Perceptual Analysis of Duration Evaluation in Mandarin

Sun Lu, Hu Yu, Wang Ren-Hua

iFlyTek Speech Laboratory
University of Science and Technology of China
{sunlu;jadefox}@ustc.edu; rhw@ustc.edu.cn

Abstract

This paper is emphasized on analyzing the evaluation ability of RMSE and correlation in duration prediction. RMSE and correlation, as two chief evaluation means of duration prediction models, have their own strong points and weaknesses. We have hence carried out a perceptual experiment to investigate the characteristics of these two means, both in comparing two prediction systems and in evaluating the performance of a single predictor. In analyzing the performance of a duration prediction system, we put forward a more effective approach to appraise the predicted duration and hence come to the conclusion that a duration prediction system can be improved through valid information of stress/unstress and phonologically important syllables.

1. Introduction

Prosody, perceived as stress, intonation and rhythm, is extremely important to our perception of natural speech, and timing is an essential part of prosody. Therefore a lot of research has been carried out on this subject, and hitherto there have been many approaches in duration prediction. However, there seems to be no universal standard of evaluation on this subject. Some researchers work to minimize the RMSE (root mean squared error), some strive to maximize the correlation between perceived and predicted segmental duration, and some give attention to both.

Till now, there have been many effective models and approaches to predict prosodic timing. For example, CART tree model [1, 2], Klatt rule model [2], Bayesian belief networks [3, 4], and SoP models proposed by Jan. Van Santen [5]. These model builders all strive to enhance the prediction ability by optimizing RMSE and correlation values.

In paper [1], Chung Hyunsong had experimented with several duration models and evaluated the result by RMSE and correlation values, which were separately calculated in vowel group and consonant group. Many researchers have adopted this approach of RMSE and correlation calculation, but nobody has ever pointed out the confounding effect of duration ranges of different phonemes in the calculation. During our study on characteristics and modeling of segmental duration, we have encountered this problem more than once, and are confused by these different values. Why is there so big a discrepancy between RMSE and correlation values given by different researchers? Is it because of the performance of different approaches? Or how were these values obtained and how much can they account for the feasibleness of the approach? And are they convincing in comparing the performance of several duration predictors? In the main body of this paper, we would make out this phenomenon through mathematical and experimental analysis.

Moreover, in the assessment of one single prediction system, is there a direct relationship between these two measures and the perceptual goodness? As natural utterances are not stable, but very flexible and changeable, is it true that only a perfect RMSE and correlation value can ensure satisfactory perception? And is there a threshold above which the prediction would sound perfect? As these problems are vital in improving an existent duration prediction system, they are also our goal in the following experiments and analysis.

This paper is arranged in the following order: Section 2 is a review of the theories of RMSE and correlation from which we demonstrate the correct calculation approach, and deduce the strong points and weakness of these two values in explaining the prediction ability. In section 3, the evaluation ability of RMSE and correlation in comparing two prediction systems is demonstrated by perceptual experiment. Then both the relationship between perceptual scoring and RMSE and the relationship between perceptual scoring and correlation are analyzed, and consequently prediction ratio is put forward for sentence evaluation. Finally in section 4 a conclusion is given and method of improvement in duration prediction is proposed.

2. Mathematical theories and analysis

As RMSE and correlation coefficient are the most popular measures in duration prediction evaluation, we first take a close look at the mathematical essence of these two values, and from which explain the strong points and weakness of these two values.

2.1. RMSE (root mean squared error)

The root mean squared error is the square root of the average squared difference between two vectors. In this application, RMSE measures how much the predicted segmental duration is deviant from the observed duration. It is calculated by:

$$RMSE = \sqrt{\frac{\sum^{n} (X_i - Y_i)^2}{N}} \qquad (1)$$

where X_i is the predicted duration of the *i*th segment and Y_i is the observed duration.

From equation (1), we can see that a large RMSE indicates a big discrepancy between the predicted and the observed duration, thus probably makes a bad performance of the predictor. However, if we take into account the observed duration, this axiom will not work infallibly. The effect of a big distort on a comparatively long segment will doubtlessly be reduced, while short segments are more vulnerable to prediction errors.

Moreover, RMSE value is highly dependent on language type and speech rate of the training samples. In Mandarin, initials and finals are usually predicted separately, and in other languages like English, German and French, vowels and consonants are predicted separately. As the vowels and consonants differ between languages, hence does their length. From the above analysis, we could affirm that the RMSE of predictors that are built for different languages could not absolutely demonstrate the soundness of the predictor. The big discrepancy between values given by different researchers lies largely on language types. In addition, faster utterance is prone to engender small RMSE while slow utterance otherwise, and utterance tempo within a certain range does little harm to the perception. [6] So RMSE solely as a measure is not sufficient, and after correlation coefficient is introduced to the measurement, this problem is alleviated to some extent.

