USING PROSODY DATABASE IN CHINESE SPEECH SYNTHESIS

Dong Minghui, Lua Kim Teng
Department of Computer Science, School of Computing
National University of Singapore
S15, 1 Science Drive 2, Singapore 117544
{dongming, luakt}@comp.nus.edu.sg

ABSTRACT
As the difficulty of revealing the relationship between text and prosody, a corpus based approach is used in prosody generation in the research. A prosody database is built as templates for prosody generation. The general idea is that we are trying to get the prosodic information from real speech examples. We first analyze given Chinese text, and form a linguistic feature vector, which describes the phonetic and lexicon characteristics of the text. Then we search the database to find the best match of the vector, which is a similar occurrence of the text. The prosody parameters of the retrieved example will be the guideline of the prosody we are going to generate.

1. INTRODUCTION
Though many efforts have been put on the Chinese text to speech synthesis, and many text-to-speech systems have been built, the naturalness of synthetic speech is still not satisfactory. The main problem is that it is difficult to generate proper prosody that resembles that of human speech.

To generate good prosody for Chinese, many researches have been done. Rule-based algorithms were used[1],[3]. But deriving the rules needs a lot of work and the derived rules may not be complete. To get higher quality prosody, more complicated rules may be needed. This makes it difficult to build a system to generate prosody of very good quality. Many systems apply word-level prosody[2],[5],[6]. But the supra-word influence is considered by using rules in the approaches. A neural network method has also been investigated to generate prosody parameters, and the reported result is good[4]. However, using a neural network method is not much flexible. It is hard to investigate the influence of a particular factor.

To make better prosody, supra-word information should be taken into consideration and rules should be avoided. In this paper, we describe an effective example-based approach in Chinese sentence prosody generation. It is a corpus-based approach with phrase level pitch pattern considered. We call it an example-based prosody generation because we have a speech database as examples for prosody generation. It is a hierarchical model, which consists of three levels, i.e. syllable level, phrase level and sentence level.

The basic principle of the approach is that we try to find similar occurrence of the speech to be generated in a real speech database, and apply its prosody to the speech we are going to generate. All the prosody information of our speech data in the corpus has been stored in the database. In the synthesis phase, we first analyze the text, and decompose it into short phrases. A phrase is then described as a feature vector using linguistic and phonetic information. Next, we search the database for the linguistic feature vector. When an exact match or the most similar match is found, its corresponding prosody will finally be used as prosody template for synthesis.

In the following sections, we will first describe the procedures of speech data and text data processing in the research. Then we describe the prosody generation model.

2. SPEECH DATA PROCESSING
To help build the prosody database, we have designed programs to do automatic speech segmentation and automatic break type labeling.

2.1 Speech data segmentation
We have a corpus of around 4000 sentences and words. The speech data are continuous speech with text transcripts. To analyze prosody parameter and generate prosody database, we need to handle the speech in a syllable level. So segmentation of the continuous speech into syllable pieces is required. To reduce the workload of segmentation work, we use HMM based recognition algorithm to automatically perform segmentation.

First we translate all the text transcripts into phonetic transcripts. For the speech recognition, we use 14th order CEP and 14th order delta-CEP as our speech recognition parameters. A frame length of 10ms is used. An HMM model is defined for each Chinese syllable. We train the HMM models using all the speech data we are going to segment.

Segmentation is done after the models built. The segmentation is actually a constrained recognition process. The phonetic transcripts are the expected recognition results. Guided by the transcripts, recognition can be more accurate. We modify the Viterbi beam search algorithm in recognition process to narrow the searching path. When the best path is found, the start and end state is obtained. Hence we get the start time and end time of each syllable.
2.2 Break type labeling

To investigate the relationship between words, we defined break types between words. In our research, we defined 3 types of break between words, which are word break, minor break and major break, similar to those defined in [2]. Word break means there is no distinct break between words. Minor break means a small break within an intonation phrase. Major break means an intonation phrase boundary.

To predict break types from text, a prediction model is required to build. The construction of the model needs pre-labeled text according to the actual speech data. To minimize the work of labeling the break types, we use program to do the work. The length of pause between two syllables is a hint for break types. However, the difficulty is that the duration is very hard to be identified. To solve the problem, we use energy to normalize the duration of syllable. The method of normalization is as in figure 3. The figure shows energy of a syllable. Sl and El are labeled start and end. Sn and En are normalized start and end. The Sn and En meet the following criteria.

\[
F(s, e) = \int E(t) dt
\]

\[
F(sl, sn) = F(en, el) = \alpha \cdot F(sl, el)
\]

where E(t) is the energy as illustrated in the figure. \(\alpha\) is a small value, eg 0.05.

