AUTOMATIC PROSODY LABELING USING BOTH TEXT AND ACOUSTIC INFORMATION

Xijun Ma , Wei Zhang , Qin Shi , Weibin Zhu and Liqin Shen

IBM China Research Lab
{maxijun,zhangzw,shiqin,zhuweib,shenlq}@cn.ibm.com

ABSTRACT

Prosody is an important factor for a high quality text-to-speech (TTS) system. The prosody is often described with a hierarchical structure. So the generation of the hierarchical prosody structure is very important both in the corpus building and the run-time text analysis. But the prosody labeling procedure is laborious and time consuming. Moreover, to keep the consistence between different labelers and even the same labeler in different time is difficult. In this paper an automatic prosody labeling system is presented, in which the decision tree plus Viterbi decoding framework proposed in [1] is used. In the system, not only the acoustic information but also the text information such as the part-of-speech (POS) of a word is used. A prosody model is built up using the automatically labeled corpus for our Mandarin TTS system. Listening test shows that the automatic prosody labeling system works pretty well.

1. INTRODUCTION

Prosody is an important factor that makes the speech generated by a TTS system more natural and understandable. It usually can be revealed by a hierarchical prosody structure. A corpus with precise prosody labels is very useful. Such a corpus [2] has been manually labeled and based on it we has got a state-of-the-art concatenative Mandarin TTS system. For our TTS system, both the prosody prediction model and the front-end prosody structure analysis model [3] are trained from the labeled corpus. However, the manually prosody labeling is laborious and time-consuming. Moreover, to keep the consistence between different labelers and even the same labeler in different time is some difficult. So automatic prosody labeling attracts more and more attention now. Some research work had been done in this field [4][5]. In this paper an automatic prosody labeling system is proposed that attempts to speed up the corpus building process.

The framework for an automatic prosody labeling system was proposed in [1], where the decision tree plus Viterbi decoding were used. The difference here is that not only the acoustic parameters but also the text information such as part-of-speech information are used.

The paper is organized as follows. In section 2, our prosody labeled corpus is introduced and the task for the automatic labeling is described. In section 3, the prosody labeling method is presented. The experiment will be reported in section 4. And the conclusion will be drawn finally.

2. IBM MANDARIN TTS CORPUS

IBM Mandarin TTS Female Corpus includes about 20,000 sentences considering all kinds of coverage. The detailed information about the corpus building is reported in [2][6]. In this experiment, 1000 sentences are used for training and 5255 sentences for testing.

2.1. Prosody Structure

Prosody labeling is used to describe the prosodic events. ToBI (Tones and Break Indices) and ToBI like systems have been used in several languages [7]. However, the reliability among labelers for some ToBI categories is low. Even though, we can still start from ToBI to form our prosody labeling system.

We describe the prosody structure into 4 hierarchical layers:

- Foot (FT)
- Prosody Word (PW)
- Prosody Phrase (PP) or Intermediate Phrase
- Intonation Phrase (IP)

Their relationships are showed in the Figure 1.

For an example, a sentence could be labeled like this according to the actual speech:

虽说(PW) 应用(FT) 界面(PP) 不如(FT) 中文(FT)
之星(PW) 丰富，(IP) 但(PW) 正在(PP) 抓紧(PW)
完善，(IP) 一定(FT) 能(PP) 后来(FT) 居上．(IP)

When the corpus is labeled, one principle we used is to label according to what you hear. The perceptive cues for judgment include:

- Pitch reset
- Pause
- Duration lengthening

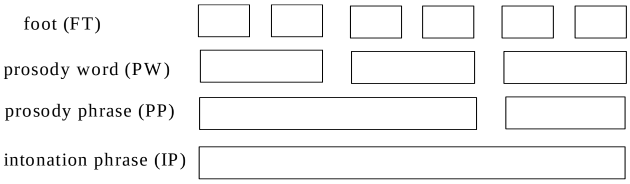

Figure 1: Prosody Structure

2.2. Task of Automatic Prosody Labeling

Given the input text and its corresponding speech, our labeling task is to generate these prosodic events. Usually, we have word segmentation and tagging tools to parse the input text and get lexical words and their POSes. So in the actual system, our starting point is the lexical words and their POSes. Practically the prosody word boundary is placed at a lexical word boundary. In other words, one or more lexical words group together to form a prosody word. After the other prosody layers (prosody word, prosody phrase and intonation phrase) are done, foot layer can be established based on some rules. So our main task is to label the prosody layers including PW, PP, IP given lexical words with their POSes and the speech data.

