AUTOMATIC PROSODIC BREAK LABELING FOR MANDARIN CHINESE SPEECH DATA

Minghui Dong, Kim-Teng Lua
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
For corpus-based speech synthesis, large quantities of labeled speech are required. Manually labeling speech data is quite labor-intensive. Therefore, automatic speech labeling is highly desired. Prosodic break detection is one of the tasks for automatic speech labeling. In the paper, we propose an automatic break detection algorithm for mandarin Chinese speech. In this approach, we use energy contour to normalize duration of syllables and use the concept of normalized transition time to represent the time interval between two syllables. A recursive algorithm is then used to select locally longer intervals as pauses. Language specific constraint rules are also used to produce a better judgment. The automatic break labeling results have been proved to be good.

1. INTRODUCTION
Corpus-based approaches have been widely used in speech related researches, especially in researches where large quantities of carefully labeled data are required to perform prosodic analysis and conduct other experiments. However, acquiring large quantities of speech data is a labor-intensive task if it is done manually. Much research has been conducted on the automatic labeling of speech data with higher accuracy. These works include automatic segmentation of speech data, automatically labeling of speech events, and so on. Of these, one of the more important speech events is the prosodic break between syllables.

The primary goal of our research is to build a labeled corpus for mandarin speech synthesis with the help of some programs. One of the tasks is to automatically detect the break between Chinese syllables. There are some related research on this [2][3][4], but Chinese-specific phenomenons have not been considered in them. As we know, Chinese is a monosyllable-based language. Each Chinese character is pronounced as a syllable and a word usually consists of one, two or more characters. A sentence is actually a series of syllables. There is usually no perceptual break between syllables in a word. Words can be grouped into prosodic groups. Breaks can appear between prosodic groups. Our main concern in this research is to determine the breaks from real speech.

When there is a break between two syllables in speech, there is usually an instinctive pause although sometimes it is not significant enough to be audible. Thus, the intuitive approach for break detection is to find the pause time between two syllables. However, to determine the pause time between syllables is not easy. There are some reasons for the difficulty. Firstly, there might be no distinct pause in then speech signal even when there is an actual break as perceived by human judgement. On the other hand, sometimes when there seems to be a break in the speech signal, it is not necessarily a break if judged by human perception. To overcome the difficulties, we adopt two quantitative measures - duration normalization by energy contour and calculation of normalized transition time. To obtain better results, we have also integrated some rules for Mandarin Chinese speech. These rules mainly include restrictions on the length of prosodic words and prosodic phrases.

2. METHODOLOGY
Our labeling work includes the following steps. First we carry out automatic speech segmentation using speech recognition techniques. Then we apply automatic break labeling techniques on the speech segments. In this section, we will describe the details of our methodology.

2.1 Speech data segmentation
We have a corpus of around 1000 sentences and phrases consisting of around 10200 syllables. The speech data are continuous speech with text transcripts. All the speech data are read by a male speaker. To analyze prosody parameters, we need to handle the speech at the 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 perform automatic segmentation.

First we translate all the text transcripts into phonetic transcripts. For the continuous speech recognition, we use 14th order CEP and 14th order delta-CEP as our speech recognition parameters. A frame length of 10ms is used. A HMM model is defined for each Chinese syllable. There are 408 models built for 408 toneless syllables in Chinese in our research [3]. We train the HMM models using all the speech data we are going to segment.

Segmentation is done after the models have been built. The segmentation is actually a constrained recognition process. The phonetic transcripts are the expected recognition results. Guided by the transcripts, recognition is expected to be more accurate. We modify the Viterbi beam search algorithm during the recognition process to narrow the search path. When the best path is found, the start and end states are obtained. Therefore, we are able to obtain the start time and end time of each syllable.

2.2 Duration normalization
Once the onset and offset of each syllable is obtained, we can deal with each syllable. In principle, the break time between two syllables can be determined by the time difference between the start time of current syllable and the end time of previous one. However, the onset and offset of syllable are usually not stable enough. There are two kinds of problems affecting the calculation of pause time. One is that the breaks are filled by lengthening of previous or succeeding syllables. There will be no distinct pause or silence part in the speech waveform. Figure 1 shows an example of this case. In the figure, there is no silence part between dao4 and wo3 although we can perceive a break by hearing. The other is that for a gradual increase of energy at the
onset and gradual decrease of energy at the offset, it is not easy to accurately determine the start and end of a syllable. So it is not easy to make all the syllables comparable.

Figure 1. Waveform of “kan4 dao4 wo3 men0” (see us). There is no distinct pause in dao4 and wo3.

