Phrase Break Prediction Using Logistic Generalized Linear Model

Lifu Yi, Jian Li, Xiaoyan Lou, Jie Hao.

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

ABSTRACT

In this paper we propose a novel phrase break prediction model for Mandarin speech synthesis. It is generalized linear models (GLM) with stepwise regression solution. We assume phrase break obeys Bernoulli distribution and then model phrase break probability by Logistic GLM. The attribute set is automatically selected by stepwise regression, which is a totally data-driven method. We also introduce speaking rate as a new attribute for prediction. The proposed method is applied to 2,150 utterances of the Mandarin speech corpus, and it achieves 5.4% higher performances than CART method in open test. The method can be extended to include more linguistic and prosodic attributes and it is very compact for application.

Index Terms: phrase break prediction, logistic generalized linear models, speech synthesis

INTRODUCTION

In TTS (Text to Speech) system, the synthesized speech is generated based on prosody information. The prosody information contains F0, duration, phrase break and etc, which are predicted based on linguistic attributes extracted from text. Because phrase break boundaries often cause dramatic changes in F0 and duration, the predicted phrase breaks are treated as an input attribute for F0 and duration prediction. Phrase break is crucial for the naturalness of the synthesized speech.

In Mandarin, linguistic research shows the prosody hierarchy system consists of 3 or 4 levels of prosody units [1][2]. A kind of prosodic phrase, also called intonation phrase, ends with an obvious phrase break. This paper focuses on the intonation phrase prediction. In the rest of this paper, we call it phrase break prediction while others may also call it pause prediction.

So far, many statistic models have been proposed for phrase break prediction, such as CART (Classification And Regression Tree) [2], MBL (Memory Based Learning) [3], HMM (Hidden Markov Model) [4][5], N-Gram [2], FST (Finite State Transducer) [6] and SVM (Support Vector Machines) [1] and ME (Maximum Entropy Model) [2][7]. Among them, CART, MBL and ME are popular methods in Mandarin TTS.

The above methods have achieved inspiring results in Mandarin phrase break prediction. However, most of methods assumed a Gaussian distribution for phrase break, other distributions were not studied yet. The linguistic attributes and attribute interactions in the methods are guided by existing linguistic knowledge, but not by totally data-driven methods. Moreover, they pay no attention on the contribution of the speaking rate to phrase break prediction.

In this paper, we propose a novel phrase break prediction model. We find that Bernoulli distribution is more reasonable for phrase break than Gaussian distribution by experiments. Hence we use Logistic GLM [8] to predict phrase break probability. The modeling attributes and attributes interactions in GLM can be automatically selected by the stepwise regression method. The attribute set selection in this paper is totally data driven.

This paper is organized as follows: firstly, we describe the novel phrase break model. Secondly, we introduce basic concept of GLM and stepwise regression for phrase break modeling. Then, we show the phrase break prediction experimental results. Finally, we draw some conclusions on this research.

1. MODELING METHOD

1.1 Phrase break modeling

Phrase break model is to predict phrase break from a sequence of contextual linguistic attributes. During training stage, a sequence of pairs of contextual linguistic attribute and tag is available:

$$(P_1, C_1), (P_2, C_2), \cdots, (P_i, C_i), \cdots, (P_m, C_m) \quad (1)$$

Where P_i is the phrase break tag, 1 or 0, after word w_i, and C_i is the contextual linguistic attributes including POS (part of speech), word length and etc. Actually, phrase break not only relies on the contextual linguistic attributes from the text, but also relies on speaker's intention, feeling and other statuses. The latter factors are difficult to model because they are subtle and always changing. The same text may have different possible phrase break positions. A Chinese example is shown as following:

1. 我们(We) # 都(all) 是(are) 研究员(researchers).
2. 我们(We) 都(all) 是(are) # 研究员(researchers).
3. 我们(We) # 都(all) 是(are) # 研究员(researchers).

('#' represents inserted phrase break)

In the above example, there are at least 3 different possible phrase break styles for the same sentence. We think, although the phrase break is a binary variable, either true (1) or false (0), it is more reasonable to treat phrase break as a probability, since speaker often changes styles. We assume the phrase break occurs independently each time with a certain probability, Pr, and Pr obeys Bernoulli distribution.

