Identifying Linguistic Segmentations in Chinese Spoken Dialogue

Yue-Shi Lee and Hsin-Hsi Chen
Dept. of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
E-mail: {leeys, hh_chen}@csie.ntu.edu.tw

ABSTRACT

In a continuous speech recognition system, a longer waveform is usually segmented into some shorter pieces based on simple acoustic criteria, such as unfilled pauses (i.e., silences). We call such a kind of segmentation as an acoustic segmentation. In general, the acoustic segmentations do not reflect the linguistic structure. They may fragment sentences or semantic units. Besides, they may also group together some unrelated units. Therefore, we need to resegment acoustic segmentations in order to output linguistically meaningful units such as clauses. We call such a kind of segmentation as a linguistic segmentation. This paper employs several acoustic and prosodic clues to resegment acoustic segmentations for identifying linguistic segmentations. Based on these clues, the experimental results show that a precision rate of 94.46% and a recall rate of 87.38% can be achieved.

1. INTRODUCTION

In a continuous speech recognition system, a longer waveform is usually segmented into some shorter pieces based on simple acoustic criteria, such as unfilled pauses (i.e., silences). We call such a kind of segmentation as an acoustic segmentation. In general, the acoustic segmentation does not reflect the linguistic structure. They may fragment sentences or semantic units. Besides, they may also group together some unrelated units. Therefore, we need to resegment acoustic segmentations in order to output linguistically meaningful units such as clauses. We call such a kind of segmentation as a linguistic segmentation.

Automatic detection of linguistic segmentation could improve the performances of many spoken language systems. For a speech retrieval system [2-3], it could process its data in more meaningful units, if the linguistic boundaries were known. Meteer and Iyer [4] also showed that the performance of a language model could be improved if both training and testing data are segmented linguistically, rather than acoustically. Besides, Bakenecker, et al. [5] also claimed that the parsing speed could also be improved when the clause boundaries of utterances were known. Moreover, proper modeling of filled pauses also requires knowledge of linguistic segment boundaries as described by Stolcke and Shriberg [6-7].

The following hand-annotated transcription extracted from our spoken corpus described in next section illustrates the characteristics of these two segmentations. It is based on the acoustic segmentation. That is, each line in the following transcription stands for an acoustic segmentation. The boundary of linguistic segmentation is marked by '<s>'

1 L: 只喝啤酒=-
2 ...就[完<@聲^啪@>]\ <s>
3 H: [對啊]\ <s>
4 都--{R1,1-1,1}
5 ..都沒錄到這樣子啊\ <s>
6 ...{(8)[那你%,-
7 ...給我們一人],-
8 L: [咦\ <s> 可是]-- <s>
9 H: ...一百塊好啦\ <s>
10 ...(1.6)談話費\ <s>

The snapshot shows each line includes three fields: (1) Reference, (2) Speaker, and (3) Utterance. Reference field shows the utterance number in the spoken corpus. Speaker field shows the identity of the speakers at a turn. The content of the utterance is shown in the third field. In this field, the three symbols, i.e., ...\(N\), ... and ..., denote an unfilled pause (silence) is long, medium and short, respectively. The symbol % denotes the glottal stop. The terminal pitch directions are denoted by three symbols, i.e., \(\downarrow\) (fall), - (level), and \(\uparrow\) (rise). The accent and lengthening are denoted by ^ and =, respectively. The quality information and vocal noises are also transcribed in this spoken corpus.
This short transcription also shows (1) some mismatches between acoustic segmentations and linguistic segmentations (e.g., utterance 1), (2) the linguistic segmentation spans several turns (e.g., utterances 6, 7 and 9), and (3) linguistic segmentation may occur within an utterance (e.g., utterance 8).

In general, the linguistic segmentations encompass different kinds of linguistic structures [1]. They include (1) complete sentences, (2) stand-alone phrases (e.g., noun phrases), (3) interjections, and (4) back-channel responses. Besides, no linguistic segmentations (i.e., <s>) will be inserted within the speech repairs (e.g., utterance 4). Thus, the repair processing will affect the segmentation performance.

This paper will employ several acoustic and prosodic clues to resegment acoustic segmentations for identifying linguistic segmentations. In the following section, section 2 introduces the spoken corpus used in this paper. Section 3 describes the acoustic and prosodic information to help the detection of linguistic segmentations.

Before concluding, we demonstrate the experimental results of our analysis methods.

2. SPOKEN CORPUS

The spoken corpus analyzed in this paper consists of two commonplace, everyday conversations among friends. Each is about forty-minute long. There are four and five speakers in these two conversations, respectively. It is originated from Professor Kaiwai Chui at National Chengchi University [8].

