AUTOMATIC GENERATION OF PROSODIC STRUCTURE FOR HIGH QUALITY MANDARIN SPEECH SYNTHESIS

Fu-chiang Chou[1], *Chiu-yu Tseng*[2] *and Lin-shan Lee*[1,3]
[1] Department of Electrical Engineering, National Taiwan University
[2] Institute of History and Philology, Academia Sinica
[3] Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
email-addr: moza@speech.ee.ntu.edu.tw

ABSTRACT

A key problem for today's speech synthesis technology is to automatically generate an appropriate hierarchical prosodic structure for text input and incorporate it into synthesized speech[1][2]. This paper presents a method for such a problem in Mandarin Chinese. This method uses a speech database for the training of a statistical model to generate the prosodic structure and determine prosodic parameters such as syllable duration, pause, energy and intonation. The experimental results show that an accuracy of 83.1% in the prediction of prosodic structure can be achieved. Furthermore, a Chinese text-to-speech system can be developed based on the proposed prosodic structure.

1. INTRODUCTION

One of the goals of current research in text-to-speech systems is to improve the naturalness of the speech output by developing algorithms for preprocessing texts in order to extract grammatical and prosodic information necessary for the generation of appropriate prosodic structure. Using parser to determine the prosodic structure for the text input has been reported[3][4]. However, precise syntactic analysis of the sentence structure is difficult and large amount of computation required, especially in Chinese. We therefore construct a statistical model to generate a hierarchical prosodic structure using Parts-of-Speech (POS's) as the main input features. This kind of prosodic structure can be equivalently represented by different phrase break markers. Four types of phrase breaks were used in our model: no break, minor break, major break, and punctuation mark break. All of these breaks were automatically labeled in the training phase using a tagged speech database. The labeled breaks and the POS's of the training corpus were then used to train a statistical model that can predict the prosodic structure of any input text.

The main purpose of the prosodic structure is to generate the necessary prosodic parameters for high quality Mandarin speech synthesis. Although many approaches have been studied in the past, it is difficult to incorporate high level linguistic information in the processing of the input text. A neural network model has been proposed by Hwang[5]. However, it is difficult to make a modification if any particular aspect of the speech output is unsatisfactory. The hierarchical prosodic structure mentioned above is proposed to deal with this problem. Different prosodic parameters can be determined on different level of this structure, including intonation, energy, tone sandhi, duration and pause.

In the following, we will describe the proposed method in three sections. First, we will describe the database used in our research and the labeling of the phrase breaks and the tagging of POS's. Second, training and testing of the proposed prosodic model will be illustrated. Finally, we will show how the prosodic parameters are determined according to the proposed prosodic structure.

2. DATABASE: LABELING AND TAGGING

A phonetically oriented speech database for Mandarin Chinese[6] was used in the experiments. The database includes two subsets: a word database (1455 words) and a sentence database (599 sentences). All of the data were recorded by 6 professional speakers at normal speaking rate. The word database was designed to incorporate all possible tonal combinations in Mandarin Chinese. We are able to extract the tone patterns and speech segments from this database. In addition, the sentence database including many longer sentences with natural intonation can be used to train the statistical model for the prediction of prosodic structure.

2.1. The Break Index Labeling

In our model, four types of the phrase breaks mentioned above were labeled at word boundaries with the break indices defined as follows: 0 for no break, 1 for minor break, 2 for major break, and 3 for punctuation mark break. No break means that no pause is inserted at word boundaries. The minor break is a very short pause between two words and the major break is a longer pause caused by respiratory effects. Meanwhile, punctuation mark break is labeled at the position of a punctuation mark. A hierarchical structure for prosodic boundaries can then be constructed using these different break indices.

These prosodic phrase breaks are associated with acoustic cues such as duration lengthening, pause insertion, and intonation markers. We designed a program for the alignment of the speech database to get the duration and pause information for labeling the break indices. At the same time, F_0 contours were extracted by the “get_f0” program in the ESPS package.

A Mandarin syllable can be conventionally decomposed into an INITIAL/FINAL format, INITIAL being the initial constant and

FINAL the vowel (or diphthong) part with an optional medial or a nasal ending. The break index at each boundary between two words can be determined primarily by three parameters: the pause duration, the length of the FINAL part of the preceding syllable, and the F_0 discontinuity. The first two parameters (pause duration and FINAL length) vary significantly for different cases (for example, some FINALs are longer and other shorter). Therefore, a normalization procedure is necessary in the analysis at this point. We defined the normalized length *n* as follows:

$$n = \frac{(d - \mu_p)}{\sigma_p} \qquad (1)$$

where *d* is the actual duration or length, μ_p and σ_p are the mean and standard deviation for the corresponding category. The value of *n* then provides very reliable information about the prosodic boundaries. The pause duration was categorized by the INITIAL of the following syllable, while the FINAL length simply by the FINALs themselves plus the tones. Mandarin Chinese is a tonal language and it is well known that the FINAL length is also dependent on the tones. The discontinuity of F_0 can be represented by the difference between the $ln(F_0)$ values of the two syllables enclosing the boundary. The log scale was chosen because of the perception factor around surround of human mechanism. The decision rules for the labeling of break index were trained by a small set of training data that was manually labeled. A typical result after the labeling of the break indices is shown in Figure 1.

