Polynomial Regression Model for Duration Prediction in Mandarin

Sun Lu, Yu Hu, Ren-Hua Wang

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

Abstract

Duration modeling is to establish a mapping relationship between the prosodic context and the segmental duration engendered in natural speech. In this paper, we first study the effect of prosodic features on segmental duration of neutral utterance in Mandarin by introducing a statistical concept---eta squared, then choose more forceful prosodic features and design interaction quantifying algorithm to study the interaction phenomenon among them, and finally determine the duration model using a polynomial and obtain the coefficients through nonlinear regression. Our research work indicates that 5 to 6 prosodic features might by and large assist a close and accurate mapping between prosodic context and the perceived duration. Compared to Wagon tree method, this one has undeniable merits.

1. Introduction

Duration information is essential part of speech prosody, and plays a critical role in improving the naturalness and understandability of synthesized speech. Till now, there are two basic methods in predicting segmental duration. One is rule-driven approach, by which syntax rules are applied to determine segmental duration and make adjustment. The other is data-driven approach, completely based on data processing, statistical studies and machine learning. With the development of computer technology, the capabilities of data storage and data processing have been greatly improved, so the data-driven method, which is realized mainly through establishing duration models, has more and more merits.

Duration modeling is to establish a mapping relationship between the prosodic context and the segmental duration engendered in natural speech. There are several methods to capture this mapping relationship. Direct model training methods include neural network [1] [2] and decision tree [3]. Or the contextual features should be analyzed first and then construct a polynomial to link them with the duration. There are many approaches to calculate the coefficients in these polynomial models. The simplest one and also the most widely used one is nonlinear regression [4]. Some researchers have also tried the method of MARS algorithm [5], EM algorithm [6] and Bayesian network [7], etc.

In this paper, we adopted the method of feature analysis, polynomial construction, and nonlinear regression method (hereafter referred to as PR model). The PR duration predictor is trained with 10000 sentences in Mandarin, and evaluated through subjective scoring. The result shows that compared with decision tree method, this model has undeniable merits.

This paper is arranged as the following: section two introduces the application of eta squared in describing the forcefulness of certain contextual features; in section three we design an algorithm to calculate the residual of feature pairs, quantify the interaction between them, and finally fix the polynomial expression; in section four the training and testing result is given and perceptual evaluation is conducted; the final section is our conclusion and anticipation in this research field.

2. Select Prosodic Features

Prosodic features are contextual and structural characteristics of a speech segment. The well-known assumption is that each feature helps determine the perceived segmental duration. Usually dozens of features can be labeled and identified, but most of them are not forceful influencing factors and some even overlap. The first step in analyzing prosodic features is to study the forcefulness of each prosodic feature and choose several effective ones. In this section, we introduce the statistical concept eta squared to perform this task.

2.1. Introduction of Eta squared

Previously we use hypothesis testing to decide whether a prosodic feature influences the duration of a speech segment, but how forceful of the feature is unknown. In this paper we introduce the statistical concept eta squared to indicate to what degree the perceived duration is influenced by a certain prosodic feature.

Eta squared is the proportion of the total variance that is attributed to a factor and is calculated by:

$$\eta^2 = \frac{SS_{total} - SS_{within}}{SS_{total} - SS_{between}} = \frac{SS_{between}}{SS_{total}}$$

As $\eta^2$ can reflect the extent to which the duration is related to a certain feature, we here apply this method to explain the forcefulness of prosodic features.

2.2. $\eta^2$ Calculation

In our labeling system, the phrase is marked by 5 layers. L0 indicates syllable layer, L1 foot layer, L2 prosodic word layer, L3 main phrase layer and L4 breath group layer.

