An HMM-Based Approach to Automatic Phrasing for Mandarin Text-to-Speech Synthesis

Jing Zhu
Department of Electronic Engineering
Shanghai Jiao Tong University
zhujing@sjtu.edu.cn

Jian-Hua Li
Department of Electronic Engineering
Shanghai Jiao Tong University
lijh888@sjtu.edu.cn

Abstract

Automatic phrasing is essential to Mandarin text-to-speech synthesis. We select word format as target linguistic feature and propose an HMM-based approach to this issue. Then we define four states of prosodic positions for each word when employing a discrete hidden Markov model. The approach achieves high accuracy of roughly 82%, which is very close to that from manual labeling. Our experimental results also demonstrate that this approach has advantages over those part-of-speech-based ones.

1 Introduction

Owing to the limitation of vital capacity and contextual information, breaks or pauses are always an important ingredient of human speech. They play a great role in signaling structural boundaries. Similarly, in the area of text-to-speech (TTS) synthesis, assigning breaks is very crucial to naturalness and intelligibility, particularly in long sentences.

The challenge in achieving naturalness mainly results from prosody generation in TTS synthesis. Generally speaking, prosody deals with phrasing, loudness, duration and speech intonation. Among these prosodic features, phrasing divides utterances into meaningful chunks of information, called hierarchic breaks. However, there is no unique solution to prosodic phrasing in most cases. Different solution in phrasing can result in different meaning that a listener could perceive. Considering its importance, recent TTS research has focused on automatic prediction of prosodic phrase based on the part-of-speech (POS) feature or syntactic structure(Black and Taylor, 1994; Klatt, 1987; Wightman, 1992; Hirschberg 1996; Wang, 1995; Taylor and Black, 1998).

To our understanding, POS is a grammar-based structure that can be extracted from text. There is no explicit relationship between POS and the prosodic structure. At least, in Mandarin speech synthesis, we cannot derive the prosodic structure from POS sequence directly. By contrast, a word carries rich information related to phonetic feature. For example, in Mandarin, a word can reveal many phonetic features such as pronunciation, syllable number, stress pattern, tone, light tone (if available) and retroflexion (if available) etc. So we begin to explore the role of word in predicting prosodic phrase and propose a word-based statistical method for prosodic-phrase grouping. This method chooses Hidden Markov Model (HMM) as the training and predicting model.

2 Related Work

Automatic prediction of prosodic phrase is a complex task. There are two reasons for this conclusion. One is that there is no explicit relationship between text and phonetic features. The other lies in the ambiguity of word segmentation, POS tagging and parsing in the Chinese natural language processing. As a result, the input information for the prediction of prosodic phrase is quite “noisy”. We can find that most of published methods, including (Chen et al., 1996; Chen et al., 2000; Chou et al., 1996; Chou et al., 1997; Gu et al., 2000; Hu et al., 2000; Lv et al., 2001; Qian et al., 2001; Ying and Shi, 2001) do not make use of high-level syntactic features due to two reasons. Firstly, it is very challenging to parse Chinese sentence because no grammar is formal enough to be applied to Chinese parsing. In addition, lack of

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 977–982,
Sydney, July 2006.

morphologies also causes many problems in parsing. Secondly, the syntactic structure is not isomorphic to the prosodic phrase structure. Prosodic phrasing remains an open task in the Chinese speech generation. In summary, all the known methods depend on POS features more or less.

3 Word-based Prediction

As noted previously, the prosodic phrasing is associated with words to some extent in Mandarin TTS synthesis. We observe that some function words (such as "的") never occur in phrase-initial position. Some prepositions seldom act as phrase-finals. These observations lead to investigating the role of words in prediction of prosodic phrase. In addition, large-scale training data is readily available, which enables us to apply data-driven models more conveniently than before.

3.1 The Model

The sentence length in real text can vary significantly. A model with a fixed-dimension input does not fit the issue in prosodic breaking. Alternatively, the breaking prediction can be converted into an optimization problem that allows us to adopt the hidden Markov model (HMM).