In total, this corpus contains 5,395 utterances (acoustic segmentations), 22,409 words and 2,602 turns. There are totally 448 speech repairs. On the average, 17% of turns contain at least one speech repair. Tables 1 and 2 list the frequency distribution of linguistic segmentations in these two conversations.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>L</th>
<th>H</th>
<th>Z</th>
<th>O</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>418</td>
<td>580</td>
<td>393</td>
<td>177</td>
<td>1568</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speaker</th>
<th>L</th>
<th>W</th>
<th>Y</th>
<th>Z</th>
<th>J</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>422</td>
<td>608</td>
<td>792</td>
<td>680</td>
<td>200</td>
<td>2702</td>
</tr>
</tbody>
</table>

From Tables 1 and 2, there are 1568 and 2702 clause boundaries in conversations 1 and 2, respectively. As mentioned in previous section, the linguistic segmentation may occur within an utterance. However, only 1.1% segments belong to this type in these two conversations. Because this paper identify linguistic segmentations based on acoustic and prosodic clues, the Chinese characters in the spoken corpus are converted into the corresponding syllables manually.

3. ANALYSIS METHODS

This section will describe several clues based on acoustic and prosodic information to help the detection of linguistic segmentations.

Clue1: Terminal Pitch Direction

For linguistic segmentation, the terminal pitch direction in the repaired segment is important. When the terminal pitch direction of an utterance is level, the utterance is often unfinished by the speaker. This also means that the linguistic structure is often not formed. Thus, no <s> will be inserted at the end of this utterance. The following shows an example.

1. L: 只喝五兩-= -
2. ...就 [介<诸@蛋^咖@>]\ <s>
3. H: [介<何]\ <s>
4. 都-- [R1.1-1.1]  
5. ...都 没 錄到 這樣子 啊\ <s>
6. ...[8] [那 你%.-  
7. ...給 我們 一 人]-  
8. L: [唉 <s> 可是-- <s>
9. H: ...一百 塊 好 喔\ <s>
10. ...(1.6)談話費\ <s>

In this example, the terminal pitch directions of utterances 1, 4, 6, and 7 are all level. Thus, no <s> are inserted at the end of these utterances. This clue can be used to tell if there is a linguistic segmentation. That is, some impossible proposals will be filtered out.

Clue2: Repair Processing

As mentioned above, no <s> will be inserted within the speech repairs. Thus, this clue will help us to filter out some impossible proposal. Because of the limitation of page size, the detailed description of this processing can refer to [9-11].

Clue3: The Length of the Utterances

When the length of the utterance (the number of syllables in the utterance) is long enough, the speaker often finishes the utterance. Two examples are listed below.

104 L: [我^覺得是外面的人]在叫四= - <s>
196 L: ...那一次要是出大一點的敘漏啊?\ <s>

The length of the utterance 104 (196) is 11 (13). Thus, we predict that the speaker has completed his utterance.
So, $<$s$>$ is inserted at the end of the utterance. In our test, when the length of the utterance is greater than 7, $<$s$>$ is usually inserted at the end of the utterance.

**Clue4: Clue Patterns**

In Chinese conversation, some syllables often mark the end of a sentence. Typical examples are phrase-final particles (e.g., 吧 (a5, a)). These syllables can be used to increase the recall rate. That is, when these syllables occur at the end of the utterance, $<$s$>$ can be inserted at this position. Two examples are listed below.

\[
\begin{align*}
47 & \quad \text{...}(8)\text{老}^\wedge\text{外} \text{習慣} \text{就這樣} \text{吧=}\text{\langle} <s> \\
55 & \quad \text{...大家} \text{坐在一起} \text{幹嘛} \text{啊}?/ <s>
\end{align*}
\]

In these two examples, the words, i.e., “\text{老}^\wedge\text{外}” and “\text{幹嘛}”, often mark the end of a sentence. Thus, when these kinds of clues occur at the end of an utterance, $<$s$>$ can also be inserted at this position.

**Clue5: The Number of Inter-Utterances**

In our analysis, if too many utterances issued by other speakers are inserted between two utterances of the same speaker, the $<$s$>$ will be introduced. An example is listed below.