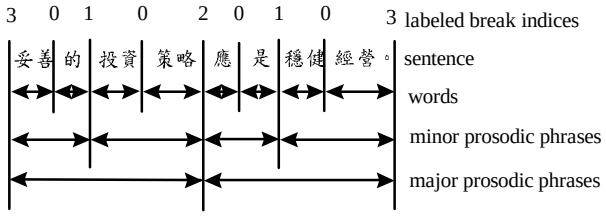

Figure 1: The prosodic structure of a Chinese sentence determined by the automatic break index labeling program.

2.2. Tagging System

A total of 44 Parts-of-Speech (POS's) primarily derived from previous syntactic analysis[7] was used in text analysis in the present study. Because all of the 44 POS's are not necessarily directly relevant to prosodic boundaries studied here, it is reasonable to classify these 44 POS's into smaller number of groups. This kind of classification can also increase the number of samples in each group in the statistical analysis described in section 3. Three different approaches for grouping these POS's were then considered. The first approach (syntactic grouping) used the syntactic knowledge from human experts for classification, and a total of 26 groups of POS's was derived. The second approach (text-corpus grouping) was based on the statistical behavior of the 44 POS's. In this approach, a feature vector was defined for each POS, whose components were the probabilities of all the preceding POS and following POS in the text of the speech database. These feature vectors were then vector quantized and clustered in the classification processes, and a total of 18 groups was obtained. Since both of the approaches did not use any prosodic information in the speech database, the third approach (speech-database grouping) attempted to use some prosodic information from the speech database. In the third approach, the means of the break indices at word boundaries measured from the speech database were used to construct the feature vectors of the POS's. The rest of the last approach is almost identical to the second approach, and a total of 18 groups was obtained.

3. TRAINING AND TESTING FOR THE PREDICTION OF PROSODIC STRUCTURE

3.1. Methods

After the POS's were classified as discussed above, the patterns for the groups of POS (for examples: Adj+N, Adv+V, and N+N+N, etc.) between two minor breaks in the speech database were collected and utilized to construct a prosodic phrase table. Frequencies of the occurrences were also included. These POS group patterns are called "minor prosodic phrases" for the moment.

At this stage, when a Chinese sentence is given, it is first segmented into words, tagged automatically with the POS groups obtained in section 2. Because there can be more than one way to segment the sentence into "minor prosodic phrases," a lattice of possible "minor prosodic phrases" can then be constructed using the prosodic phrase table obtained above. A typical example of such a lattice can be found in Figure 2. A dynamic programming procedure is then performed to determine the best path in the lattice based on the scores obtained from the frequencies of the "minor prosodic phrases." The longer "minor prosodic phrases" are preferred in this procedure and higher priority values are given with some special measures.

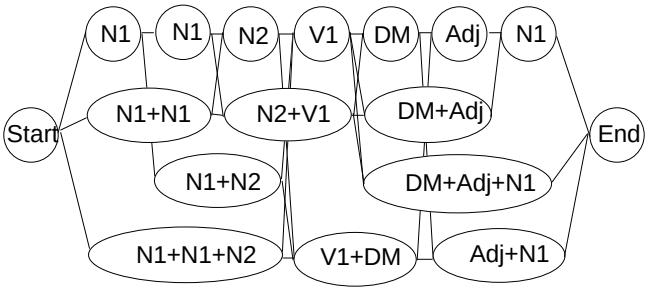

Figure 2: An example of "minor prosodic phrases" lattice.

After the "minor prosodic phrases" have been determined in this way, the probabilities that the boundary between two "minor prosodic phrases" is a major break are estimated in a similar way with a second dynamic programming procedure and associated frequency scores. The probabilities for numbers of constitution syllables between two major breaks are also estimated statistically. These probabilities are used to realize the length

constraints due to the respiratory effects, and have been integrated in the dynamic programming procedure.