An HMM for discrete symbol observations is characterized by the following:

- the state set $Q=\{q_i\}$, where $1 \le i \le N$, N is the number of states
- the number of distinct observation symbol per state M
- the state-transition probability distribution $A=\{a_{ij}\}$, where

$$a_{ij}=P[q_{t+1}=j|q_t=i],\ 1 \le i,j \le N$$

- the observation symbol probability distribution $B=\{b_j(k)\}$, where

$$b_j(k) = P[o_t = v_k \mid q_t = j],$$

$1 \le i,j \le N$

- the initial state distribution $\pi=\{\pi_i\}$, where $\pi_i = P[o_t=v_k|q_t=j]$, $1 \le i,j \le M$.

The complete parameter set of the model is denoted as a compact notation $\lambda=(A,B,\pi)$.

Here, we define our prosodic positions for a word to apply the HMM as follows.

0 phrase-initial
1 phrase-medial
2 phrase-final
3 separate

This means that Q can be represented as $Q=\{0,1,2,3\}$, corresponding to the four prosodic positions. The word itself is defined as a discrete symbol observation.

3.2 The Corpus

The text corpus is divided into two parts. One serves as training data. This part contains 17,535 sentences, among which, 9,535 sentences have corresponding utterances. The other is a test set, which includes 1,174 sentences selected from the Chinese *People's Daily*. The sentence length, namely the number of words in a sentence varies from 1 to 30. The distribution of word length, phrase length and sentence length(all in character number) is shown in Figure 1.

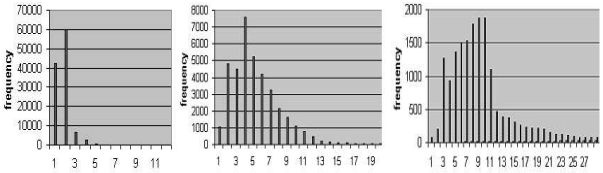

Figure 1. Statistical results from the corpus

In a real text, there may exist words that are difficult to enumerate in the system lexicon, called "non-standard" words (NSW). Examples of NSW are proper names, digit strings, derivative words by adding prefix or suffix.

Proper names include person name, place name, institution name and abbreviations, etc. Alternatively, some characters are usually viewed as prefix and suffix in Chinese text. For instance, the character 伪 (pseudo-) always serves as a prefix, while another character 般 (-like) serves as a suffix. There are 130 analogous Chinese characters have been collected roundly. A word segmentation module is designed to identify these non-standard words.

3.3 Parameter estimation

Parameter estimation of the model can be treated as an optimization problem. The parametric methods will be optimal if distribution derived from the training data is in the class of distributions being considered. But there is no

known way so far for maximizing the probability of the observation sequence in a closed form. In the present approach, a straightforward, reasonable yet, method to re-estimate parameters of the HMM is applied. Firstly, statistics for the occurring times of word, prosodic position, prosodic-position pair are conducted. Secondly, the simple ratio of occurring times is used to calculate the probability distribution. The following expressions are used to implement calculations,

State probability distribution

$$P[q_i] \approx \frac{F_i}{\sum_{j=1}^{N} F_j} \quad , \quad 1 \le i \le N$$

F_i is the occurring times of state q_i

the state-transition probability distribution $A = \{a_{ij}\}$,

$a_{ij} \approx \frac{F_{ij}}{F_i}$, $1 \le i, j \le N$, F_{ij} is the occurring times of state pair (q_i,q_j).

Observation probability *distribution* $B = \{b_j(k)\}$,

$$b_j(k) \approx \frac{F(q = j, o = v_k)}{F_j}$$

$$b_j(k) \propto \frac{F(q = j, o = v_k)}{P[q_j]}$$

where $F(q = j, o = v_k) = \sum_t F(q_t = j, o = v_k)$ is the concurring times of state q_j and observation v_k.

With respect to the proper names, all the person names are dealt with identically. This is based on an assumption that the proper names of individual category have the same usage.

