Totally Data-driven Duration Modeling Based on Generalized Linear Model for Mandarin TTS

Lifu Yi, Jian Li, Xiaoyan Lou, Jie Hao

Toshiba (China) Research and Development Center
{yilifu, lijian, louxiaoyan, haojie}@rdc.toshiba.com.cn

ABSTRACT

This paper proposes a totally data-driven duration modeling method for Mandarin TTS, which uses Generalized Linear Models (GLM) to model duration and stepwise regression to automatically select the attribute set with statistical measurements. This method can get a good tradeoff between model complexity and goodness of fit. Besides, speaking rate is introduced as a new modeling attribute, which not only achieves higher performance but also provides a novel approach to adjust speaking rate when synthesizing. We also propose to use R^2 to fairly evaluate the modeling performances on different databases, since R^2 refers to the fraction of corresponding variance explained by a model. Experiments show the performance of GLM is significantly higher than that of CART. With our much smaller models and corpus, the proposed method also achieves comparable results reported by other excellent researches.

Index Terms: duration modeling, data-driven, generalized linear models, speech synthesis

1. INTRODUCTION

Duration model is an important part of speech synthesis. It predicts the reasonable duration of speech units according to the linguistic and phonetic attributes.

The goal of duration model is to predict value $\hat{d}$ by q-dimensional attribute vector a as close as possible to real value d. Sum-of-Products (SOP) [1] is a popular method for duration modeling, defined as follows:

$$\hat{d} = \sum_{i \in T} \prod_{j \in I_i} (\beta_{i,j} a_j) \quad (1)$$

SOP is a generalization of the multiplicative model [2]. When $|T| = 1$ and $I_1 = \{1, 2, \ldots, q\}$ in Eq.(1), the SOP model is turned into a multiplicative model. CART is another popular method for modeling duration, which successfully minimizes the prediction error by partitioning the attribute space by a binary decision tree [2]. Other methods, such as MARS[3], EM[4], ANN[5] and BNF[6], can also model duration very well.

Although above methods are inspiring for duration modeling, the imbalance problem between size of database and attribute interactions still assails us [6]. When training data is fixed, the more attribute interactions, the more complex model. If the model is too complex for the training data, it memorizes parts of the noise as well as learns the true problem structure. This will cause overfitting problems. Conversely, it causes underfitting problems. To make tradeoff between goodness of fit and model complexity, it's necessary to find the most important attributes and attributes interactions. However the attribute sets of most existing models are preset by experience or analysis. Some promising methods [7] [8] [9] are proposed on this topic, but they didn't provide totally automatic solutions for attribute selection.

The GLM and stepwise regression we proposed just gives a totally automatically solution to the attribute selection problem. The attributes and attributes interactions are automatically selected by stepwise regression based on Bayes information criterion (BIC) and F-test. With these statistical methods, the model structure and prediction coefficients are optimized at the same time. In addition, this paper introduces speaking rate as a new attribute. We also introduce the R^2 to fairly compare the prediction errors on different databases.

This paper is organized as follows: section 2 introduces the modeling methods. The corpus and the attributes will be introduced in section 3. The training and evaluation experiments will be described in section 4, followed with section 5 of conclusion and future work.

2. MODELING METHOD

This paper models the duration based on GLM, optimizes the attributes and attributes interactions by stepwise regression based on F-test and BIC, and uses R^2 to evaluate model.

2.1 GLM introduction

GLM based model is a generalization of multivariate linear regression model [10]. The GLM model predicts the duration $\hat{d}$ from attribute set $\boldsymbol{A}$ of speech unit s by:

$$\hat{d} = h^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_i a_i) \quad (2)$$

Where h is a link function, $(\beta_0, \beta_1, \cdots, \beta_p)$ is the vector of regression coefficients and p is the dimension of the regression coefficient vector, a_i is an attribute or attribute interaction.

Using different link functions, we can get different exponential family distributions of d. a_i in Eq.(2) can be a second order attribute interaction, such as $(a'_{in} \times a'_{im})$, a'_{in} and a'_{im} are linear attributes from attribute vector A. We assume Gaussian distribution for duration, accordingly, h equals I (identity function). With above explanations, Eq.(2) can be denoted as follows.

$$\hat{d} = \beta_0 + \sum_{i=1}^{p} \beta_i a_i = \beta_0 + \sum_{i=1}^{p'} \beta_j a_j + \sum_{i=p'+1}^{p} \beta_i (\prod_{j \in I_i} a_{ij}) \quad (3)$$

We can see an obvious difference between SOP of Eq.(1) and GLM of Eq.(3), that is GLM treats attribute interactions as a "new" linear attribute, while SOP does not. In SOP, an attribute interaction item needs multiple coefficients but GLM need only one coefficient. GLM also can be used as non-linear model, such as introducing some exponential parts. Coefficients of GLM are estimated by iterative maximum likelihood estimation method, not by least squares estimation method.

