An HPM-based Prosodic Analysis of Disfluencies for Spontaneous Mandarin Speech

Cheng-Hsien Lin¹, Chung-Long You¹, Chen-Yu Chiang², Yih-Ru Wang¹, Sin-Horng Chen¹

¹Dept. of Electrical Engineering, National Chiao Tung University, Taiwan
²Dept. of Communication Engineering, National Taipei University, Taiwan
jolin@itriti.org.tw, eychiang@mail.ntpu.edu.tw, yrwang@cc.nctu.edu.tw, schen@mail.nctu.edu.tw

Abstract

This paper presents a prosodic analysis of speech disfluencies on MCDC corpus labelled with two types of prosody tags, syllable-juncture break and syllable prosodic state, by a hierarchical prosodic modeling (HPM) approach proposed previously. The prosodic properties of two major types of speech disfluency, repetition and repair, are explored via examining the prosodic tags labelled on all MCDC utterances. The prosodic phrase structure of a disfluency is analyzed by examining the contextual break types. The prosodic state patterns of prosodic phrases in disfluencies are also examined. Lastly, the relation between prosodic properties of disfluencies and their pragmatic functions are discussed.

1. Introduction

Speech disfluencies, such as filled pauses, repetitions, repairs and particles, are frequent in spontaneous speech [1]. Exploration of the prosodic properties of disfluencies is useful for many spoken language processing tasks including sentence segmentation [2], disfluency identification [2-4], pragmatic function determination [5]. In the past, most prosodic analysis studies used the unlabeled prosodic-acoustic features directly to build statistical relations between prosodic features and disfluencies [1-5]. Only few studies [6-9] used prosody-labeled corpus with tags of breaks and tones by ToBI [10,11].

In this paper, a prosodic analysis of speech disfluencies on Mandarin Conversational Dialogue Corpus (MCDC) [12,13] is studied. The corpus is labeled in advanced with two types of prosody tags, syllable-juncture breaks and syllable prosodic states, by the hierarchical prosodic modeling (HPM) approach proposed previously [14,15]. The HPM adopts a 4-layer structure to describe the prosody hierarchy of an utterance. The HPM structure is composed of four constituents syllable/particular unit (SYL/PU), prosodic word (PW), prosodic phrase (PPh), and breath group/prosodic phrase group (BG/PG). Two types of prosodic tags, break type of syllable juncture and prosodic state of syllable, are used. A set of seven break types, B={B0, B1, B2-1, B2-2, B2-3, B3, B4}, is used to delimit prosodic constituents of these 4 layers. Here, B0 and B1 represent an intra-PW boundary with adjacent syllables/PU's being tightly and normally coupled, respectively; B2-1, B2-2 and B2-3 are PW boundaries with obvious F0 reset, perceived short pause and pre-boundary lengthening, respectively; and B3 and B4 represent major breaks with medium and long pause durations, respectively. Besides, two more dummy break types, Bp and Bm, are added for utterance beginning and end. Three types of prosodic states are employed to describe the variations of syllable pitch levels, durations, and energy levels. Since the affecting values of those prosodic states are the normalized and global-scaled prosodic features which have been compensated by local influential factors such as tone and base-syllable, they can be regarded as parameters describing the influences of high-level linguistic features.

In the study, the prosodic properties of two major types of speech disfluency, repetition and repair, are explored via examining the prosodic tags labelled on all MCDC utterances. The prosodic phrase structure of a disfluency is analyzed by examining the contextual break types. The prosodic state patterns of prosodic phrases in disfluency are also examined. Lastly, the relation between prosodic properties of disfluencies and their pragmatic functions are discussed.

The paper is organized as follows. Section 2 describes the MCDC corpus and the method. Section 3 discusses the experimental results. Some conclusions are given in the last section.

2. DATA and Methods

The MCDC corpus comprises eight dialogues collected by the Institute of Linguistics of Academia Sinica, Taiwan. Its total length is about eight hours. The eight dialogues were uttered by nine female and seven male speakers, and transcribed into Chinese texts with some tags including discourse marker (DM), particles, and pauses by professional linguist annotators. Some important spontaneous speech phenomena were also annotated, including disfluencies (long pause, repair, repetition, restart), pronunciation variations (contraction and lengthening), discourse-related items (marker and particle), sociolinguistic phenomena (English, Min-nan), and non-speech (laughter, breath, click).

Utterances of the corpus were segmented into syllable strings by forced alignment with manual error correction. Syllable durations and energy levels in each utterance were then mean-and-variance normalized to compensate the speaking rate and volume variations. F0s were first detected by the ESPS algorithm with manual error correction. Then, all F0 values of each speaker were mean-and-variance normalized. Last, each syllable pitch contour was represented by four coefficients of orthogonal transform using the Legendre polynomials up to the 3rd order [16].