One way to solve the problem is to use normalized duration by using the mean and standard deviation for each syllable. However, the resulting syllable may experience the problems of strength and tone. Same pronunciation with different tones may have different durations in Chinese. An accented syllable typically holds a longer duration, whereas a weakened one does the opposite.

Figure 2. Syllable duration normalization

To remove the possible lengthening effect and the ambiguity of starting and ending point, we use a duration normalization method to recalculate the onset and offset of syllables. Energy contour is used as a guide for the normalization process. The method of normalization is shown in figure 2. 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. \( F(s,e) \) means the area between s and e.

In this way, we removed the low energy part at two ends of a syllable and only cover the main part of the waveform of it. The normalized onset and offset time is quite stable. In the later calculation, we use the normalized onset and offset time.

2.3 Transition time

Another phenomenon in speech is that when there is no perceptual break between syllables, there is a break in speech signal or energy contour. This is mainly caused by some plosive initials of Chinese syllable because humans need time to adjust speech organs to pronounce some consonants. Figure 3 shows an example of this case. There seems a break between bu4 and tong2, but we do not perceive it as a break.

Figure 3. Waveform of “bu4 tong2 de0” (different). There seems a break between bu4 and tong2.

Each Chinese syllable is composed of an initial and a final (sometimes the initial syllable is optional). Finals are vowels or vowels with nasal consonants. They usually possess high energy. The normalized duration mainly covers the final part of a syllable. Different kinds of initials may have different intrinsic length in time because in articulating speech, human needs different amount of time to pronounce different sounds.

To normalize the pause time between syllables, we use the concept of transition time, which means the time for human to change the shape of speech organ from one syllable to another. This is the pause time between the normalized end of one syllable and the normalized start of succeeding syllable. We assume the time is mainly determined by the initials of syllables.

In the research, we first calculated the mean and standard deviation of transition time for different initials. Then we normalize the transition time in the following way.

\[ Tn(i) = \frac{T(i) - \mu_j}{\sigma_j} \]

where \( Tn(i) \) is the normalized transition time for syllable i, \( T(i) \) is the original transition time of syllable i, which is the start time of syllable i minus the end time of syllable i-1. \( \mu_j \) and \( \sigma_j \) are the mean and standard deviation of the transition time for initial j (The initial of syllable i is j).

2.4 Recursive detection algorithm

One of the problems we need to highlight is that the speaking rate may vary in different sentences. This means the average transition time is longer and shorter in some utterance and the absolute break time may be different on different occasions. Using absolute values for break detection is not appropriate because of the variation of the pause time in different speaking rates. In the research, we used a recursive break detection algorithm to find the prosodic breaks.

The main idea of recursive break detection is to find the longest pause in the whole sentence. We set it as a break. Thus the sentence is divided into two parts. For each part, we find the longest pause as break. This is a recursive process. The process is terminated when some conditions are met (see 2.6). The conditions mainly include the number of breaks and the length of prosodic group. The algorithm works as in algorithm 1. The meaning of the functions and variables are self-explainable.
2.5 Break types
For the purpose of analysis of prosody, we have defined break types for the speech data. In our research, we defined 3 types of break between syllables, which are no break, minor break and major break (similar to defined in [6]). No break means there is no perceptually distinct break between two syllables. Minor break means a small break within an intonation phrase, in which a syllable is affected by its surrounding ones. Major break means a longer break, which usually indicates an intonation phrase boundary. The actual time lengths of the two types of break are dependent on the speaking rate.

Although pitch value is a cue for identification of prosodic phrase boundary, currently we classify the break types mainly based on the length of the transition time because pitch value change is usually accompanied by a longer break time.

2.6 Constraint rules
We have used some language specific rules to restrict the possible mistakes in our break detection process. These rules include:

Length of prosodic word: We assume the minimal length of a word is two. So there are few cases where two breaks are adjacent unless there is a significant pause time. This assumption is based on two observations. One is that most Chinese words are bi-syllable words. The other is that monosyllable words are often combined with their adjacent words to form larger prosodic groups.

Length of prosodic phrase: We assume that the major break only appears in the utterances whose length is longer than seven syllables. This is based on the observation that when a sentence has more than seven syllables, people are more likely to break it into shorter prosodic groups.

Word boundary from text: We use the result of word segmentation for Chinese text. Breaks can only appear in word boundary. Therefore there should be no break between two syllables within a word.

3. EXPERIMENTS
Based on the abovementioned methodology, some experiments have been conducted.

3.1 Experiment 1: Transition time calculation
We have a corpus of around 1000 sentences and phrases, consisting of around 10200 syllables. First all the sentences are segmented into syllables, and we also corrected the mistakes in the automatic segmentation. To calculate the transition time, we remove the first syllable in each sentence because we did not define a transition time for it.