3.4 Parameter adjustment

Note that the training corpus is discrete, finite set. The parameter set resulting from the limited samples cannot converge to the "true" values with probability. In particular, some words may not be included in the corpus. In this case, the above expressions for training may result in zero valued observation-probability. This, of course, is unexpected. The parameters should be adjusted after the automatic model training. The way is to use a sufficiently small positive constant ε to represent the zero valued observation-probabilities.

3.5 The search procedure

In this stage, an optimal state sequence that explains the given observations by the model is searched. That is to say, for the input sentence, an optimal prosodic-position sequence is predicted with the HHM. Instead of using the popular *Viterbi* algorithm, which is asymptotically optimal, we apply the Forward-Backward procedure to conduct searching.

Backward and forward search

All the definitions described in (Rabiner, 1999) are followed in the present approach.

The forward procedure

forward variable: $\alpha_t(i) = P(o_1 o_2 \cdots o_t, q_t = i \mid \lambda)$

initialization: $\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N.$

induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \quad 1 \le t \le T-1, \quad 1 \le j \le N \ .$$

termination: $P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$

where T is the number of observations.

The backward procedure

backward variable:

$\beta_t(i) = P(o_{t+1} o_{t+2} \cdots o_T \mid q_t = i, \lambda)$

initialization $\beta_T(i) = 1, \quad 1 \le i \le N$

induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad t = T-1, T-2, \ldots 1, \quad 1 \le i \le N$$

The "optimal" state sequence

posteriori probability variable: $\gamma_t(i)$, this is the probability of being in state i at time t given the observation sequence O and the model λ. It can be expressed as follows:

$$\gamma_t(i) = P(q_t = i \mid O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

most likely state q_t^* at time t:

$$q_t^* = \arg \max_{1 \le i \le N} [\gamma_t(i)] \quad 1 \le t \le T \ .$$

Here comes a question. It is, whether the optimal state sequence means the optimal path.

Search based on dynamic programming

The preceding search procedure targets the optimal state sequence satisfying one criterion. But it does not reflect the probability of occurrence of sequences of states. This issue is explored based on a dynamic programming (DP) like approach, as described below.

For convenience, we illustrate the problem as shown in Figure 2.

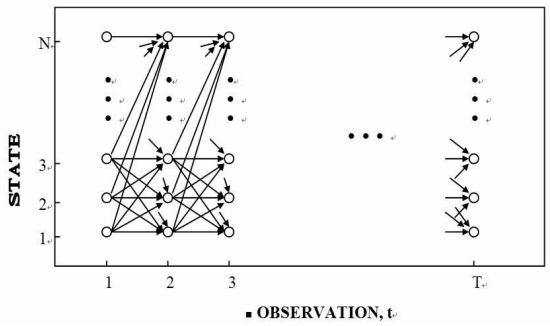

Figure 2. Illustration of search procedure in trellis (quoted from [Rabiner, 1999])

From Figure 2, it can be seen that the transition from state *i* to state *j* only occurs in the two consecutive stages, namely time synchronous. Totally, there are *T* stages, N^2T arcs. Therefore, the optimal-path issue is a multi-stage optimization problem, which is similar to the DP problem. The slight difference lies in that a node in the conventional DP problem does not contain any additional attribute, while a node in HMM carries the attribute of observation probability distribution. Considering this difference, we modify the conventional DP approach in the following way.

In the trellis above, we add a virtual node (state), where the start node q_s corresponding to time 0 before time 1. All the transitions from q_s to nodes in the first stage (time 1) equal to *1/N*. Furthermore, all the observation probability distributions equal to *1/M*. Denoting the optimal path from q_s to the node q_i of time *t* as *path(t,i)*, *path*(*t,i*) is a set of sequential states. Accordingly, we denote the score of *path(t,i)* as *s(t,i)*. Then, *s(t,i)* is associated with the state-transition probability distribution and observation probability distribution. We describe the induction process as follows.