We consider the general structures of disfluencies:

Repetition
<pre-context> # [repeat sequence] *...* [repeat sequence] @ <post-context>

Repair
<pre-context> # [reparandum] * <editing term> [reparan] @ <post-context>
Here, repeat sequence is the voice segment with multiple repetitions, reparandum is the voice segment to be abandoned, repetition is the correction segment, edition term is an optional meaningless segment, * denotes the interrupt point, # and @ denotes the beginning and ending points of the disfluency, pre-context and post-context are the voice segments preceding and following the disfluency, respectively. We will analyze the break type at *, # and @, and explore the relations with their pragmatic functions. The prosodic state patterns of the syllable sequence of the disfluency will also be discussed. In this study, the nine break types will be categorized into non-paused class B\text{non}={B0,B1,B2-3} and paused class B\text{pau}={B2-2,B3,B4,B5,B6} with B2-1 being regarded as a non-decisive transition class. B\text{non} is further divided into minor-pause class B\text{non}={B2-2} for PW boundary and major-pause class B\text{non}={B3,B4,B5,B6} for PPh and BG/PG boundaries.

3. Experimental Results

Two types of disfluencies, repetitions and repairs, are analyzed through examining the prosodic tags labeled on MCDC utterances by the HPM.

3.1. Repetition

A repetition disfluency is defined as a voice segment with a word or a syllable sequence being pronounced several times consecutively within an utterance. By examining their pragmatic functions, they can be classified into 5 types: restart of phrase/sentence (RES), agreement (AGR), emphasis (EMP), compound of two or more repetitions (COM), and planning or word retrieval difficulty (PLA). Table 1 shows their counts in MCDC.

Table 1: Counts of five types of repetitions in MCDC

<table>
<thead>
<tr>
<th>Prag. Function</th>
<th>RES</th>
<th>AGR</th>
<th>EMP</th>
<th>COM</th>
<th>PLA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>381</td>
<td>199</td>
<td>35</td>
<td>61</td>
<td>20</td>
<td>696</td>
</tr>
</tbody>
</table>

A. Restart (self-correction)

For an RES repetition disfluency, the repeat sequence is repeatedly pronounced several times with the last one being combined with post-context to form a phrase/sentence. So it is a kind of repair with reparandum being equal to reparandum. Since the last repeat sequence is the beginning part of a phrase/sentence, it is usually pronounced to be smoothly concatenated with the following word. So, the ending point @ tends to be non-paused (B\text{non}=60.9%, B2-1=31.5%, B\text{pau}=7.6%). Besides, the interrupt point * before the last repeat sequence tends to be paused (B\text{non}=49.9%, B2-1=32.5%, B\text{pau}=17.6%).

Some other characteristics of RES are listed below:
- Most RESs are of one repeat: 93.7%.
- For one-repeat RESs, most of their repeat sequences are monosyllabic (48.5%) or disyllabic (41.5%) words.
- All repeat sequences are words. Most of their POSs are Pronoun (42.2%), Verb (23.1%) and Adverb (19.1%).
- Most begin points with B\text{non} are corresponding to sentence beginning (B0).

An example is shown below:

**Example 1**

...ok！[B3] 那 # 我，[B2-1] * 我 @ 叫 順彥辰。[B2]

...ok！[B3] That # 1 * I @ named lai-yan-chen。[B2]

Fig. 1 displays the pitch state patterns of RESs for two cases of break type at # being (a) B\text{non} and (b) B3/B4.

B. Agreement

The syllable “Dui” in Mandarin is frequently used in conversation to show the agreement with something or to give a positive answer to someone, and then take the speaking turn back. It may also combine with a particle like “A” (Dui-A) to present the same semantic meaning. In MCDC, we found that “Dui” and “Dui-A” form the majority of AGRs, and they mainly repeat once (31.2%), twice (46.2%), or triple (11.6%).

The characteristics of AGR are listed below:
- Most interrupt points * are non-paused (B\text{non}=91.9%, B2-1=62.2%, B\text{pau}=1.9%).
- They appear at the beginning of an utterance with very high probability: 64.3%.
- For one-repeat AGRs, the ending point @ tends to be a transition break (B\text{non}=21%, B2-1=51.6%, B\text{pau}=19.4%); while for two- and three-repeat AGRs, they tend to be paused (B\text{non}=57.4%, B2-1=36.5%, B\text{pau}=6.1%).
- The POS of the first word in post-context has high probabilities to be Adverb and Noun (27.4% and 27.4% in one repeat, 19.6% and 20.7% in two repeat). Usually, they represent the beginning words of sentences.
- We also observed that speakers may insert a Particle/Marker (17.4%) at @ for “Dui-Dui-Dui”.