2.2 Attribute selection

When the model structure is determined, the performance of the model is tightly related to the attribute sets. The best attribute set turns out the highest performance. We use BIC and stepwise regression to automatically find the "best" model.

2.2.1 BIC

BIC is a widely used statistical criterion, which gives a measurement integrating both model complexity and the goodness of fit. It is defined as:

$$BIC = n\log(SSE/n) + p\log n \quad (4)$$

Where *SSE* is the sum of squared prediction errors and n is the amount of training data. The first part of right side of the equation 4 indicates the precision of the model and the second part indicates the penalty for the model complexity. When the number of training sample n is fixed, the more complex the model is, the larger the dimension p is, the more precise the model can predict for training data, and the smaller the *SSE* is. So the first part will be smaller while the second part will be larger, and vice versa. The increase of one part always leads to the decrease of the other part. When the summation of the two parts is minimized, the model is optimal.

2.2.2 Stepwise regression

If there are q components in attribute vector $\boldsymbol{A}$, there should be 2^q–1 linear combinations of the attributes. If non-linear combinations are considered, the number should be infinite. It is very time consuming to select the optimal model by evaluating all the possible models. Stepwise regression introduced here is very efficient to solve the problems.

Stepwise regression avoids training all the possible models and hence saves much time by repeating steps described in Fig.1.

This training is an off-line process. We can always get the "best" model for a given corpus of the narrator. For example in the Fig.1, suppose that the duration is only affected by attributes "phone" and "tone", note the model as: duration~β_0+β_1×phone+β_2×tone+β_3×(tone*phone),
"tone*phone" means the interaction (combination) of phone and tone. This is the Initial model. Then we calculate F-test value of each item. Maybe the "tone*phone" item is the least important, if so, we remove it. Now we retrain the model: duration~β_0+β_1×phone+β_2×tone, calculate the BIC, maybe the BIC is minimized, if so, we can stop here and get the optimal model.

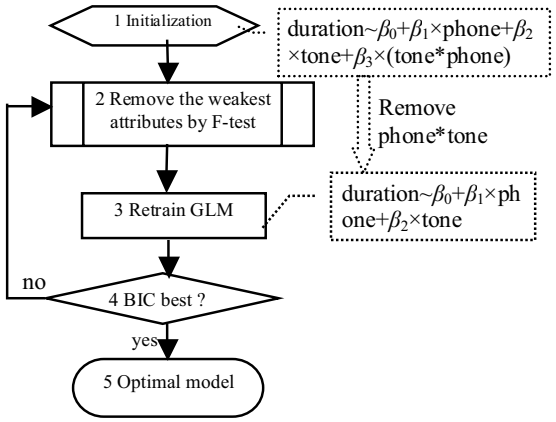

Fig.1: Flowchart of Stepwise regression for duration model training

If the initialized model is too complex to compute, we can decompose the training into two stages to reduce model dimension. Firstly, we select the most important linear attributes. Secondly, we can select the optimal model with these attributes and their second order combinations. Both stages follow the same steps described above.

2.3 Evaluation metrics

The duration models may be built on different corpora. The corpus differences should be considered when comparing the model performances.

Root of Mean Square Error (RMSE) and correlation (Corr) are frequently used in evaluation of duration modeling. RMSE is a measurement of the absolute prediction error and it highly depends on the corpus. Correlation measures the degree to which two variables are linearly related but it cannot measure the non-linear relationship. Neither RMSE nor Corr provides the way for eliminating corpus differences straightforwardly.

We find that R^2 is a more suitable evaluating metric for the requirement. R^2 is called the coefficient of determination, and it refers to the fraction of corresponding variance explained by a model. R^2 is defined as:

$$R^2 = 1 - \frac{\sum (\hat{d}_i - d_i)^2}{\sum (d_i - \bar{d})^2} = 1 - \frac{RMSE^2}{\sigma^2} \quad (5)$$

R^2 normalizes RMSE by the duration variances of different corpora. RMSE, Corr and R^2 are the 3 evaluation metrics used in this paper.