3. THE LABELING ALGORITHM

In the labeling system, the framework proposed in [1] is followed, using the decision tree plus Viterbi decoding.

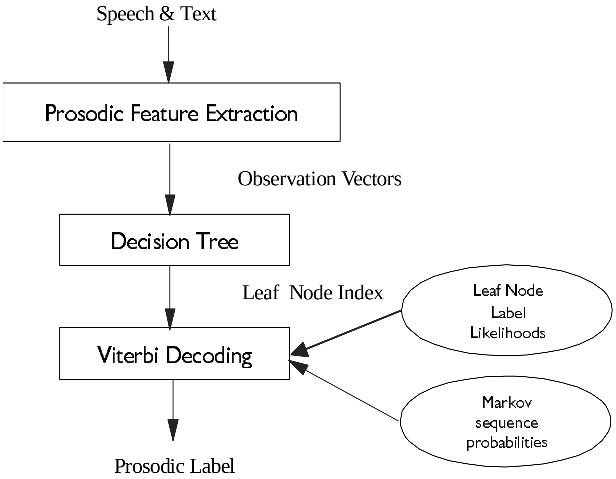

Figure 2: Block diagram for the basic labeling algorithm

3.1. General Algorithm

Our task is to map a sequence of feature vectors $X_1^n = \{x_1, ..., x_n\}$ to a sequence of prosodic labels $\alpha_1^n = \{\alpha_1, ..., \alpha_n\}$. For a Markov or simple bigram model,

$$p(\alpha_1, ..., \alpha_n) = p(\alpha_1)\prod_{i=2}^{n} p(\alpha_i \mid \alpha_{i-1})$$

Combining the decision tree with the Markov sequence assumption gives the overall probabilistic model of the form:

$$p(x_1, ..., x_n, \alpha_1, ..., \alpha_n)$$

$$= p(x_1 \mid \alpha_1) p(\alpha_1) \prod_{i=2}^{n} p(x_i \mid \alpha_i) p(\alpha_i \mid \alpha_{i-1}) \quad (1)$$

$$= [\prod_{j=1}^{n} p(x_j)] L(\alpha_1 \mid x_1) p(\alpha_1) \prod_{i=2}^{n} L(\alpha_i \mid x_i) p(\alpha_i \mid \alpha_{i-1}) \quad (2)$$

where

$$L(\alpha \mid x) = \frac{p(\alpha \mid x)}{p(\alpha)} = \frac{p(x \mid \alpha)}{p(x)}$$

$p(\alpha \mid x)$ is given by the decision tree, $p(\alpha)$ is the marginal probability of α , and $p(\alpha_i \mid \alpha_{i-1})$ is given by the Markov model. The term $\prod_{j=1}^{n} p(x_j)$ can be ignored when we just want to find the best prosodic label sequence. The goal is to choose the most likely label sequence:

$$\hat{\alpha}_1^n = \arg \max_{\alpha_1^n} p(\alpha_1^n, x_1^n) \quad (3)$$

$$= \arg \max_{\alpha_1^n} L(\alpha_1 \mid x_1) p(\alpha_1)$$

$$\cdot \prod_{i=2}^{n} L(\alpha_i \mid x_i) p(\alpha_i \mid \alpha_{i-1}) \quad (4)$$

The maximization problem (4) can be solved with a dynamic programming algorithm, as follows.

1) For all α , compute $L(\alpha,1) = \log[\ L(\alpha \mid x_1) p(\alpha)]$
2) For each time $i = 2, ..., n$ and all α , compute

$$L(\alpha, i) = \max_{\alpha'} \{\log[L(\alpha \mid x_i) p(\alpha \mid \alpha')] + L(\alpha', i-1)\}$$

$$prev(\alpha, i) = \arg\max_{\alpha'} \{L(\alpha', i-1) + \log[L(\alpha \mid x_i) p(\alpha \mid \alpha')]\}$$

3) $\hat{\alpha}_n = \arg\ \max_{\alpha} = L(\alpha, n)$
4) For each time $i = n, ..., 2$

$$\hat{\alpha}_{i-1} = prev\ (\hat{\alpha}_i, i)$$

3.2. Feature Vectors Extraction

Decision tree is used here to classify a feature vector x by asking a series of questions about the elements of x, and finally map it to a leaf node. The decision tree can provide the conditional probability distribution of the labels at the leaf nodes. As we know, the decision tree could ask heterogeneous questions. So besides the acoustic features extracted from speech data such as F0, duration and energy discontinuity, the text information is also used, such as the lexical word length and POSes of the lexical words before and after each lexical word boundary, the tone context and phone context of the boundary, etc.