2.2. Correlation coefficient

Correlation is a statistical technique, which can show whether and how strongly pairs of variables are related. The square of the coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. The calculation equation is like this:

$$CORR = \frac{\sum^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum^{n} (X_i - \overline{X})^2 \sum^{n} (Y_i - \overline{Y})^2}} \qquad (2)$$

where X_i and Y_i denote the same meaning as in equation (1), and $\overline{X}$, $\overline{Y}$ are the average of these two vectors.

Correlation coefficient, on the other hand, is not sensitive to speech rate or language type. If used as the measure of duration prediction evaluation, it allows for flexibility of speech rate. However, there will be great fraudulence if all phonemes are added into the calculation process, as most researchers have done. For example, in one prediction model we have configured, the correlation coefficient of all initials is 0.952, while, the average of the correlation of each initials is 0.568. This phenomenon is illustrated and explained by the following example.

Figure 1 is the box plot of the duration distribution of all 21 initials in Mandarin. The red boxes contain 50% of all sample duration and the solid lines show the reach of the longest and shortest sample duration.

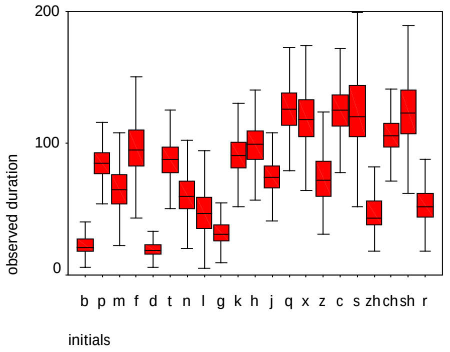

Figure 1: Box plot of initials duration

From Figure 1, we can see that the variance of duration between different identities is very large with a standard deviation of 40.072. This is a case in point. The correlation coefficient represents more of the general envelope, while neglects the small variance in each initials. Thus the correlation coefficient we obtained could not explain how exactly the duration of each initials is predicted. Therefore we've concluded that this value is of great fraudulence. For finals, although this gap is not so big due to the small variance in the duration of different finals, the fraudulence still exists.

3. Perceptual experiments

To investigate the describing ability of the two measures, RMSE and correlation coefficient, we carried out some perceptual experiments. The experiments have been done both for between-system analysis and within-system analysis.

3.1. Data preparation

We adopted a polynomial regression approach (hereafter referred to as PR approach), which is similar to Sum-of-Product model [5], and Wagon tree method [2] to predict Mandarin segmental duration. The objective evaluation is carried out for each sentence with syllable as the basic unit, and we averaged to obtain the final evaluation as listed in Table 1.

Table 1: Objective evaluation of PR and Wagon-tree prediction systems

	RMSE	Correlation
PR	33.432451	.764682
Wagon	40.964951	.704614

For the perceptual test, utterances are temporally adjusted by PSOLA in praat. [7] At the same time, the RMSE and correlation coefficient of each sentence is calculated.

3.2. Scoring criteria

As our experiment is about speech temporal perception, and there is no adjustment in fundamental frequency, the emphasis of scoring is put on temporal naturalness, thus a scoring criterion, which is similar to MOS, is designed as in Table 2. The subjective scoring is given with one significant digit of precision.

Table 2: Subjective scoring criteria

Score	Characteristics
5	Sounds natural, and can't be distinguished from the original utterance;
4	Only a few of the syllables sounds inconsistent, but doesn't spoil the whole utterance;
3	Some prosodic phrase sounds unnatural, and causes damage to the utterance, a little annoying;
2	Inconsistent throughout the whole utterance, sounds annoying, hard to accept;
1	Unacceptable.

3.3. RMSE and correlation coefficient of Mandarin prediction systems

50 sentences are randomly chosen for hearing test. Their segmental duration is separately predicted by PR predictor and Wagon-tree predictor. Then, six people who are professional in speech perception are invited to score these 50 sentences according to the above scoring criteria as listed in Table 1. Ultimately, the average score of these six people is taken as the valid score for each sentence.

The average score for PR prediction system is 4.27, Wagon-tree 3.50 and the original sentences 4.45. Referring to the objective evaluation displayed in Table 1, we should see that the difference described by RMSE and correlation of the two Mandarin prediction systems can be perceived.

