PROSODIC ANALYSIS OF FILLERS AND SELF-REPAIR IN JAPANESE SPEECH

Felix C M. Quimbo   *Tatsuya Kawahara*   *Shuji Doshita*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

The prosodic features of filled pauses (fillers) and self-repair are investigated with a view towards the detection of disfluencies. First, we compare the prosodic features of typical fillers and their fluent homonyms using read sentences of identical phoneme sequences. It is confirmed that the fillers (1) have at least 2 times longer duration than their non-disfluent counterparts, (2) tend to be followed by definitely longer pauses, and (3) have much smaller movement in their pitch contours. Then, the spontaneous fillers segmented out from a dialogue corpus are also analyzed. The same tendency is confirmed, but some samples lie halfway between the read fillers and their fluent homonyms. The abruptly cut-off endings in self-repair are also analyzed by comparing with the ordinary endings of words. It is found that a short phoneme ending coupled with a relatively short succeeding pause indicates the abrupt cutoff.

1. INTRODUCTION

The presence of disfluencies is one of the major problems in the recognition of spontaneous speech. This is because current speech recognizers are based on language and acoustic models which assume grammatical and well articulated input. Since disfluent segments are not likely to be matched well with the conventional acoustic-phonetic model, even if given their transcriptions, the prosodic features can be helpful to detect them. For example, human beings can recognize the disfluencies out of unfamiliar foreign language speech from the prosodic cues.

We deal with filled pauses (fillers) and self-repair, which are the two most common instances of disfluencies. Specifically, we investigate how prosodic features such as speech segment duration, pause length, pitch as well as phonetic information can be used in the detection of fillers and self-repair.

2. PROSODIC DIFFERENCES OF FILLERS AND THEIR HOMONYMS

Since there are only a limited number of words used as fillers, we take a bottom-up approach. Namely, fillers are recognized as individual words in the input. However in Japanese, where homonyms are common, these words have to be differentiated from their homonyms using prosodic information. We compare the prosodic features of the frequent fillers in Japanese ‘ano’, ‘eeto’, ‘ee’ and ‘to’ with their homonyms in both read speech and spontaneous speech.

We prepared 5 sets of sentences with almost identical phoneme sequences. Each set contains one filler and one fluent homonym. The pairs are as follows. English translation is attached for reference.

```
1a. ano, yotei ha touroku sarete imasu ka.
    (Uh, is the appointment scheduled)
 b. ano yotei ha touroku sarete imasu ka.
    (Is that appointment scheduled?)
2a. tto, kyou ha dou desu ka.
    (Uuhm, how about today?)
 b. Tokyo ha dou desu ka.
    (How about Tokyo?)
3a. eeto, kyou ha dou desu ka.
    (Uuuhm, how about today?)
 b. ee, Tokyo ha dou desu ka.
    (Uuh, how about Tokyo?)
4a. eeto, moji de kaite kudasai.
    (Uuuhm, please write it with words.)
 b. e to moji de kaite kudasai.
    (Please write it with pictures and words.)
5a. eeto, ten to iu kissaten ga arimasu.
    (Uuuhh, there is a coffee shop called Ten.)
 b. eeto ten to iu kissaten ga arimasu.
    (There is a coffee shop called Eight Ten.)
```

A total of 11 samples for every set was recorded. By listening to these samples, it is observed that (1) the duration of fillers are usually longer than ordinary words, (2) relatively long silences succeed them, and (3) they have a relatively flat pitch.

We can easily imagine that speakers prolong pronunciation of fillers, trying to buy time or keep his

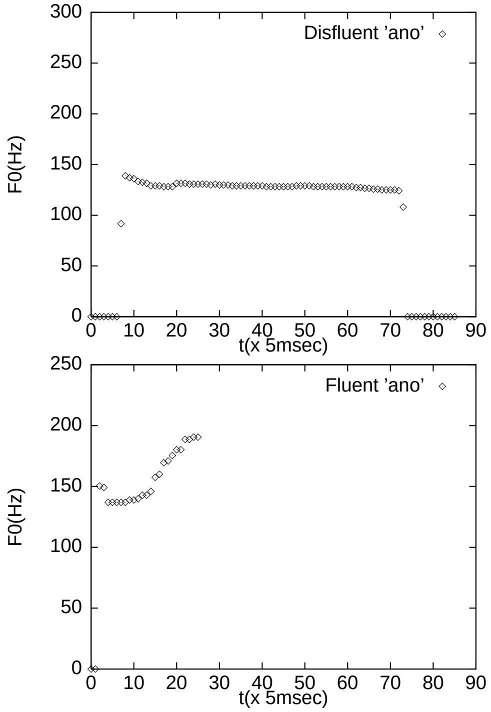

Figure 1: Pitch contour for fluent and disfluent 'ano'

Table 1: Comparison of fluent and disfluent 'ano'