For initials, the labeled features include identity of the initials, type of the initials, identity of finals in current syllable, type of finals in current syllable, identity of previous finals, tone type, tone type of previous syllable, tone type of next syllable, previous boundary, next boundary, position of current syllable in the sentence, length of each prosodic layers. For finals, the labeled features are identity of current finals, type of current finals, identity and type of initials in the current syllable, identity of initials in the next syllable, tone
type, tone type of previous syllable, tone type of next syllable, previous boundary, next boundary, position of current syllable in the sentence, length of each prosodic layers.

The statistical tool SPSS [8] is used to calculate the eta squared of each prosodic feature, and the results are listed below:

Table 1: $\eta^2$ of initials and finals

<table>
<thead>
<tr>
<th>Prosodic</th>
<th>$\eta^2$</th>
<th>Prosodic</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initials</td>
<td></td>
<td>Finals</td>
<td></td>
</tr>
<tr>
<td>Ini_ID</td>
<td>0.839</td>
<td>Final_ID</td>
<td>0.839</td>
</tr>
<tr>
<td>Ini_type</td>
<td>0.781</td>
<td>Final_type</td>
<td>0.781</td>
</tr>
<tr>
<td>Final_ID</td>
<td>0.239</td>
<td>Ini_ID</td>
<td>0.239</td>
</tr>
<tr>
<td>Final_type</td>
<td>0.017</td>
<td>Ini_type</td>
<td>0.017</td>
</tr>
<tr>
<td>PFinal_ID</td>
<td>0.008</td>
<td>PIni_ID</td>
<td>0.008</td>
</tr>
<tr>
<td>PFinal_type</td>
<td>0.001</td>
<td>PIni_type</td>
<td>0.001</td>
</tr>
<tr>
<td>Tone</td>
<td>0.058</td>
<td>Tone</td>
<td>0.058</td>
</tr>
<tr>
<td>PTone</td>
<td>0.003</td>
<td>PTone</td>
<td>0.003</td>
</tr>
<tr>
<td>NTone</td>
<td>0.002</td>
<td>NTone</td>
<td>0.002</td>
</tr>
<tr>
<td>PreBoundary</td>
<td>0.016</td>
<td>PreBoundary</td>
<td>0.016</td>
</tr>
<tr>
<td>PostBoundary</td>
<td>0.019</td>
<td>PostBoundary</td>
<td>0.019</td>
</tr>
<tr>
<td>Position in L2</td>
<td>0.006</td>
<td>Position in L2</td>
<td>0.006</td>
</tr>
<tr>
<td>Position in L3</td>
<td>0.001</td>
<td>Position in L3</td>
<td>0.001</td>
</tr>
<tr>
<td>Length of L2</td>
<td>0.003</td>
<td>Length of L2</td>
<td>0.003</td>
</tr>
<tr>
<td>Length of L3</td>
<td>0.001</td>
<td>Length of L3</td>
<td>0.001</td>
</tr>
</tbody>
</table>

According to the table listed above, some of the weak factors can be removed from the data and kept are the more influential contextual features including: initials identity, identity of finals in the current syllable, tone type, previous boundary and posterior boundary for initials; final identity, identity of initials in the current syllable, tone type, posterior boundary, position in L3 and length of L3 for finals.

Through forcefulness evaluation and feature selection, we have cut down the vector space to a much smaller and acceptable size to reduce the damage to the training procedure that may be caused by unbalanced data.

3. Interaction between prosodic features

Though prosodic features influence the segmental duration, the do not act independently; some of them may interact [9]. For example, in Mandarin, syllables with tone 3 or occurring in the end of a sentence usually have much longer duration, however when these two feature co-occur, their length will not be double-lengthened but rather be shortened. We design an algorithm to quantify this interactive phenomenon for each initials and finals.

3.1. Interaction Quantification

The quantified interaction between two factors is calculated through constructing a table with the combination of the factor values for which the interaction is to be studied as the row headings, and the all combinations of values of the other factors as column headings. All segment samples are fitted into each cell according to the values of their prosodic features. In each cell, the mean duration of all samples is calculated and number of them is counted. Actually, due to the sparsity of speech data, it is impossible that all cells are filled with sufficient data. To remove the randomness of samples in some cells, we clear up cells with less than 5 samples.