1.2 Logistic GLM Model

GLM is a generalization of multivariate linear regression model. The GLM model predicts the probability of phrase break from attributes by:

$$Pr_i = \hat{P}r_i + e_i = h^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_j C_{ij}) + e_i \quad 0 < i \le N \tag{2}$$

Where Pr_i is probability of phrase break, e_i is prediction error, h is a link function, β_j is the unknown regression coefficient, C_{ij} is the context linguistic attribute, p is the dimension of the regression coefficient vector. N is the number of training samples, i is index of a sample. C_{ij} can be not only linear attributes, but also attribute combination or attribute interactions. GLM treats an attribute interaction as a new linear attribute with only one regression coefficient, but SOP (sum-of-products) model treats an attribute interaction with multiple regression coefficients. When h equals identity function, the GLM is a plain GLM and phrase break obeys Gaussian distribution. When h equals to *logit* function, the GLM model is a logistic GLM model and phrase break obeys Bernoulli distribution [8]. The *logit* function is defined as:

$$h^{-1}(z) = e^z/(1+e^z) \tag{3}$$

$$h(\hat{P}r_i) = \text{logit}(\hat{P}r_i) = \log[\hat{P}r_i/(1-\hat{P}r_i)] = \beta_0 + \sum_{j=1}^{p} \beta_j C_{ij} \tag{4}$$

Both plain GLM and Logistic GLM attempt to estimate the posterior probability $\Pr(P|C)$ and have linear classification boundaries. In Logistic GLM, $\Pr(P|C)$ is nonlinear function of context $\boldsymbol{C}$. Logistic GLM guarantees $\Pr(P|C)$ to range from 0 to 1 while plain GLM does not.

Logistic model has been widely used in many statistical fields of classification and regression. The regression coefficients in Logistic GLM can be trained by iterative maximum likelihood estimation method. More details can be found in [8].

2.3 BIC criterion

To estimate the regression coefficients in Logistic GLM based on limited training data, we use Bayes Information Criterion (BIC) to weigh together complexity and goodness of fit of the models. BIC is defined as:

$$BIC = N\log(SSE/N) + p\log N \tag{5}$$

Where *SSE* is the sum of squared prediction errors. So the first part of right side of the Eq.(5) indicates the goodness of fit of the model. And the second part is a penalty for the model complexity. When the summation of the both parts is minimized, we get good goodness of fit and prevent overfitting and underfitting as well.

2.4 Stepwise regression

We suppose the distribution of phrase break obey Bernoulli distribution and we only keep all the first and second order attribute items. In Eq. (6), C_{ij} is an attribute from the attribute vector $\boldsymbol{C}$, $C_{im} \times C_{in}$ is an attribute interaction (combination) item that is treated as a common linear attribute in coefficient estimations.

$$\text{logit}(\hat{P}r_{ik}) = \beta_0 + \sum_{j=1}^{p} \beta_j C_{ij} + \sum_{m=1}^{p} \beta_{mn} C_{im} \sum_{n=m+1}^{p} C_{in} \tag{6}$$

The stepwise regression method introduced here can select the most important attributes and the attribute interactions by an iterative training process as shown in Fig.1.

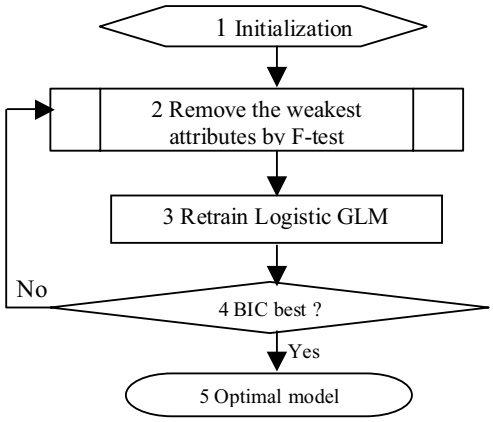

Fig.1: Flowchart of Stepwise regression

This training is an off-line process. We can obtain BIC "best" model for a given corpus. For example, suppose that phrase break is only affected by attributes "POS" (C_1) and "word length" (C_2), note the model as in Eq.(7).

$$\text{logit}(\hat{P}r_i) = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_{12}(C_1 \times C_2) \tag{7}$$

Where $C_1 \times C_2$ means the combination of POS and word length. Eq.(7) is the initial model. Then we calculate F-test values of each item, maybe $C_1 \times C_2$ is the least important item, if so, we remove it. Now we retrain the model in Eq.(7). Then we calculate the BIC, maybe now the BIC is minimized, if so, we can stop here and get the optimal model in Eq.(8).

$$\text{logit}(\hat{P}r_i) = \beta_0 + \beta_1 C_1 + \beta_2 C_2 \tag{8}$$

2. EXPERIMENTS

3.1 Corpus

The model described above is trained and tested using a 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 textual and linguistic information, such as Chinese word segmentation boundaries, POSs, and acoustic information. The phrase breaks are initialized as the segmental labels for silence that are generated by the force-alignment technology in speech recognition, then are manually corrected.

The total number of word boundaries in the corpus is about 24,700. Among them, the total number of phrase break boundaries is about 4,900. About 19.83% word boundaries are phrase break boundaries.

