1% |

**Table 2: consistency rates of neighboring rounds**

| Round vs Round | 1 vs 2 | 2 vs 3 | 3 vs 4 |
|---|---|---|---|
| Annotator A | 66.7% | 80.9% | 83.2% |
| Annotator B | 54.1% | 67.9% | 78.9% |

Table 2 shows the consistency rates of each neighboring rounds of the two annotators. Take annotator A for examples, the second round has an consistency rate of 66.7% with the first round, but after the third round, the agreement with the second round becomes 80.9%, and the fourth round makes limited progress to the third one. Annotator B's results are similar to annotator A's.

Finally, only the speeches with the same annotations by two annotators in the last round are selected for experiments.

### 3.1. Experimental Results and Discussion

In our experiments, classification error rate is used to assess the performances of different features. GMM is used as the classifier, and 5-fold cross validation is carried out to fetch the error rate.

Figure 3 shows the results of three different intonation features. It shows that correlation is better than RMSE, and SEV is better than correlation.

In the first group of Figure 3, only the tone curve is used to calculate intonation similarity features. SEV performs much better than correlation and RMSE. Because in the pitch extraction procedure, it is hard to avoid doubling and halving problem, and there is usually a creaky voice when pronouncing the third tone of standard Chinese. Although the pitch is smoothed and interpolated with a spline function, it seems that correlation and RMSE are more sensitive to these errors. But SEV are relatively steadier than the other two. This is the best benefit of SEV.

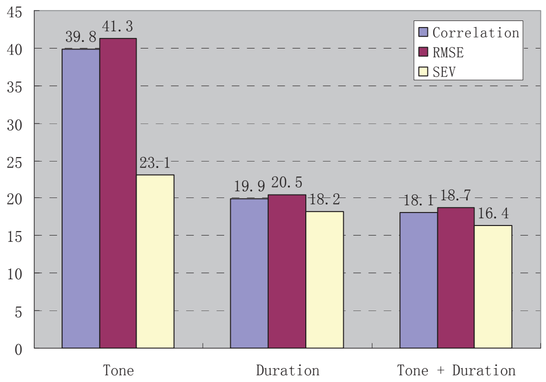

Figure 3: Error Rates of different intonation features

In the second group of Figure 3, only the duration similarity features are used. It shows that SEV also performs the best. By comparing with the first group, it is clear that duration features are better than tone features.

In the third group of Figure 3, both tone similarity and duration similarity features are used. Better results are achieved when comparing to the first and the second group. Experiment results show that SEV is still the best for intonation assessment.

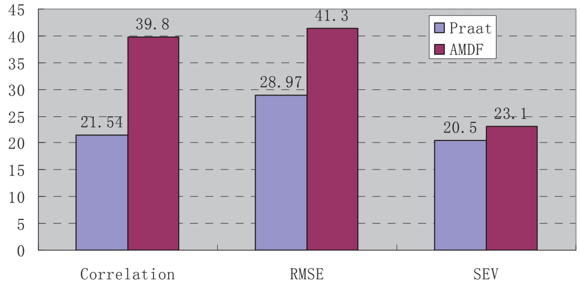

Figure 4: Error Rates of different pitch extraction methods

In the above experiments, we can see that duration feature is better than F0 feature. Then we apply different approach of F0 extraction method to see the change. Figure 4 shows the experimental results of two pitch extraction methods. It is clear that different extraction method really come to different results. It also shows that SEV is more robust than correlation and RMSE method. In fact, our AMDF method is worse than Praat method. In this case, the performances of correlation and RMSE degrade rapidly, while the performance of SEV is nearly the same.

Both correlation and RMSE use pre-set coefficients to weight different degrees of mismatching errors. Some of the distinguishable information lost in the progress. But SEV keeps much more original messages than correlation and RMSE. It has four kinds of benefits:

First, SEV can deal with sentence of different length. No matter how many syllables are there in the sentence, the size of SEV is a constant.

Second, SEV does not assume any weight to any error, so the weights may be optimized by the classifier.

Third, SEV does not assign any relationship from the errors to the final results, no matter linear or non-linear. The relationship between the best hearing results and the input vector is let to be optimized by classifier. Selecting optimal classifier may lead to optimal results.

Fourth, SEV feature is more robust then other features. It is not sensitive to the performance of different pitch extracting algorithms.

## 4. CONCLUSION AND FUTURE TASK

To the best of our knowledge, this was the first paper to propose Sorted Error Vector (SEV) feature for intonation assessment of standard Chinese. Experimental results show that no matter in which group of the experiments, SEV outperforms the correlation feature and RMSE feature.

As is mentioned in section 2, SEV results can be further improved by using suitable distance function $D(X,Y)$ and suitable transformation function $f(\bullet)$ . Finding suitable expressions for $D(X,Y)$ and $f(\bullet)$ may be our future task.

## 5. REFERENCES

[1] Y. Morlec, G. Bailly and V. Aubergé, "Synthesis And Evaluation Of Intonation With A Superposition Model", EuroSpeech 1995

[2] Gerit P.Sonntag and Thomas Portele, "Comparative Evaluation of synthetic prosody with the PURR method", ISCSLP 1998

[3] D. J. Hermes, “Measuring the perceptual similarity of pitch contours”. Journal of Speech, Language, and Hearing Research, 41:73–82, February 1998.

[4] Robert A.J. Clark and Kurt E. Dusterhoff, "Objective Methods For Evaluating Synthetic Intonation", EuroSpeech 1999

[5] Kim C. and W. Sung, “Implementation of An Intonational Quality Assessment System,” In Proceedings of International Conference on Spoken Language Processing, 2002, pp.1857-1860.

[6] Albert Rilliard And Veronique Auberge, "Prosody Evaluation as a Diagnostic Process: Subjective vs. Objective Measurements", International Journal Of Speech Technology 6, 409–418, 2003

[7] Joseph Tepperman, Abe Kazemzadeh, and Shrikanth Narayanan, "A Text-free Approach to Assessing Nonnative Intonation", InterSpeech 2007

[8] Huibin Jia and Jianhua Tao, "Automatic Prosody Quality Evaluation of Mandarin Speech", O-COCOSDA2007

[9] Huibin Jia, Jianhua Tao and Xia Wang, "Prosody Variation: Application to Automatic Prosody Evaluation of Mandarin Speech", Speech Prosody 2008