\[
\begin{align*}
142 & \quad \text{O:} \quad \text{...不要} \text{倒到} \text{桌面上} \text{就好} \langle <s> \\
143 & \quad \text{Z:} \quad \text{...為什麼} \text{會} \text{倒}^\wedge \text{？} \text{[R1,1-1,1]} \\
144 & \quad \text{...[^\wedge]倒在} \text{桌面上} \text{啊=}\text{\langle} <s> \\
145 & \quad \text{L:} \quad \text{[我}} \text{進度} \text{最後} \text{很} \text{\langle} <s> \\
146 & \quad \text{H:} \quad \text{...[1.2]你就} \text{在} \text{那麼} \text{一直} \langle <s> \\
147 & \quad \text{不} \text{知} \text{在} \text{幹什麼} \text{啊} <s> \\
148 & \quad \text{...[1.1]不} \text{努力} \text{這樣子} \text{啊=}\text{\langle} <s> \\
149 & \quad \text{Z:} \quad \text{[0]}@\text{@@} \text{@[R1,1-1,1]} <s> \\
150 & \quad \text{H:} \quad \text{你看} \langle <s> \\
151 & \quad \text{...臨時} \text{還} \text{佛腳} <s> \\
152 & \quad \text{就是} \text{很} \text{辛苦} \text{的} \text{啊=}\text{\langle} <s> \\
153 & \quad \text{O:} \quad \text{[嘆} \text{可} \text{是]} \text{劃拳} <s> [R1,2-1,2] \\
154 & \quad \text{...} \text{劃拳} \text{哪有} \text{說} <s> \\
155 & \quad \text{...還有一個人} \text{一} \text{罐} \text{的} \langle <s>
\end{align*}
\]

In this example, because 10 utterances issued by other speakers are inserted between two utterances (utterances 142 and 153) of the speaker O, the $<$s$>$ will be inserted at the end of the utterance 142. According to our analyses, when more than 2 utterances pronounced by other speakers interrupt the speech of a speaker, we insert $<$s$>$ at the end of the utterance.

**Clue6: Inter-Segmentation**

Although only 1.1% segmentations occur within the utterance, we also need to process them. However, this type of segmentation is relatively easy to detect because it usually occurs after the interjection. That is, when an interjection occurs, a $<$s$>$ can be put after it.

### 4. EXPERIMENTAL RESULTS

For evaluating the performance, some conditions are listed below.

- **Condition 1 (C1): clue 1**
- **Condition 2 (C2): clue 1 + clue 2 + clue 3**
- **Condition 3 (C3): clue 1 + clue 2 + clue 3 + clue 4**
- **Condition 4 (C4): clue 1 + clue 2 + clue 3 + clue 4 + clue 5**
- **Condition 5 (C5): clue 1 + clue 2 + clue 3 + clue 4 + clue 5 + clue 6**

Based on these conditions, Table 3 lists the experimental results for identifying linguistic segmentations.

<table>
<thead>
<tr>
<th>Clue</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>64.05%</td>
<td>69.51%</td>
<td>79.18%</td>
<td>86.44%</td>
<td>87.38%</td>
</tr>
<tr>
<td>Precision</td>
<td>93.38%</td>
<td>93.78%</td>
<td>94.36%</td>
<td>94.40%</td>
<td>94.46%</td>
</tr>
<tr>
<td>P &amp; R</td>
<td>0.76</td>
<td>0.8</td>
<td>0.86</td>
<td>0.9</td>
<td>0.91</td>
</tr>
</tbody>
</table>

In this table, three metrics, i.e., precision, recall and P&R, are used for evaluating the system performance. The precision is the number of proposed linguistic segmentations that were correctly identified compared to the number of total proposals. The recall is the number of proposed linguistic segmentations that were correctly identified compared to the total number of linguistic segmentations. In addition to precision and recall scores, we also report Van Rijssbergen’s F-Measure [12] which combines these two scores into a single measure. The F-measure (also called P&R) allows the differential weighting of precision and recall. With precision and recall weighted equally, it is computed by the following formula.

\[
F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

The experimental results show that C1 achieves a high precision rate, i.e., 93.38%, but it has a relatively low recall rate, i.e., 64.05%. Thus, more clues are needed. Obviously, the precision rate (93.78%) and the recall rate (69.51%) are all increased under condition C2. After clue 4 is added (C3), we find that this clue is very useful because the recall rate is increased about 10% (compare to C2). At the same time, the precision rate is also slightly increased (about 0.6%). When clue 5 is also used in our system (C4), the experimental results show that the recall rate is increased about 7% (compare to C3).
and the precision rate is slightly increased 0.04%. After all clues are used (C5), the final experimental results show that 94.46% precision rate and 87.38% recall rate can be achieved.

5. CONCLUSION

This paper employs several acoustic and prosodic clues to resegment acoustic segmentations for identifying linguistic segmentations. Based on six important clues, the experimental results show that a precision rate of 94.46% and a recall rate of 87.38% can be achieved. Although our system performs well in this task, it still has some spaces to improve. For example, the duration of pauses may be a good clue that can be used in this task. Intuitively, when the duration of pauses is long enough, a linguistic segmentation is formed. Currently, the pauses are only classified into three cases, i.e., long, medium, and short, in the spoken corpus. Thus, tagging detailed duration informations is needed and its effects should be investigated further.

ACKNOWLEDGMENTS

We are grateful to Professor Kawai Chui for her kindly providing the spoken corpus to us. Research on this paper was partially supported by National Council grant NSC88-2213-E-002-052.

6. REFERENCES