But, how about the difference of RMSE and correlation in one prediction system?

3.4. Subjective scoring and RMSE for PR predictor

For the experiment of within-system perception, we choose 100 sentences of the same length (21 syllables) to be the subject. These 100 sentences are temporally adjusted by PR predictor, and scored in the same way as in the above experiment. The ultimate score is also set by the average of the scores from the six people.

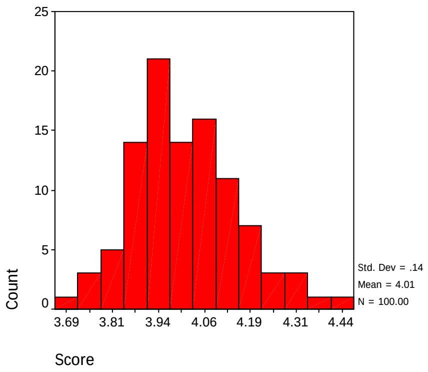

Figure 2. Distribution of average score

Figure 2 shows the distribution of average score, and the mean value of the average score is 4.01. As the satisfactory threshold of our scoring criteria is 4, we may assume that sentences with an average score of above 4 are satisfactory.

In the following, we worked to establish some relationship between subjective scoring and the objective evaluation by RMSE. In the above theoretical analysis about RMSE, we have deduced that, the smaller RMSE is, the better the sentence sounds. However, Figure 3, which is the scatter of subjective scoring and RMSE evaluation, has showed us a different fact. Obviously, there is no clear negative correlation between RMSE and the subjective scoring. To make the statement more reasonable, we have calculated the correlation between these two values by Pearson method, which is one of the simplest one in common use. The Pearson correlation calculation showed that there be no significant correlation between them. (p>0.002)

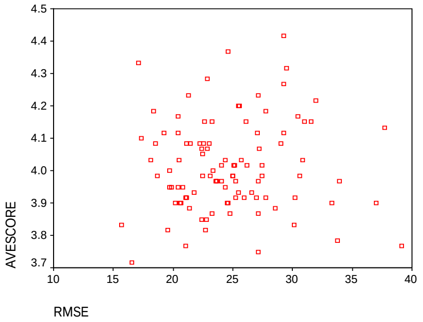

Figure 3. Scatter of average score and RMSE of the 100 sentences adjusted by PR predictor

3.5. Subjective scoring and correlation for PR predictor

Similarly, in the scatter of average scoring and correlation coefficient, there is neither any obvious negative relation, and the correlation coefficient provided by Pearson correlation calculation between these two values is again of no significance. (p>0.002)

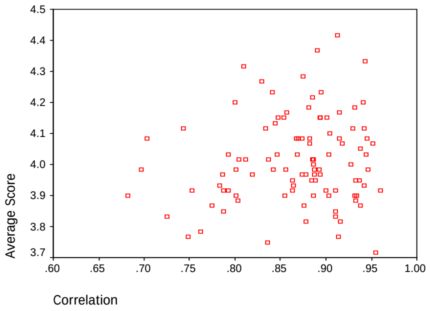

Figure 4. Scatter of average score and correlation coefficient of the 100 sentences

Figure 3 and Figure 4 disclosed that there is some great defect in measuring the duration prediction of a sentence in that we can not capture any direct relationship between the subjective and objective scoring. For RMSE, if the segmental length of most syllables is comparatively long, the RMSE value we've obtained will be inevitably bigger, while the perceptual scoring is still high. As to correlation coefficient, it is vulnerable to the deviation of a few numbers of syllables, while sometimes these syllables are of no importance in perception.

However the relationship between RMSE and correlation coefficient is still strong, they are negatively correlated, and the correlation is significant (p=0.00). That's to say, there is no problem in the calculation of RMSE and correlation coefficient, yet they are not fitful for evaluation of a certain sentence. In the following experiments, we have tried some other measure.

3.6. Subjective scoring and ratio averages for PR predictor

Since most researchers undoubtedly choose RMSE and correlation as the evaluation criteria and the theoretical foundation is so forceful, the above conclusion is rather a surprise and more of a puzzle for us. In application, some theoretical approaches perhaps don't come out so efficient.

To reduce the effect caused by different syllable length, we replace the predicted duration with prediction ratio, which calculated by Equation (3):

$$Ratio = abs(1 - \frac{\Pr ed}{Orig}) \tag{3}$$

where $\Pr ed$ is the predicted syllable duration, and $Orig$ is the original syllable duration. Each syllable corresponds to a ratio value, and the average of all ratio values in a sentence is taken as the average prediction ratio of a sentence.