For each column, we calculate the average of all samples in cells which belong to the column, and subtract this average from the mean duration value in each cell in this column to get the residual ($resid_{m}^{n}$, where m and n are the valid column and row numbers of this cell) of samples in each cell. Then the squared sum of all the residuals ($SQ_{Residual}^{m}$) in the column are obtained following formula (2) to represent the interaction of the studied features under the prosodic context denoted by the column heading.

$$SQ_{Residual}^{m} = \sum_{n} (resid_{n}^{m})^2$$

Finally following formula (3), the average squared residuals are summed up to obtain the final value as the quantified interaction between two factors regardless of prosodic context determined by other factors.

$$Residual = \frac{\sum_{m} SQ_{Residual}^{m}}{num_{m}}$$

where $num_{m}$ is the number of samples in the mth column.

In another word, the interaction between two factors is quantified through calculating how far away the engendered duration can be dragged by the combined effect of the two factors. This method is not as crucially subject to data sparsity as traditional figure explanation.

3.2. Interaction Pattern Classification

The interaction between any two forceful factors is calculated for each initials and finals.

For all initials, strong interaction occurs between current finals and tone type, current finals and previous boundary, current finals and posterior boundary as illustrated in Figure 1.

Figure 1: Residual distribution of Initials

For different finals, the interaction pattern acts differently. We classify the finals into three groups according to the distribution pattern of interactive pairs. The first group includes i, u, ai, ao, ian, ing, uan, a, e, ie, uo, ou, iou, uei, an, en, ang, eng, in, iang, uen, and strong interaction occurs between posterior boundary and position in L3, tone type and position in L3, current initials type and position in L3, as illustrated in Figure 2. The second group includes ii, iii, ei, o,
iong, ua, uai, van, ia, v, ong, ve, and strong interaction occurs between current initials type and other features, posterior boundary and position in L3, as illustrated in Figure 3. The third group has only one finals: er, and strong interaction occurs between posterior boundary and position in L3, as illustrated in Figure 4.

For initials, there is only one model expression: \( \text{ALL ADD} + \text{current initials type} \times \text{tone} \times \text{position in L3} \times \text{ALL MULT} \). Here, the independent influence of all factors is embodied in the \( \text{ALL ADD} \) item. And in finals models, we use \( \text{ALL MULT} \) to represent the overall interaction between all factors.

For finals, in determining the polynomial expression, we should also consider the number of samples for training, because for nonlinear regression the number of samples should be sufficient to ensure validity of the regression. Thus two finals (uang and vn) are taken from the second group to form a special group in the polynomial model of which the \( \text{ALL MULT} \) item is neglected to cut down the number of coefficients. Thus the four polynomial models for finals are decided as:

1. Group one: \( \text{ALL ADD} + \text{current initials type} \times \text{position in L3} \times \text{current initials type} \times \text{ALL MULT} \);
2. Group two (except uang and vn): \( \text{ALL ADD} \times \text{current initials type} \times \text{current initials type} \times \text{position in L3} \times \text{ALL MULT} \);
3. Group three: \( \text{ALL ADD} \times \text{position in L3} \);
4. Group two (only uang and vn): \( \text{ALL ADD} \times \text{current initials type} \times \text{current initials type} \times \text{position in L3} \);

4 Model Training and Testing

4.1 Model Training and objective Testing

The total number of training sentences is 10000 and testing sentences is 4500. For each initials and finals a model is trained by nonlinear regression. The training and testing result is calculated by putting all initials together and all finals together. Table 2 is the result of model training and testing.