3.2. Simulation Results

First, the effect of different POS grouping methods for the prediction of prosodic structure was compared (see Table 1). Five hundred sentences were used for training and ninety-nine sentences for testing. The experiment results in Table 1 show that the grouping method based on the information derived from the speech database achieved the highest accuracy. This means that prosodic phrases cannot be accurately predicted from the syntactic structure only. We found that prosodic phrase breaks do not always coincide with syntactic phrase boundaries, and the relationship between prosody and syntax is still not well understood, yet.

Methods	average accuracy
(1) syntactic grouping: 26 groups	80.9%
(2) text-corpus grouping: 18 groups	78.3%
(3) speech-database grouping: 18 groups	83.1%

Table 1: The average prediction accuracy with different POS grouping methods.

Next, the confusion table for the predicted break indices was established (see Table 2). We are unable to compare the performance with other reported results due to the difference in the language and database used. Another important consideration in evaluating the model is that prosodic phrase structure is not deterministic, either; speakers can produce a sentence in several ways without altering naturalness or meaning. Therefore, the quality of the speech output using the predicted prosodic structure would be the best available measurement of this statistical model itself.

Predicted / actual labeled	no break (0)	minor break (1)	major break (2)
no break (0)	85.7%	10.5%	3.8%
minor break (1)	10.8%	81.2%	8.0%
major break (2)	4.6%	16.7%	78.7%

Table 2: The confusion table for the predicted break indices using the POS's grouping by speech database.

4. DETERMINATION OF PROSODIC PARAMETERS

The main purpose of the prosodic structure described above is to determine the prosodic parameters for Mandarin speech synthesis. The basic prosodic parameters involved in our system are fundamental frequency(F_0), energy, duration, and pause. Acoustic analysis of speech production shows that many prosodic features have to be considered in the implementation of a text-to-speech system. Phenomena such as the insertion of pause at grammatical junctures, the declination tendency of F0 contour, the diminution of F0 range, and the lengthening at the end of a word or sentence can all be found in our speech database. The prosodic structure is then used to realize these effects in the output of synthetic speech.

The prosodic parameters are decided according to the prosodic structure on different levels. Global intonation and energy contours are applied to the major prosodic phrases depending on the type of punctuation mark. Only four types of contours are used in the present initial study:

- Type 0: sentence middle (without punctuation mark)
- Type 1: comma (，)
- Type 2: period mark (。)
- Type 3: question mark (？)

These contour patterns were extracted from the speech database according to punctuation marks or some special sentence-final particles such as 嗎,呢… in Chinese. A dynamic range contour was also extracted to realize the diminution effect of F_0 range.

An important characteristic of Mandarin Chinese is that it is a monosyllabic based tonal language. Each syllable is associated with a specified lexical tone. These tones are categorized as:

Tone 1	—	high and level	55
Tone 2	ˊ	rising	35
Tone 3	ˇ	fall-rise	214
Tone 4	ˋ	falling	51
Tone 0	•	variable	

Although the interaction among sequences of tones is phonologically predictable at both the word and phrase levels, we in fact extracted the F_0 patterns of all possible tonal combinations from the designed word speech database in order to study possible non-phonological factors that may very much affect the realization of tones in actual speech. These tone patterns are used to form the basic fundamental frequency of the words located in a minor prosodic phrase. The usage of larger units for the tone patterns will increase the fluency of the speech output.

The duration of the syllable is determined by the following equation:

$$d = d_i \bullet r_t \bullet r_p \bullet r_b$$

where d_i is the average duration of syllable i and the other factors are the ratios according to the tone, position, and break index. The length of the pause between two words is determined by the break index. Furthermore, a small random number is used to adjust the pause length to avoid the effect of mechanical rhythm. The whole mechanism to determine the prosodic parameters is illustrated in Figure 3.

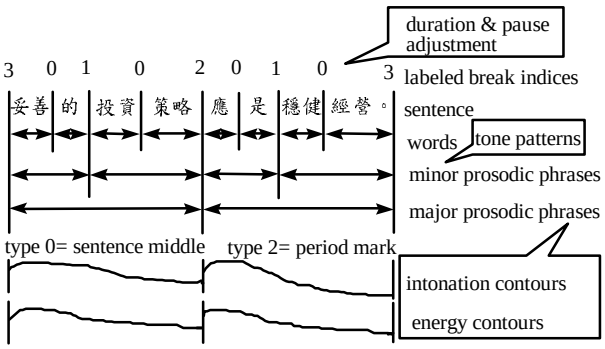

Figure 3: The mechanism for the determination of the prosodic parameters according to the prosodic structure.