3. CORPUS AND ATTRIBUTES

Speech corpus is the foundation of duration modeling. The corpus contains the attributes of speech units and the durations of them.

3.1 Corpus

The models described above are trained and tested using our mandarin corpus. The corpus is narrated by a professional female broadcaster and contains 2,150 utterances sampled at 22.05 kHz. The corpus also consists of text information, such as Chinese word segmentation boundaries, acoustic information such as phoneme segmentations.

We perceived that the speaking rate of the corpus varies to some extent after listening to speech data of the corpus. The speaking rate is defined as the average number of syllables per second. The mean value of speaking rate is 4.47 syllables per second, and the standard variation is 0.47 syllables per second. The distribution of speaking rate is not very sharp as shown in Fig.2.

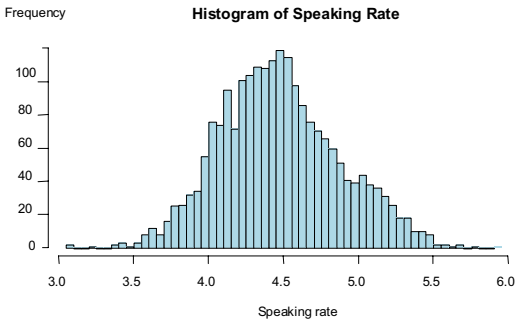

Fig.2: Histogram of Speaking Rate

Since the distribution of speaking rate is not very sharp, it's reasonable to model speaking rate into the duration models.

3.2 Attribute set

Theoretically, all linguistic and phonetic attributes are likely to influence duration [9] [11] [12]. Generally considered attributes are the information of the current phoneme and neighboring phonemes [9] [11]. The attributes used in this paper are listed in the table 1.

As we know, there are about 21 Initials and 39 Finals in Mandarin. Each Initial is just a single consonant phoneme and a Final may consist of one to three vowels. We use the 21 Initials and 39 Finals as the basic speech units for modeling.

Here speaking rate is adopted as a new attribute. In the training process, we get the speaking rate from the corpus, and in the synthesizing process, the speaking rate maybe obtained by the system or user configurations. So the speaking rate is known for both training and prediction phases.

Table 1: Definitions of the attributes

Attribute	Description
Pho	ID of current subsyllable
ClosePho	ID of another subsyllable in the same syllable
PrePho	ID of the neighboring subsyllable in the previous syllable
NextPho	ID of the neighboring subsyllable in the next syllable
Tone	Tone of the current syllable
PreTone	Tone of the previous syllable
NextTone	Tone of the next syllable
POS	Part Of Speech
DisNP	Distance to the next pause
DisPP	Distance to the previous pause
PosWord	Syllable position in the lexical word
ConWordL	Combined lengths of the current, previous and next lexical word
SNumW	Number of syllables in the lexical word
SPosSen	Syllable position in the sentence
WNumSen	Number of lexical words in the sentence
SpRate	Speaking rate

Furthermore, speaking rate maybe interacts with other attributes so that we can improve the precision of duration prediction by modeling speaking rate. The duration predicted in this way is adapted to the total length of the synthesized sentence, which is more reasonable than that by linear lengthening or shortening. Some other researches indicate that the effect of speaking rate on duration is different from phoneme to phoneme [12]. In other words, speaking rate does interact with other attributes.

4. TRAINING AND TESTING

The modeling method was described in section 2. Only two duration models are built by the two-stage method, one for all Initials and the other for all Finals.

In the first stage, we can get the most important attributes for Final: Pho, ClosePho, SpRate, DisNP, NextPho, Tone, PosWord, ConWordL. And the most important attributes for Initial are: Pho, ClosePho, SpRate, PosWord, PrePho, ConWordL, NextPho and Tone. The speaking rate is the third most important attribute for both Initial and Final. That indicates that speaking rate is very important for modeling duration.

In the second stage, the method of selecting interaction attributes is same as that in the first step. When the interaction attributes are selected, we have the optimal GLM models for evaluation.

We use the 75% data for training and the other 25% data for testing. To make a comparison, we also build CART duration model on the same datasets. The modeling performances are shown in Table 2.

Table 2: Performance comparison of GLM model and CART model (open test)

Model	GLM		CART	
	Initial	Final	Initial	Final
RMSE	13.82	32.05	14.76	36.18
Correlation	0.928	0.801	0.905	0.744
R^2	0.858	0.687	0.838	0.605

The unit of RMSE in this paper is millisecond. Table 2 shows that GLM model outperforms the CART model. One of the possible reasons is CART cannot utilize the interaction of attributes.