<table>
<thead>
<tr>
<th></th>
<th>Initials</th>
<th>Finals</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>Training</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>12.21ms</td>
<td>12.23ms</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.952</td>
<td>0.952</td>
</tr>
</tbody>
</table>

To make a comparison, we also build a Wagon tree model using the same prosodic features to predict segmental duration. 1000 sentences with the same length are chosen to be predicted by both methods and the RMSE and correlation of each sentence is calculated and averaged to obtain the averaged objective evaluation.

<table>
<thead>
<tr>
<th></th>
<th>model</th>
<th>PR</th>
<th>Wagon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RMSE</td>
<td>26.0ms</td>
<td>26.4ms</td>
<td></td>
</tr>
<tr>
<td>Mean correlation</td>
<td>0.858</td>
<td>0.856</td>
<td></td>
</tr>
</tbody>
</table>
We can see no big improvement in objective evaluation, however what matters in practical system is not objective data but rather subjective hearing. We then conducted a subjective scoring to further evaluate the two methods.

4.2 Subjective Scoring

50 sentences with the same length (21 syllables) are chosen for hearing test. The segmental duration of each sentence 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 a 5 grade criteria as fixed in our previous work [11]. Ultimately, the average score of these six people is taken as the valid score for each sentence.

Table 4: Subjective scoring

<table>
<thead>
<tr>
<th></th>
<th>PR</th>
<th>Wagon tree</th>
<th>Wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score</td>
<td>4.27</td>
<td>4.08</td>
<td>4.45</td>
</tr>
</tbody>
</table>

In table 3, wav is the original recorded speech. From the perceptual test, we can conclude that PR predictor performs better than wagon tree.

However, considering the data listed in table 3, people should doubt the principles behind the two different scoring methods. Through careful scrutiny, we find that in wagon tree predicted sentences, the error of most syllables are very small, but there always exist one or two syllables which are awfully predicted, and it is this small number of syllables that have spoiled the perceptual naturalness of the whole sentence. The main reason of wagon tree’s occasional performance is that during the wagon tree training procedure, data is prone to be overfitted, which refers to the possibility that the model is adjusted to fit the training data too exactly without sufficient generalization. So it is unavoidable that new data come to be trapped in a certain branch and result in big error. Thought there are methods to control overfitting, it is hard to adapt to duration prediction.

Comparatively, PR model can make very stable prediction, that is, there is seldom any big discrepancy because nonlinear regression is performed in an ittnerative approach and the coefficients are updated in each iteration, avoiding the case that some parameters are determined by a single sample or a few number of samples.

Conclusion

In this paper, we have introduced the statistical conception--eta squared to evaluate the forcefulness of prosodic features in influencing segmental duration, designed the interaction quantifying algorithm, and put forward the idea of classifying phonemes in fixing polynomial models. As interaction also occurs in other prosodic realization, this set of analyzing method might also be adapted to analysis in fundamental frequency, accent and stress, etc.

From the above analysis, it is obvious that PR model can make much stable prediction compared with wagon tree. Also as the principle of nonlinear regression is very simple, it does not take much space and resources in predicting segmental duration though the initial analyzing procedure is very complicated.

There are two noticeable problems in duration prediction. When we study the predicted duration of initials, we find that for affricatives and fricatives the predicting error is comparatively small and correlation is high, but for others like stops, laterals and nasals, whose intrinsic duration is short (about 20-30ms for stops, 45ms for laterals and about 60ms for nasals), the prediction is dissatisfying. According to our discovery in previous work [11], these initials are not subjective to prosodic context, mainly the finals type of the current syllable, as affricatives and fricatives, but rather to unpredictable factors. Whatsoever, the inaccurateness of initials duration prediction does not affect much on perception.

Another problem is that, in the PR prediction system, duration is better predicted for completely neutral utterances, which refers to utterances without any stress or emotion. Otherwise, there is perceptible discrepancy. As our major goal is to model neutral utterance, work on the several emotional utterances will exceed the topic of this paper. Nevertheless, duration modeling for emotional utterances is also a challenging research topic.

References