Theoretically, all linguistic and phonetic attributes around phrase break boundaries are likely to influence phrase break. For a certain phrase break P_i after word w_i in Eq.(9), the word lengths, POSs, etc. of current word $w_{i,}$ the preceding 2 words

w_{i-1}, w_{i-2}, the subsequent words w_{i+1}, w_{i+2} are combined as input attributes. The contextual window length is set to 5, which is similar with other researches [2]. The first Initial after the word boundary and the last Final before the word boundary are included in the attribute set. This is our basic attribute set (BASATTR).

$$...,(P_{i-2},C_{i-2}),(P_{i-1},C_{i-1}),(P_i,C_i),(P_{i+1},C_{i+1}),(P_{i+2},C_{i+2}),.... \quad (9)$$

On the other hand, we perceived that the speaking rate of the corpus varies to some extent. The speaking rate is defined as the 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 sharp. In this paper, another attribute set (SPATTR) including BASATTR and speaking rate is also evaluated.

As we know, the six Chinese punctuation marks, comma, colon, semi-colon, period, exclamation mark and question mark (“，：；。！？”) directly cause phrase breaks in speech. In order to get extra punctuation information of the larger corpus, we use some Chinese text with word segmentations from the 1998 People’s Daily corpus. We count the frequencies of each word separated by these punctuation marks, and only keep the top 1,000 high-frequency words. The word frequency is combined with BASATTR, which results in another attribute set (WFATTR).

3.2 Evaluation metrics

As for objective evaluation, we measure the goodness of fit in term of precision, recall and f-score in this paper, which are often used in evaluation of phrase break modeling [2][6] and defined as follows.

$$precsion = \frac{number\ of\ correctly\ predicted\ pauses}{number\ of\ predited\ pauses}$$
$$recall = \frac{number\ of\ correctly\ predicted\ pauses}{number\ of\ correct\ pauses} \quad (10)$$
$$F-score = \frac{2 \times precision \times recall}{precision + recall}$$

F-score gives a balance between precision and recall, thus we treat it as the main objective evaluation metric.

3.3 Experiment results

Our experiments in this paper are performed using 75% of the corpus for training and the other 25% part for testing.

We also apply C4.5 and CART methods for phrase break prediction for comparison with GLM-based methods. The three attribute sets, BASATTR, SPATTR and WFATTR are tested. Table 1 shows the recalls, precisions and F-scores.

Table 1: Performance of different phrase break models

Method	Attribute	Precision	Recall	F-score
C4.5	WFATTR	0.4923	0.6606	0.5642
CART	WFATTR	0.5457	0.6176	0.5795
Plain GLM	BASATTR	0.5757	0.6703	0.6194
Logistic GLM	BASATTR	0.5486	0.7212	0.6232
Logistic GLM	SPATTR	0.5452	0.7324	0.6251
Logistic GLM	WFATTR	0.5844	0.6923	0.6337

From Table1, we can see the GLM-based methods are better than decision-tree based methods C4.5 and CART. The best F-score made by Logistic GLM is 0.6337. This F-score is about 5.42% higher than that of CART and 6.95% higher than that of C4.5. The speaking rate and the word frequency bring some improvement in phrase break prediction.

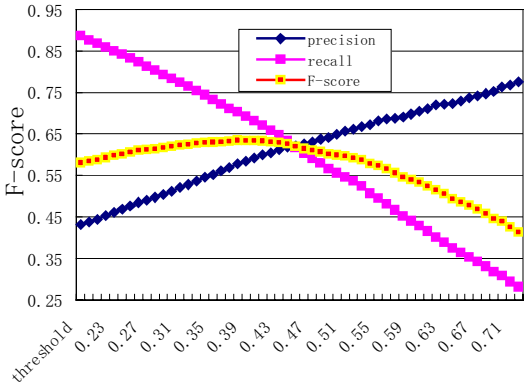

Fig.2: Performances of Logistic GLM with different threshold values

When predicting, if the predicted probabilities of word boundaries are larger than a given threshold, we take them for phrase break boundaries, otherwise they are just word boundaries. Fig.2 illustrates the performance of attribute set WFATTR under different thresholds.