There is a significant relationship between the average prediction ratio of each syllable and the subjective scoring of the sentence (p = 0.00, corr = -0.200).

As we have mentioned above that the general perception of a sentence might be ruined by perhaps several syllables, that's to say, the "most badly" predicted several syllables might have a heavier weight in subjective scoring. Hence we've chosen and average the big ratio values, and obtained the following list of correlation as in Figure 5. The x-coordinate denotes the number of biggest prediction ratio values, and the y-coordinate is the correlation coefficient between the average of the several biggest prediction ratio values and the subjective scoring.

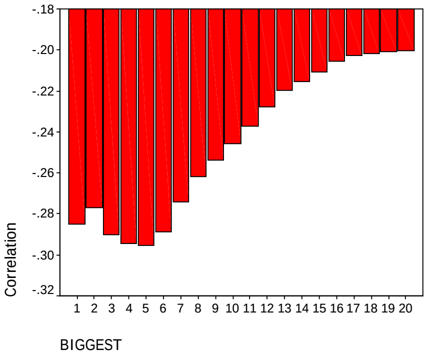

Figure 5. Correlation coefficient between the average of several biggest ratios and subjective scoring

All of these correlations are significant at the 0.002 level (p = 0.000). Compared with RMSE and correlation coefficient, this measure is more rational in sentence-wise evaluation.

To estimate the threshold above which the predicted sentence sounds satisfactory, we divide the satisfactory sentences, the average scoring of which is above 4.0, from the others, but no clear watershed could be found. Therefore, it is believed that there be no threshold of satisfactory prediction even by the measurement of prediction ratio.

What's more, as we should notice, the average of the five biggest prediction ratio values is most closely related with subjective scoring. That's to say, the subjective perception is vulnerable to five "big" corruptions.

What have caused those big corruptions? We have analyzed several sentences on this regard and found that the prediction ratio values of unstressed syllables are comparatively higher than that of stressed syllables, because we lack stress/unstress information in the predictor configuration. Also, we have noticed that if the duration of those syllables that carry more information than others are "badly" predicted, the scoring will be greatly influenced.

So, to improve the performance of a duration prediction system, valid information of stress/unstress and phonologically important syllables is indispensable.

4. Conclusion

This paper has presented our investigation on the evaluation ability of RMSE and correlation coefficient through mathematical and experimental analysis. From the first experiment, we illustrated that RMSE and correlation coefficient of different duration prediction systems can be perceived by listeners, thus these two values are valid in the comparison of different prediction approaches.

However, in the following experiments, we've found that RMSE and correlation coefficient of sentences predicted by one single prediction system is not discernible. So in the adaptation and improvement of some duration prediction systems, we should introduce other measures. Prediction ratio has been proved to be an effective assessment in sentence prediction evaluation. As we have also noticed the function of the biggest five values, it may be concluded that, the perceptual assessment of a sentence is vulnerable to several "most badly" predicted syllables.

This has been an important clue in improving duration prediction. In investigating the temporally adjusted utterances, we've discovered that the prediction ratio values of unstressed syllables are comparatively higher than that of stressed syllables, because we don't have effective stress information. Also, the perception of the whole sentence may be ruined by the "bad" prediction of some important syllables that carry more information than others. Therefore, to perfect the prediction system, valid information of stress/unstress and phonologically important syllables should be provided.

5. References

[1] Chung, Hyunsong, 2002. Duration models and the perceptual evaluation of spoken Korean. *Proceedings of Speech Prosody 2002*, 219-222, Aix-en-Provence, France.

[2] Brinckmann, C.; Trouvain, J, 2003. On the role of duration models and symbolic representation for timing in synthetic speech. *International Journal of Speech Technology* 6, 21-31.

[3] Goubanova, O.; Taylor, P.. 2000. Using Bayesian Belief Networks for model duration in text-to-speech systems. *CD-ROM Porceedings ICSLP2000,* Beijing.

[4] Goubanova Olga. 2001. Predicting segmental duration using Bayesian belief networks. In *SSW4-2001*, paper 139.

[5] J.P. H. van Santen. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, Vol. 8, 95-128.

[6] Fant G. Kruckenberg A. 1996. On the quantal nature of speech timing. *Proceedings of the International Conference on Spoken Language Processning*, 1996, 2044-2047.

[7] P. Boersma; D. Weenink. 2002. Praat: doing Phonetics by Computer, Version 4.0.26. **http://www.fon.hum.uva.nl/praat**.