After determining the prosodic parameters, the Time-Domain Pitch Synchronous OverLap-Add (TD-PSOLA) algorithm was used to synthesize the output speech. The system is currently under development on a personal computer running Windows 95. A set of function library was developed for Mandarin speech synthesis. These functions can be integrated in many different applications using voice as the output media. An example of the output speech signal for the input string "妥善的投資策略應是穩健經營。" is shown in Figure 4.

5. CONCLUSIONS

In this paper, we have focused on the synthesis application of the prediction of prosodic structure. The model proposed here is to generate a hierarchical structure that represents a sentence in different forms of prosodic phrases. The model is stochastic, which accounts for the natural variability in human speech, and can be automatically trained to reflect the speaking style of a speaker. The grouping methods used to design the best predictor of prosodic structure can also provide new insight into the relationship between prosody and syntax. We have found that the best prediction result of phrase breaks was achieved with the POS's grouping by the speech database. Therefore, a low complexity prediction algorithm was developed without precise syntactic analysis or detailed Parts-of-Speech labeling. Natural synthetic speech was generated in the initial simulations with the prosodic parameters derived from the proposed prosodic structure.

6. REFERENCES

1. Shigery Fujio, Yoshinori Sagisaka and Norio Higuchi, "Prediction of Prosodic Phrase Boundaries Using Stochastic Context-free Grammar", ICSLP, pp. 839-842, 1994.
2. Ostendorf and N. Veilleux, "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location", Computational Linguistics 20(1), pp. 27-54, 1994.
3. Merle Horne and Marcus Filipsson, "Generating Prosodic Structure for Swedish Text-to-Speech", ICSLP, pp. 711-714, 1994.
4. Marcel Tatham and Eric Lewis, "Prosodics in a Syllable-Based Text-to-Speech Synthesis System", ICSLP, pp. 1179-1182, 1992.
5. Shaw-Hwa Hwang and Sin-Horng Chen, "A Prosodic Model of Mandarin Speech and Its Application to Pitch Level Generation for Text-to-speech", ICASSP, pp. 616-619, 1995.
6. Chiu-yu Tseng, "A Phonetically Oriented Speech Database for Mandarin Chinese", ICPHS, pp. 326-329, 1995.
7. "The Analysis of Chinese Part-of-speech", Technical Report no. 93-06, Chinese Knowledge Information Processing Group, Institute of Information Science, Academia Sinica, ROC.

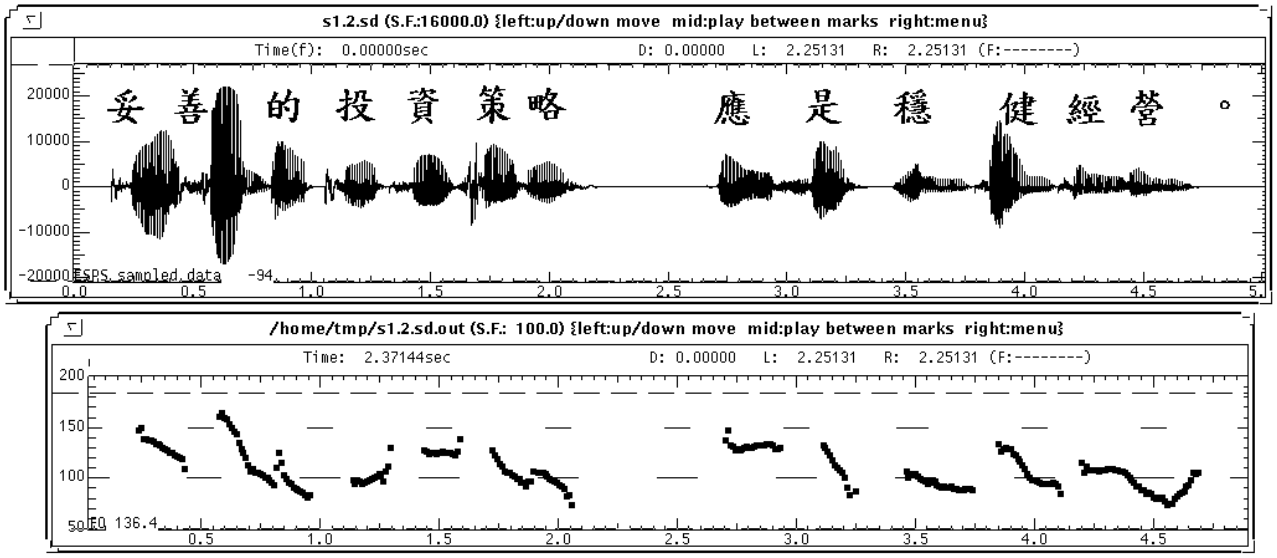

Figure 4: The waveform and fundamental frequency of the synthetic speech "妥善的投資策略應是穩健經營。".