Sun also built excellent polynomial regression (PR) duration models in Mandarin [7] [8]. We give the comparison results between the GLM models and PR models as shown in Table 3. Since the two models are built on different corpora, we use R^2 to compare them.

Table 3: Comparison of GLM and PR [7] [8]

Model	GLM		PR [7] [8]	
	Initials	Finals	Initials	Finals
Corpus size	41k	48k	200k	200k
Model number	**1**	**1**	**21**	**39**
Correlation	0.928	0.801	0.952	0.826
SD	36.75	**57.56**	40.2	**44.92**
RMSE	13.82	32.05	12.23	25.44
R^2	0.858	**0.687**	0.907	**0.679**

SD in Table 3 is standard deviation of duration, and its unit also is millisecond, same as RMSE. From Table.3, we can see that there are some differences between two experiments. The first is our corpus size of syllable is about 1/4 of Sun's. The second is our model is much smaller in size than Sun's, because we adopt phoneme-independent modeling approach. We build only one model for all Initials and one for all Finals. However Sun's method builds 60 prediction models in all, one model for each Initial or each Final. SD of Finals in our corpus is much larger than Sun's. Larger SD means duration of our corpus varies more widely than Sun's. Nevertheless, from the point of view R^2, the proposed GLM method outperforms the PR method for Finals.

5. CONCLUSION AND FUTURE WORK

This paper proposes a totally data-driven method for duration modeling in Mandarin TTS. The GLM duration models adopt stepwise regression to automatically optimize the attribute set. For the first time, speaking rate is applied as a new attribute. We train only one Initial model and one Final model for all 60 mandarin phonemes, which reduces the model size significantly compared with the phoneme-dependent modeling methods.

From the above experiments results, the proposed GLM model is very compact, but reliable in performance. It significantly outperforms CART. With much smaller corpus, GLM method provides similar or even higher performance compared with the PR method, since Mandarin Finals are more important than Initial for the human auditory perception [7].

Besides, the PR method uses prosodic layers information [7] as input attribute. We will include this attribute into our method in the future. We also plan to use more attributes for further improvements, such as semantic, accent and emotional information. Duration model is a kind of prosodic models in TTS, and we hope the method proposed in this paper can be applied to other prosodic modules.

6. REFERRENCES

[1] Venditti, Jennifer J., Santen, Jan P. H. van, "Modeling Final duration for Japanese text-to-speech synthesis", In *ICSLP-1998,* pp.786-789.

[2] Batusek, Robert., "A Duration Model for Czech Text-to-Speech Synthesis", in *Speech Prosody 2002*., pp.167-170.

[3] Riedi M., "Modeling Segmental Duration with Multivariate Adaptive Regression Splines", in *EUROSPEECH97*, Rhodes, Greece, Vol.5, pp.2627-2630, 1997.

[4] Wen-Hsing Lai, Sin-Horng Chen, "A Novel Syllable Duration Modeling Approach for Mandarin Speech", in *ICASSP2001*, pp.706-709.

[5] S.H. Chen, S.H. Hwang, et al., "An ANN-based prosodic information synthesizer for Mandarin text-to-speech", *IEEE trans. Speech Audio Processing*, Vol.6, No.3, pp226-239, 1998.

[6] Goubanova, Olga, "Predicting segmental duration using Bayesian belief networks", In *SSW4-2001*, pp.139-142.

[7] Sun Lu, Yu Hu, Ren-Hua Wang, "Polynomial regression model for duration prediction in Mandarin", in *INTERSPEECH-2004*, pp 769-77.

[8] Sun Lu, et al.,"The Statistics and Analysis of Initials and Finals Duration based on Eta squared", in *NCMMSC2003*, Xiamen, China. 2003.

[9] Min Chu and Yong-Qiang Feng, "Study on Factors Influencing Durations of Syllables in Mandarin", in *Eurospeech2001*, Aalborg, 2001, pp.927-930.

[10] McCullagh P and Nelder JA, *Generalized Linear Models*, Chapman & Hall, 1989.

[11] Tao Jianhua, Ni Xin, "Auditive Learning Based Chinese F0 Prediction", in *ICASSP2003*, Hongkong, pp.500-503

[12] Chilin Shih, Wentao Gu, Jan P. H. van Santen, "Efficient adaptation of TTS duration model to new speakers", in *ICSLP-1998*, pp.177-180.