initialization:

$$s(0,i) = \frac{1}{M \times N}, \quad 1 \le i \le N$$
$$path(0,i) = \{q_s\}.$$

induction:

given

$$j,\ s(t,j) = \max_{1 \le i \le N} [s(t-1,i) \times b_i(o_t) \times a_{ij}],\ 1 \le t \le T \quad ,$$

denotes

$$k = \arg \max_{1 \le i \le N} [s(t-1,i) \times b_i(o_t) \times a_{ij}] \quad ,$$ then

$path(t,j) = path(t-1,k) \cup \{k\}$.

termination:

at time T, $k = \arg \max_{1 \le i \le N} s(T,i)$.

then $path(T,k) - \{q_s\}$ *is the optimal path.*

Basically, the main idea of our approach lies in that if the final optimal path passes a node *j* at time *t*, it passes all the nodes in *path*(*t,j*) sequentially. This idea is similar to the forward procedure of DP. We can begin with the termination *T* and derive an alternative approach. As for time complexity, the above trellis can be viewed as a special DAG. The state transition from time *t* to time *t+1* requires $2N^2$ calculations, resulting in the time complexity $O(TN^2)$.

Intuitively, the optimal path differs from the optimal state sequence generated by the Forward-Backward procedure. The underlying idea of Forward-Backward procedure is that the target state sequence can explain the observations optimally. To support our claim, we can give a simple example ($T=2$, $N=2$, $\pi = [0.5, 0.5]^T$) as follows:

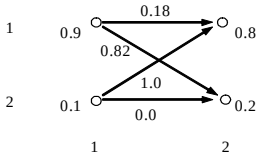

Figure 3. Optimal state sequence vs. optimal path

Apparently, the optimal state sequence is (1,1), while the optimal path is {1,2}.

4 Experimental Results

Before reporting the experimental results, we first define the criterion of evaluation and the related issues.

4.1 The evaluation method

After analyzing the existing evaluation methods, we feel that the method proposed in (Taylor and Black, 1998) is appropriate for our application. By employing this method, we can examine each word pair in the test set. If the algorithm generated break fully matches the manually labeled break, it marks correct. Similarly, if there is no labeled break and the algorithm does not place a break, it also marks correct. Otherwise, an error arises. To emphasize the effectiveness of break prediction, we define the adjusted score, S_a, as follows.

$$S_a = \frac{S - B}{1 - B}$$

where

S is the ratio of the number of correct word pairs to the total number of word pairs;

B is the ratio of non-breaks to the number of word-pairs.

4.2 The test corpora

From the perspective of perception, multiple predictions of prosodic phrasing may be acceptable in many cases. At the labeling stage, three experts (*E1, E2, E3*) were requested to label 1,174 sentences independently. Experts first read the sentences silently. Then, they marked the breaks in sentences independently. Table 1 and 2 show their labeling differences in terms of *S* and S_a, respectively.

	E1	*E2*	*E3*
E1	*1.00*	*0.87*	*0.87*
E2	*0.87*	*1.00*	*0.86*
E3	*0.87*	*0.86*	*1.00*

Table 1. Three experts' matching scores

	E1	*E2*	*E3*
E1	*1.00*	*0.74*	*0.67*
E2	*0.74*	*1.00*	*0.66*
E3	*0.72*	*0.72*	*1.00*

Table 2. Three experts' adjusted matching scores

Table 1 indicates that any two can achieve a consistency of roughly 87% among three experts.

4.3 The results

To evaluate the approaches mentioned above, we conducted a series of experiments. In all our experiments, we assume that no breaking is necessary for those sentences that are shorter than the average phrase length and remove them in the statistic computation. For the approaches based on HMM path, we further define that the initial and final words of a sentence can only assume two state values, namely, (*phrase initial, separate*) and (*phrase final, separate*), respectively. With this definition, we modify the approach *HMM-Path* to *HMM-Path-I*. Alternatively, to investigate acceptance, we also calculate the matching score between the approaches and *any* expert (We assume the prediction is acceptable if the predicted phrase sequence matches any of three phrase sequences labeled by the experts). By employing the preceding criterion, we achieve the results as shown in Table 3 and 4.