3. PROSODY GENERATION

There are many factors in prosody, among which pitch contour and duration parameters are the two most important ones. Therefore, our research mainly concentrates on the two parameters.

3.1 Syllable level prosody

Chinese is a monosyllabic language, in which each Chinese character is a syllable. It is quite natural to use syllable as the basic unit in our research. As Chinese is a tonal language, tone information plays an important rule in the investigation of syllable prosody. In our research, syllable level prosody includes the duration of a syllable and the pitch contour of syllable. Statistical methods were used to calculate the default value of the parameters.

3.1.1 Duration of syllable

Each Chinese syllable is composed of an initial and a final. But to accurately separate initial and final in a syllable is not easy. Instead of trying to consider the initial and final directly, we divide the duration of a syllable into two parts, voiced and unvoiced parts. The unvoiced part is usually contributed by initial, and the voiced by final. The influence of tone on duration will be considered in the final part, which consists of vowel.

We group all the unvoiced part by the initials of syllables. Then take average in each group to get the default duration of initial. To get default duration of voiced part, we group all the voiced part with same final and same tone. By taking the average duration in each group, we obtained the value.

In synthesis, the default duration of syllable can be calculated this way.
\[ D_{\text{def}} = DU(\text{Initial}) + DV(\text{Final, Tone}) \]

where \( D_{\text{def}} \) means default duration of a syllable, \( DU \) \( DV \) are functions that determine the default duration of initial and final parts.

The default duration is used as a guideline in synthesis process. Modification can be made to the default duration by multiplying a control factor, which will be described later.

### 3.1.2 Pitch contour of syllable

The pitch contour of a syllable is usually influenced by the tones of its surrounding syllables. Therefore, we take the surrounding tones of a syllable into consideration in the research. We group the syllables according to the combination of their tone, previous tone and next tone. When the previous one or the next one does not exist, we define it a null tone. The three tones form a vector. We define it a tone vector.

For the pitch contour of each syllable, we evenly sample the pitch contour. The sampled values form a vector, which is defined as tone pitch contour vector in the research. We put the syllables with same tone vector into one group, then take the average of the vectors in each group. A tone pitch contour vector is produced for the tone vector.

A tone pitch contour table is built, in which each record includes a tone vector and a pitch contour vector. This way, we get the pitch contour for a tone, taking its surrounding tones into consideration. The smoothness of the pitch transition between the syllables is guaranteed. And complicated tone sandhi rules are not needed. All rules have been implied in the table. The two vectors can be represented using following way. \( P_i \) means sample values of pitch.

\[
TV = (\text{PrevTone, Tone, NextTone}) \\
TPCV = (P_1, P_2, …, P_n)
\]

In synthesis, when coming up with a syllable, we first form a tone vector for it. Then we search the tone vector in the tone pitch contour table to find the corresponding pitch contour vector. Based on the duration and the pitch contour vector, the pitch contour will be generated by using interpolation algorithms.

### 3.2 Phrase level prosody pattern

Phrase level prosody pattern is the general pattern for a phrase. It includes the pitch level (average pitch value), duration of each syllable, etc.

To determine the prosodic pattern of a phrase, a phrase prosodic pattern database is established. In the prediction phase, first phrase boundaries are determined using part of speech (POS) information. Then the prosodic pattern is retrieved according to the linguistic feature vectors that describe phonetic and lexicon information. Pattern match technique is used to determine the best match of the linguistic feature and hence to determine the prosodic pattern for a phrase. The prosodic parameters will be calculated according the prosodic pattern. Figure 3 shows how the prosody of a phrase is generated.

![Figure 3. Process of phrase level prosody generation.](image)

#### 3.2.1 Break type detection

Phrase boundary detection is made based on the part of speech information. The text has been labeled with POS types and different prosodic break marks according speech data as described in previous part. For each pair of POS combination, we calculate the probabilities of different break types. In detection, the break between two words will be assigned a type that holds the highest probability for the POS pair.

#### 3.2.2 Setting up prosody pattern database

Our corpus consists of around 4000 sentences and words, which are uttered by a male speaker. The speech and text of each sentence are both analyzed. Text is translated into phonetic symbols to form a linguistic feature vector. Each Chinese character is turned into a pinyin (Chinese pronunciation syllable) marked with break types before and after it. Speech is analyzed to form a prosodic feature vector, which includes the pitch level and duration of syllable. The linguistic feature vector (LFV) and speech feature vector (SFV) can be represented in the following way.

\[
\text{LFV} = (LF_1, LF_2, \ldots, LF_n) \\
\text{SFV} = (SF_1, SF_2, \ldots, SF_n) \\
\text{LF}_i = (\text{Pinyin, Tone, BreakBefore, BreakAfter}) \\
\text{SF}_i = (\text{PitchLevel, Duration})
\]

#### 3.2.3 Prosody pattern match

For a given Chinese text, word segmentation is first done. Part of speech tagging is performed next. Then break types between words are determined. Text to phonetic translation is made. After all of these, a linguistic feature vector is formed for each phrase. Then the database is searched against the linguistic feature vector. A best match is found to represent the prosodic pattern of the phrase.