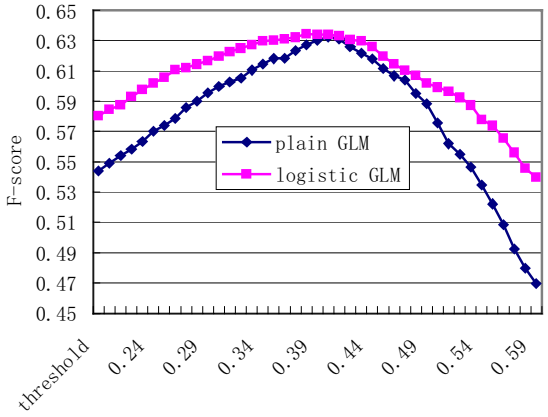

Fig.3: Performances of plain GLM and logistic GLM with different threshold values

Fig.3 compares the performance of plain GLM and Logistic GLM under different thresholds. Again, the attribute set WFATTR is adopted. We can see the F-score curve of Logistic GLM is always above that of plain GLM. And the curve of Logistic GLM is much flatter than that of plain GLM.

Therefore, Logistic GLM is more suitable for phrase break prediction than plain GLM.

Table 2: Comparing results with Li's approaches

Method	Utterance size	F-score	Relative improvement
Proposed DT	2,150	0.5642	---
Proposed GLM	2,150	0.6232	10.4%
Li's DT [2]	42,000	0.5962	---
Li's ME1[2]	42,000	0.6291	5.5% [2]

Table 2 presents comparison results with Li's approaches [2]. Both our methods and Li's methods include no punctuation information in the training attributes and are to predict phrase break for Mandarin TTS. In table 2, DT means decision tree (C4.5) method and ME means maximum entropy model, the proposed DT represents the C4.5 method in Table 1., the proposed GLM represents the Logistic GLM with our basic attribute set (BASATTR). Li's ME1 method also uses only POS and word length attributes that similar to our BASATTR. From table 2 we can see, our GLM method makes a relative improvement of 10.4% over CART, while Li's ME1 makes a relative improvement of 5.5% over DT. Nevertheless, we believe the proposed method achieves as good F-score as Li's ME1 approach in about 1/20 data size. The numerical results are inspiring.

A few differences still exist among corpus, attributes and the test conditions between our methods and Li's methods, therefore the comparing results should be taken care of. Li also improved his ME1 method by using additional word-dependent lexical attributes (ME2 in [2]) and post-processing technique after directly predicting (MES in [8]), which indicates us to collect larger corpus and use more detailed attributes and more complex techniques. An advantage of our GLM model is compactness, since it contains only about 160 coefficients.

4. CONCLUSIONS

In this paper, we propose a novel logistic GLM based phrase break prediction approach under Bernoulli hypothesis of phrase break. It's suitable and portable for phrase break prediction. Furthermore, we bring up stepwise regression for attribute selection based on BIC. As we shown, the whole modeling method proposed in this paper is a totally data driven method and make an improvement of 5.4% over CART.

We expect that using more linguistic and lexical information will improve the performance of the phrase break prediction. Some post-processing methods, such as sliding window smoothing [8] after predicting, may also contribute to performance.

5. REFERENCES

[1] Zhao Sheng, Tao Jianhua, Cai Lianhong, "Prosodic Phrasing with Inductive Learning", in *Proc. of ICSLP2002* Denver, USA, pp.2417-2420.

[2] Jian-feng Li, Guo-ping Hu, Wan-ping Zhang, Ren-hua Wang. "Chinese Prosody Phrase Break Prediction Based on Maximum Entropy Model", in *Proceedings* ICSLP 2004, Oct 4-8, Korea, pp. 729-732.

[3] Sun, X. and Applebaum, T.H., "Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model", in *Proceedings Eurospeech2001*, Denmark, Vol 1, pp. 537-540.

[4] Bertjan Busser, Walter Daelemans, Antal van den Bosch, "Predicting phrase breaks with Memory-Based Learning", *Proceedings 4th. ISCA Tutorial and research Workshop on Speech Synthesis,* Perthshire Scotland, 2001.

[5] Alan W. Black and Paul A. Taylor, "Assigning phrase breaks from part-of-speech sequences", in *Proceeding Eurospeech1997*, Rhodes, Greece, Vol 2, pp. 995-998.

[6] Veilleux N.M., Ostendorf M., Price P.J., et al. "Markov Modeling of Prosodic Phrase Structure", in *Proceeding of the 1990 International Conference on Acoustics,* Speech and Signal Processing, 1990, Vol 2.

[7] Antonio Bonafonte, Pablo Daniel Agüero, "Phrase Break Prediction Using a Finite State Transducer", *in 11th International Workshop "Advances in Speech Technology 2004",* Maribor, Slovenia. July 2004.

[8] Jian-Feng Li, Guo-Ping Hu, Ren-Hua Wang, Li-Rong Dai, "Sliding Window Smoothing For Maximum Entropy Based Intonational Phrase Prediction In Chinese", in *Proceeding of ICASSP2005*, Philadelphia, PA, USA, pp. 285-288.

[9] McCullagh P and Nelder JA, *Generalized Linear Models*, Chapman & Hal, London, 1989.