	E1	*E2*	*E3*	*Any*
HMM	*0.78*	*0.77*	*0.77*	*0.85*
HMM-path	*0.79*	*0.77*	*0.78*	*0.85*
HMM-path-I	*0.82*	*0.80*	*0.82*	*0.88*

Table 3. Matching scores of 3 approaches

	E1	*E2*	*E3*	*Any*
HMM	*0.55*	*0.53*	*0.44*	*0.66*
HMM-path	*0.52*	*0.54*	*0.44*	*0.67*
HMM-path-I	*0.62*	*0.60*	*0.55*	*0.74*

Table 4. Adjusted matching scores of 3 approaches

A sentence consumes less than 0.3 ms on average for all the evaluated methods. So they are all computationally efficient. Alternatively, we compared the HMM-based approach base on word format and some POS-based ones on the same training set and test set. Overall, *HMM-path*-I can achieve high accuracy by about 10%.

5 Conclusions/Discussions

We described an approach to automatic prosodic phrasing for Mandarin TTS synthesis based on word format and HMM and its variants. We also evaluated these methods through experiments and demonstrated promising results. According to the experimental results, we can conclude that word-based prediction is an effective approach and has advantages over the POS-based ones. It confirms that the syllable number of a word has substantial impact on prosodic phrasing.

References

Black, A.W., Taylor, P., 1994. "Assigning intonational elements and prosodic phrasing for

English speech synthesis from high level linguistic input", *Proc. ICSLIP*

Chen, S.H., Hwang, S.H., Wang, Y.R., 1998. "An RNN-based prosodic information synthesizer for Mandarin text-to-speech", *IEEE Trans. Speech Audio Processing*, 6: 226-239.

Chen, Y.Q., Gao, W., , Zhu, T.S., Ma, J.Y., 2000. "Multi-strategy data mining on Mandarin prosodic patterns", *Proc. ISCLIP*

Chou, F.C., Tseng, C.Y., Lee, L.S. 1996. "Automatic generation of prosodic structure for high quality Mandarin speech synthesis", *Proc. ICSLP*

Chou, F.C, Tseng, C.Y, Chen, K.J., Lee, L.S, 1997. "A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and non-uniform units", ICASSP'97

Klatt, D.H., 1987, "Review of text-to-speech conversion for English", *J. Acoust. Soc. Am.*, 182: 737-79

Gu, Z.L, Mori, H., Kasuya, H. 2000. "Prosodic variation of focused syllables of disyllabic word in Mandarin Chinese", *Proc. ICSLP*,

Hirschberg, J., 1996. "Training intonational phrasing rules automatically for English and Spanish text-to-speech", *Speech Communication*, 18:281-290

Hu, Y., Liu, Q.F., Wang, R.H., 2000, "Prosody generation in Chinese synthesis using the template of quantified prosodic unit and base intonation contour", *Proc. ICSLIP*

Lu, S.N., He, L., Yang, Y.F., Cao, J.F., 2000, "Prosodic control in Chinese TTS system", *Proc. ICSLP*,

Lv, X., Zhao, T.J., Liu, Z.Y., Yang M.Y., 2001, "Automatic detection of prosody phrase boundaries for text-to-speech system", *Proc. IWPT*

Qian, Y., Chu, M., Peng, H., 2001, "Segmenting unrestricted Chinese text into prosodic words instead of lexical words", *Proc. ICASSP*.

Rabiner, L., 1999, *Fundamentals of Speech Recognition*, pp.336, Prentice-Hall and Tsinghua Univ. Press, Beijing

Taylor P., Black A.W., 1998, "Assigning phrase breaks from part-of-speech sequences", *Computer Speech and Language*, 12: 99-117,

Wang, M.Q., Hirschberg, J., 1995, "Automatic classification of intonational phrase boundaries", *Computer Speech and Language*, pp.175-196, Vol. 6,

Wightman, C.W., 1992, "Segmental durations in the vicinity of prosodic phrase boundaries", *J. Acoust. Soc. Am.*, 91:1707-1717

Ying, Z.W., Shi, X.H., 2001, "An RNN-based algorithm to detect prosodic phrase for Chinese TTS", *Proc. ICASSP*