(sample No.)		filler 'ano' (1a)	word 'ano' (1b)
duration	range (sec)	0.28-0.57	0.14-0.22
	average (sec)	0.42	0.17
pause length	occurrence	82%	0%
	range (sec)	0.02-0.47	N.A.
	average (sec)	0.30	N.A.
pitch movement	range (Hz)	2.8-20.5	8.0-30.9
	average (Hz)	10.4	14.9

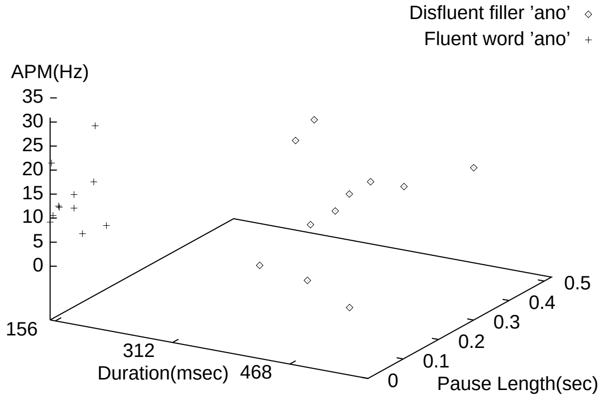

Figure 2: Duration vs. Pause length vs. Pitch movement for 'ano'

turn. As Kawamori et al. pointed out[1], fillers have a flat long contour and does not have a sharp drop at the end. An example of pitch contour for ordinary word 'ano' and filler 'ano' are shown in Figure 1.

In this paper, quantitative analysis is performed using the following prosodic features

1. Duration of voiced segments

   Fillers are manually segmented out of utterances and the duration of the segments is measured.

2. Pause length

   Similarly, voiceless portions are segmented out manually and the duration is measured.

3. Average Pitch Movement (APM)

   Pitch is extracted every 5 msec with the autocorrelation analysis and smoothed by windowing. Pitch movement is defined to be the absolute difference in value between two adjacent frames. The average of the difference during the particular word/filler segment is then calculated.

2.1. Analysis of Read Speech Samples

At first, disfluent filler 'ano' (1a) is compared with fluent pronoun word (1b). Table 1 lists statistics of prosodic features for 11 samples each. When we plot the values in three-dimensional space of duration, pause length and average pitch movement (APM) in Figure 2, the discriminant ability of these parameters is clear.

Similarly, the filler 'eeto' (3a,4a,5a) is compared with the fluent counterparts (4b,5b) in Table 2. The plot for their distributions in three dimensions looks much like the one for 'ano'. In this case, we also make use of the intermediate pause that is inserted right after 'ee'. In the fluent words, this pause corresponds to the closure in articulating the consonant 't', thus cannot be so long. In the disfluent filler, however, the pause tends to be significantly longer. The plot by the intermediate pause length vs. average pitch movement (APM) as shown in Figure 3 clearly discriminates the filler and the fluent homonym (4b).

With the analysis of all 110 samples of read sentences, the following results are derived.

1. Duration

   The average duration of fillers is at least 2 times longer than that of the fluent counterparts. There is very little overlapping in their distributions.

Table 2: Comparison of fluent and disfluent 'eeto'

(sample No.)		filler (3a,4a,5a)	noun (5b)	noun-conj. (4b)
duration	range	0.34-1.09	0.22-0.39	0.19-0.33
	ave.	0.63	0.32	0.24
pause length (after 'ee')	occur.	100%	100%	100%
	range	0.08-0.41	0.04-0.08	0.04-0.09
	ave.	0.17	0.06	0.06
pause length (after 'to')	occur.	88%	100%	9%
	range	0.09-1.17	0.06-0.11	0.15
	ave.	0.32	0.07	0.15
pitch movement	range	12.1-42.7	37.5-66.5	47.4-101.1
	ave.	27.1	51.8	72.2

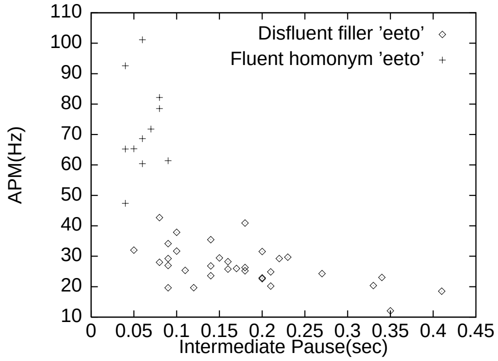

Figure 3: Intermediate pause length vs. Average pitch movement (APM) for 'eeto'

2. Pause length

   Pauses are present after the majority of fillers. Pauses can be present in the fluent counterparts, but the average duration is much shorter than the ones after the fillers with very little overlapping in the distributions.

3. Average Pitch Movement (APM)

   The APM of fillers is significantly smaller than that of the fluent homonyms, although there is overlapping in their distributions.