Two examples are shown below:
Example 2

---

…# 對！* 對！* 對！[B3] @ 啊！[B2-1] 可是呢！[B3]
電台也不贊乘客的錢的。[B2]
…# dui! * dui! * dui! [B3] @ particle-A! but particle-Ne ! [B3]
radio-station also not earn passenger de money de * [B3]

Example 3

---


Mean pitch pattern of “Dui-A Dui-A” is displayed in Fig.2 for the case of break type at # being B2. In the figure, the mean pitch pattern start at a high pitch level owing to utterance beginning, then falling to meet the disyllabic word pitch pattern of the first “Dui-A”, then rising to emphasize the second “Dui-A”, and lastly falling to start the post-context. It is worthy to note that “Dui-Dui-Dui” has more stable durations, while a significant drop in duration state can be observed at the beginning syllable of the second “Dui-A”.

C. Emphasis

For an EMP repetition disfluency, the repeat sequence is repeatedly pronounced several times to emphasize something mentioned previously (22.9%), currently (37.1%), afterward (8.6%), or to respond (14.3%) or request (11.4) with strong mood. Usually, repeat sequences are pronounced smoothly with no significant pauses existed in their beginning (B0=84.6%) and interruptive junctures (B2=66.1%), whereas the ending point @ tends to be paused (B0=45.7%).

An example is shown below:

Example 4

---

…MHM！[B3] 所以現在#很多，*很多，@那個，[B3]…
…MHM！[B3] so now # many，* many，@ that，[B3]…

D. Compound

For a COM repetition disfluency, two or more repetitions occurred consecutively. It can be regarded as a repair with the last repetition being taken as reparandum and all preceding repetitions as reparandum. An example is shown below:

Example 5

---


high [B2-1] in [B3] nan-hai school-park opposite…

The characteristics of COM are listed below:

- Most interrupt points * are non-paused (B0=47.7%，B2-1=22.5%，B2=29.8%)．

- The beginning point # tends to be B2 or non-paused (B0=34.5%，B2=19.7%，B2=30.0%，B3=38.9%)．

- The ending point @ tends to be non-paused (B0=45.9%，
B2=31.1%，B2=35.0%)．

Fig. 3 displays the pitch state patterns for Example 5. It is found from this figure that rising-falling-rising pitch pattern is observed in this COM repetition. Moreover, high declining pitch pattern is found at the post-contexts of it. This shows a restart occurs at the beginning of the post-context of a COM repetition disfluency.

E. Planning

For a PLA repetition disfluency, the speaker has a word retrieval difficulty. The whole disfluency is given up without any useful function in the post-context. PLAs in MCDC can be classified into two types. For the first type, a restart is occurred at post-context to resume the normal speech flow. Example 6 shown below is of this type. For the second type, the speaker cannot find a suitable speech to follow the PLA repetition disfluency so that the planning is lasted and sustained in the post-context. In this case, the post-context is rife with fillers. Example 7 is belonging to this type.

Example 6

---

[B2-2] 就算，[B4]…

…EI！[B2-1] then [B2-2] # I，* I [B2-2] @ because have me，
[B2-2] even-if，[B4]…

Example 7

---


The characteristics of PLA are listed below:

- Most interrupt points * are paused (B0=15.0%，B2-1=5.0%，B2=80.0%)．

- The beginning point # tends to be non-paused (B0=50.0%，
B2-1=15.0%，B2=30.0%，B3=5.0%)．

- The ending point @ tends to be paused (B0=25.0%，
B2=25.0%，B2=40.0%，B3=10.0%)．

Fig. 4 displays the mean patterns of syllable pitch and duration state for PLAs with one repeat of monosyllabic word. It is found from the figure that both the pitch and duration...
state patterns at the second monosyllabic word (Syl-2) are high-valued. This shows that Syl-2 is accented with pitch raising and duration lengthening. Besides, we also find that hesitations are frequently found at @.

Example 8

...不 是 # 好 \[B2-1]* 变好 \[B2-1]@ 就是 變更 \[B2-1] 壞。 \[B3]... 
...not is # good \[B2-1]* become-good \[B2-1]@ that is become-more \[B2-1] bad \[B3]...

Fig. 5 displays the mean patterns of syllable pitch and duration states for repairs with monosyllabic reparandum (Rep-1) and disyllabic reparan (Cor-1, Cor-2). From Fig. 5(a), a declining pitch pattern is observed when * is B\textsubscript{50}, whereas a falling-rising pattern is found when * is B2-1. From Fig. 5(b), significant contraction and lengthening of syllable duration are found respectively at reparandum and reparan when * is B\textsubscript{50}.

4. Conclusions

In this paper, the prosodic property of two major disfluencies, repetition and repair, in MCDC is explored based on the prosodic tags labeled by the HPM. Prosodic phrasing and prosodic state patterns are analyzed for different functional classes of disfluencies. Many meaningful findings are presented.

5. Acknowledgements

This work was supported by MOST of Taiwan under the contract “MOST 101-2221-E-009-078-MY3”. The authors want to thank Dr. S. C. Tseng of Academia Sinica for providing the MCDC corpus.
6. References