To find the best match in the database, we use pattern match approach. We try to match the linguist vector of the input
text with all the linguistic vectors in the database. The similarity between two linguistic vectors \( S \) and \( T \) can be defined using the following equations.

\[
Sim(S, T) = \sum_{i=1}^{n} \left( Sim(S_i, T_i) \right)
\]

\[
Sim(S_i, T_i) = a \cdot Sim(IT_i, IS_i) + b \cdot Sim(FT_i, FS_i) + c \cdot Sim(TT_i, TS_i) + d \cdot Sim(BT_i, BS_i)
\]

The notations in the above are described in the following.

- \( S \) is a phrase that consists of syllables \( S_1, S_2, \ldots, S_n \).
- \( T \) is a phrase that consists of syllables \( T_1, T_2, \ldots, T_n \).
- \( IT_i, FT_i, TT_i \) are the initial, final and tone of the \( T_i \) and \( BT_i \) is break type after \( T_i \).
- \( IS_i, FS_i, TS_i \) are the initial, final and tone of the syllable.
- \( BS_1 \) and \( BS_2 \) are break types before and after the syllable.
- \( Sim, Sim1, Sim2, Sim3, Sim4 \) are the similarity functions.
- \( a, b, c \) and \( d \) are weights for different features.

When we find the best match of the linguistic feature vector, the corresponding speech prosody feature vector will be used as the prosody template for the speech we are going to generate.

If a phrase is too long, the phrase can be cut into small parts, we can take each part as a phrase and follow the same process as to a phrase. This increases the probability to find the most similar match. If phrases are too short, they can be combined as a long phrase. This makes fewer joints in a sentence. Hence prosody of sentence will be better.

As we have described in the previous section, there is a default pitch contour and default duration for each syllable. After getting the prosody parameter, we can make modification to the pitch contour and duration. The modification is made this way.

\[
D_{tg} = D_{def} \ast R(PY_{sc}, PY_{tg}, D_{sc})
\]

\[
PL_{tg} = PL_{def} + Diff(T_{sc}, T_{tg}, PL_{sc})
\]

Where \( T_{tg} \) means the target syllable that we are going to generate. \( D_{def} \) means default value for the syllable. \( D_{sc} \) means the source syllable in our example. \( D_{tg} \) means duration. \( PL_{def} \) means pitch level. \( PY \) means phonetic information. \( T \) means tone of syllable. \( R \) and \( Diff \) are functions to adjust the duration and pitch level parameters respectively.

### 3.3 Sentence level prosody

Sentence level prosody is actually the combination of syllable level prosody and the phrase level. The phrases are concatenated together and pauses are inserted between phrases. Pitch level should be adjusted at the joint of two phrases. Changes also should be made to make the whole sentence keep a declination trend.

![Figure 4. Comparison of generated prosody (upper) and synthetic prosody(below).](image)

### 4. RESULT AND DISCUSSION

We applied our generated prosody to PSOLA bases synthesis component. Informal test showed the generated prosody is quite good. Figure 4 shows a comparison of a natural prosody and a generated prosody as an example. The pitch contour is for the text “ma1 ma5 you3 yi1 ben3 xiang4 ce4 (Mother has an album)”. The upper one is for natural speech. The lower is for synthetic speech. From the figure we can see that the pitch contour and durations are quite similar.

However, there are also shortcomings for the approach. Some synthetic speech may have improper prosody due to lack of similar pattern in the database. Therefore the database needs to be carefully designed to have a better cover of phrase patterns. Another problem is that the pattern match process is rather slow. So some measures should be taken to speed the match process.

### 5. CONCLUSION

We briefly described our method to use prosody database in the generation of prosody for mandarin Chinese. We use automatic speech segmentation program and automatic speech break type labeling program to help speech processing. Word segmentation and POS tagging are performed for text processing.

The prosodic model is a hierarchical model. The prosody of a sentence is considered in 3 levels. In syllable level default duration and pitch contour for syllable are generated. In phrase level, phrase prosody pattern database is built as phrase pitch pattern examples. Phrase pitch level is generated according the best match in the database. Sentence level prosody is the combination of the whole phrase level prosody and syllable level prosody.
6. REFERENCES