Figure 3 shows the feature extraction procedure.

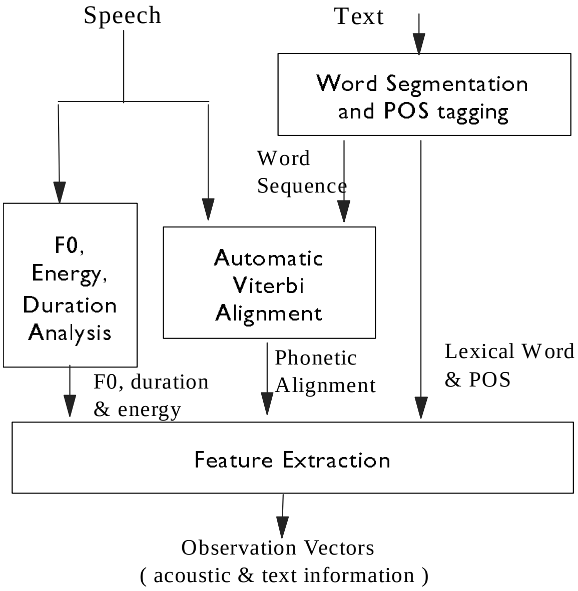
Speech
Text
Word Segmentation and POS tagging
Word Sequence
F0, Energy, Duration Analysis
Automatic Viterbi Alignment
Phonetic Alignment
Lexical Word & POS
F0, duration & energy
Feature Extraction
Observation Vectors (acoustic & text information)

Figure 3: Feature Vector Extraction

The text is prepared for the corpus. The automatic word segmentation and POS tagging tool is used to generate the word sequence with the phonetic alphabet (Pinyin being used for Chinese) and the POS information. The speech data recorded by a professional speaker and the phoneme sequence are used to do automatic phonetic alignment by using IBM speech recognition tools. Also an acoustic parameters analysis module is required to get the acoustic information of the speech.

Based on these work, the prosody feature vectors can be extracted. The feature extraction module gets the observation vectors at each lexical word boundary. It means that every lexical word boundary indicates an observation vector with an underling state (prosody layer).

An observation vector contains the following information:

- Acoustic: F0 reset, Duration Lengthing, Pause, Energy discontinuity, etc.
- Text: POS and length of the lexical word before and after this boundary, the phone and tone context of the boundary, etc.

3.3. Training and Testing

Four states corresponding to the prosodic layers (LW, PW, PP & IP) are defined. In the training procedure, 1000 sentences in the IBM Mandarin TTS Female Corpus are used. For each lexical word boundary in every sentence, the observation vector and its state can be obtained as described in 3.2. Note that the boundaries have an inclusion relation. The boundary greater than LW, such as PW, is also known as a LW boundary.

The prosodic labels are given in the manually labeled training data and therefore the states/labels are not hidden. So the states transition probabilities and states initial probabilities can be estimated from the training data directly.

After the training data collected and questions designed, a decision tree is built up using the greedy growing algorithm. The algorithm splits nodes according to which node optimizes the tree design criterion over all possible questions at that node.

To label new corpus automatically, the dynamic programming (Viterbi decoding) algorithm is used to get the best states/labels sequence given the observation vectors.

4. EXPERIMENTS & DISCUSSION

5255 sentences in IBM Mandarin TTS Female Corpus are automatically labeled using the system described above. These sentences are used to train a prosody model and a voice set for our Mandarin TTS system is built. The listening test is conducted to compare the subjective quality ratings for the automatically and manually labeled systems. The results show the voice quality using the automatically labeled corpus is a little better than that of the manually labeled system. Figure 4 gives out the MOS of the two systems. 15 people participate in the test and 9 pairs of testing sentences are used.