3.2 Experiment 2: Break detection with absolute thresholds and pause rank.
Among all the speech data, we manually labeled around 400 sentences and phrases as our test set. The experiment is performed using two absolute thresholds to classify three types of breaks. We used one restriction, which is that we only consider the syllable intervals with large transition time as candidate breaks.

We sort transition time in descending order. Only first half of the syllable intervals are considered as candidates for breaks. In this set, we use the absolute thresholds to classify the breaks into major, minor breaks or no break. The algorithm is as in the algorithm 2. Here the values of threshold1 and threshold2 are 0.2 and 0.7 respectively.

The experimental result is shown in table 2. The table compares the automatic labels with the manual labels. The accuracies of minor break and major break are 63.0% and 75.2% respectively.

```
Algorithm 1. Recursive break detection

BreakDetection (Start, End
/
if (ShouldTerminate()) return;
BreakPos=FindTheLongestPause(Start, End);
KeepTheBreak(BreakPos);
BreakDetection(Start, BreakPos-1);
BreakDetection(BreakPos+1, End);
}

Algorithm 2. Thresholds with pause rank

If (Rank(TransitionTime) > LengthOfSentence/2)
BreakType = NoBreak;
Else if (TransitionTime > Threshold2)
BreakType = MajorBreak;
Else if (TransitionTime > Threshold1)
BreakType = MinorBreak;
Else BreakType = NoBreak;
```
Most of the errors are misclassification of no-break into minor break and minor break into major break. Some possible reasons may explain the errors. First, the normalization using energy contour cannot ensure that the transition time between two syllables fully reflects the intended break of speaker. Second, we assume that transition time is only dependent on the initials of a syllable. This is only an approximate solution. It might also be dependent on the finals of previous syllable and the speaking rate. The absolute thresholds may not be proper for all breaks in a sentence. We try to reduce the number of mistakes in experiment 3.

```c
BreakDetection (Start, End)
{
    if (End-Start<3) return;
    CalculateAllPossibleBreaks(Start, End);
    RemoveBreaksWithinWords(Start, End);
    BreakPos=FindLongestBreak (Start, End);
    BreakType= NoBreak;
    If (TransitionTime > Threshold2)
        BreakType = MajorBreak;
    Else If (TransitionTime > Threshold1)
        BreakType = MinorBreak;
    If (End-Start < 7)
        BreakType=MinorBreak;
    KeepTheBreakType(BreakPos, BreakType);
    BreakDetection(Start, BreakPos);
    BreakDetection(BreakPos+1, End);
}
```

Algorithm 3. Recursive detection with rules.

<table>
<thead>
<tr>
<th>Manual labels</th>
<th>Automatic labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No break</td>
</tr>
<tr>
<td>No break</td>
<td>1848</td>
</tr>
<tr>
<td>Minor break</td>
<td>73</td>
</tr>
<tr>
<td>Major break</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2. Break detection using absolute thresholds.

### 3.3 Experiment 3: Recursive detection with rules.

In this experiment we use recursive break detection algorithm and the constraint rules. The algorithm is as in algorithm 3. The meaning of the function and variables can be explained by its name. We use the same testing set as in experiment 2. The result is as in table 3. The misclassification errors have been reduced significantly. There is no confusion between no-break and major break. The accuracies of minor break and major break are 77.4% and 82.7% respectively. From the improved result, we can see that recursive algorithm and the rules play an important role in the accuracy of break type detection.

For the mistakes, we think they arise from following reasons. First, The rules cannot cover all phenomena in the Mandarin Chinese. Second, the perceptual breaks cannot be correctly judged only by energy. Pitch information should also be taken into consideration.

<table>
<thead>
<tr>
<th>Manual labels</th>
<th>Automatic labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No break</td>
</tr>
<tr>
<td>No break</td>
<td>2072</td>
</tr>
<tr>
<td>Minor break</td>
<td>57</td>
</tr>
<tr>
<td>Major break</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3. Results using recursive detection with rules.

### 4. CONCLUSION

We studied the problems of automatic prosodic break detection for Chinese speech. Normalization of duration of syllables is used to obtain a stable starting and ending time for a syllable. The transition time is used to describe the effect of initials. Normalization of the transition time is to solve the problem of significant differences of transition time among different syllables. Recursive detection method is to give a relative comparison of break time. And the language specific rules are used to further reduce the number of possible mistakes. The experiments show that the performance of these measures is good. The automatic program can help us to label large speech corpus for our further research of speech synthesis.

### 5. REFERENCES