Several other experiments with different feature vector extraction methods are performed. Table 1,2 and 3 give out the confusion matrix of 3 experiments. In Table 2, only the text information is used. In Table 3, the POS information is not used. At all cases, the labeling results were not so good as using both the text and acoustic information. When the confusion matrix is calculated, the manually labeled sentences are used as reference.

From the confusion matrix, we notice that the confusion between PW and PP is large, but this should not always be seen as error. This is only the conformity between the manually label and automatic label. In many cases, both the manually and automatically labeled results are acceptable. And we also notice that the acoustic information currently used is more useful in differentiating large boundaries (PP & IP) and text information is more useful in differentiating small boundaries.

There is still lot of research work to do to find out which kind of acoustic parameters is more useful and how to normalize them. Precise acoustic parameters extraction and normalization are always some difficult.

5. CONCLUSIONS

In this paper, an automatic prosody labeling system is presented, where the decision tree plus Viterbi decoding method is used. Both the acoustic and the text information are used to build the decision tree. Listening test shows the system is workable, which means we can manually label a small amount of corpus, train an automatic labeling system and then label more corpus automatically based on that.

The following directions are currently being explored to improve the performance of the system:

1. Research on the feature vector extraction to find the features more suitable for labeling.
2. Research on the robustness of the labeling algorithm. Use the labeling model to label the corpus of a new speaker.
3. Try the unsupervised training method instead of the current supervised training method.

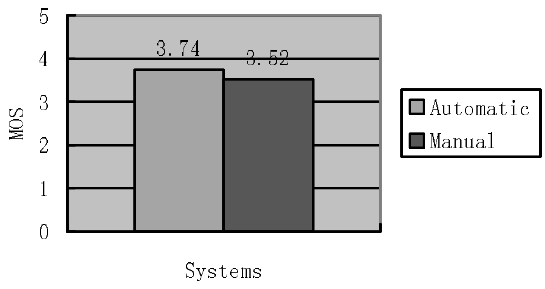

Figure 4: Mean Opinion Score for the 2 systems using automatically labeled corpus and manually labeled corpus

Automatic Manual	LW	PW	PP	IP
LW	8654	2333	496	25
PW	1456	27160	4273	654
PP	326	4496	6951	1437
IP	42	735	1249	8314

Table 1: Confusion matrix of the automatic labeling using both the text and acoustic information

Automatic Manual	LW	PW	PP	IP
LW	8729	2241	511	27
PW	1324	26816	4712	691
PP	377	4625	6764	1444
IP	61	821	1169	8289

Table 2: Confusion matrix of the automatic labeling using only the text information

Automatic Manual	LW	PW	PP	IP
LW	8035	2615	806	52
PW	1027	26938	4576	1002
PP	384	4837	6409	1580
IP	64	802	1122	8352

Table 3: Confusion matrix of the automatic labeling using the acoustic and text info. but no POS information

6. REFERENCES

[1] C.W. Wightman and M. Ostendorf, “Automatic Labeling of Prosodic Patterns,” Proc. ICASSP, October 1994.

[2] Weibin Zhu, Wei Zhang, Qin Shi, Fangxin Chen, “Corpus Building for Data-Driven TTS System,” 2002 IEEE TTS Workshop.

[3] Qin shi, Xijun Ma, Weibin Zhu, Wei Zhang and Liqin Shen, “Statistic Prosody Structure Prediction Based on Annotated Corpus,” 2002 IEEE TTS Workshop.

[4] C.W. Wightman, A.K. Syrdal, G. Stemmer, “Perceptually Based Automatic Prosody Labeling and Prosaically Enriched Unit Selection Improve Concatenative Text-to-Speech Synthesis,” Proc. ICSLP, Beijing, October 2000.

[5] A. Conkie, G. Ricardi and R.C. Rose, “Prosody Recognition From Speech Utterances Using Acoustic and Linguistic Based Models of Prosodic Events,” Proc. EUROSPEECH, September 1999.

[6] Haiping Li, Fangxin Chen, Liqin Shen, “The Context Variation Unit Vector,” Proc. ICSLP, Denver, 2002.

[7] K. Silverman. “ToBI: A Standard for Labeling English Prosody,” Proc. ICSLP, 1992.