2.2. Analysis of Spontaneous Speech Samples

Next, we apply the analysis to fillers embedded in spontaneous utterances, as the prosodic features of spontaneous speech are different from those of read speech[2]. For spontaneous speech, we use samples collected from our scheduling task corpus[3]. This corpus has 556 utterances in total. Of these, 216 utterances have a filler and 48 contain self-repair. However, there are only 3 samples of 'ano'. Thus, we focus on the most frequent filler 'eeto'.

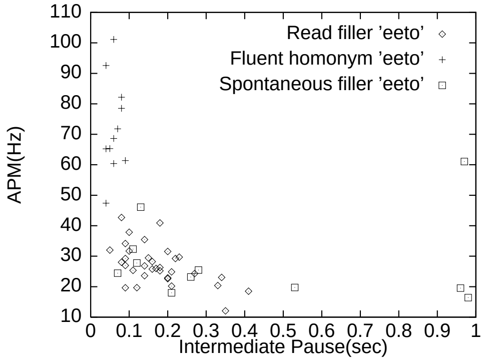

Figure 4: Intermediate pause length vs. Pitch movement for read and spontaneous 'eeto'

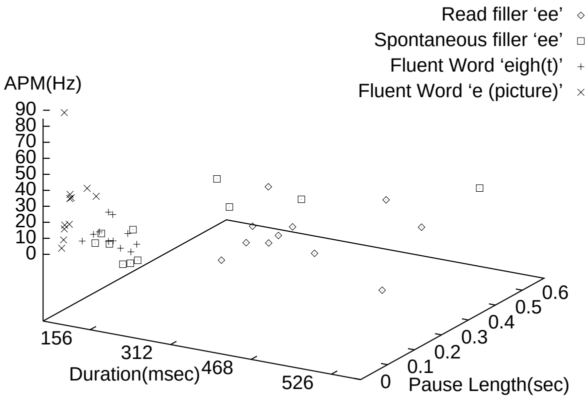

Figure 5: Duration vs. Pause Length vs. Pitch Movement for read and spontaneous 'ee'

We segment out 11 samples of the filler 'eeto' from the corpus and calculate their duration, pause length and average pitch movement (APM). They are compared with the previous samples of read speech, and superimposed in Figure 4. It is confirmed that the distributions of the spontaneous filler 'eeto' are much the same as that of read speech, and they can still be discriminated from the fluent words. However, some of the spontaneous one have extremely long pauses, because the subjects were truly hesitating.

The features of the segment 'ee' are also calculated and plotted in Figure 5. It is observed that some of the spontaneous filler 'ee' lies halfway between the read filler and the non-disfluent one. This may be because the read fillers are articulated consciously to be ideal ones.

The results show that spontaneous fillers have the same characteristics as the read fillers, but some samples are not as stereo-typically 'filler-like' as the read ones. It suggests the possibility of a disfluent filler detector that puts priority on a high precision rate rather than a high recall rate.

3. PROSODIC FEATURES OF ABRUPT ENDINGS IN SELF-REPAIR

Next, we deal with self-repair. Nakatani and Hirschberg[4] pointed out the three biggest indicators of the presence of self-repair.

1. presence of a pause
2. presence of a fragment
3. presence of a filled pause (filler)

A pause (1) can easily be detected easily. The detection of a filled pause (3) has just been discussed in the previous section. We will now look closely at the prosodic characteristics of fragments (2), which are caused by abruptly cutoff of words.

Here, we use the same spontaneous samples of the scheduling task corpus for analysis. There, 39.6% (19/48) of the self-repairs have fragments.

In Japanese, the dominant syllabic structure is the C+V (consonant + vowel) form. The property might be helpful for detecting fragments. If an input ends with a consonant, then it must not be an ordinary word ending, ergo it is a fragment. In our corpus, however, only 3 out of 19 samples end with consonants.

Therefore, prosodic features of the abrupt endings are investigated. Specifically, the duration of the final phoneme is measured manually. It was our hypothesis that short phoneme endings indicate the abrupt endings which Hindle used as his marker for the Interruption Site[5]. While the duration of all three consonants is around 30 msec, that of vowels has wide range from 30 to 300 msec.

These fragmentary endings are compared to ordinary word endings. The ordinary endings are picked up from the phrase boundaries, and the duration of the final vowels is measured. The duration of the following pauses is also measured. We plot the final vowel duration against the pause length in Figure 6 There is a region in the lower left side which consists mostly of fragmentary endings. Although fragmentary endings are also seen in other regions, the fragments in the lower left region are hardly mixed with ordinary word endings. The ordinary word or phrase usually ends with a sufficiently long vowel. If the final vowel is observed as short, it is due to the weak power level of fading out portion of the utterance and must be followed by a long pause.

Thus, short phoneme endings coupled with relatively short succeeding pauses indicate the abrupt cutoff which characterizes fragments of self-repair.

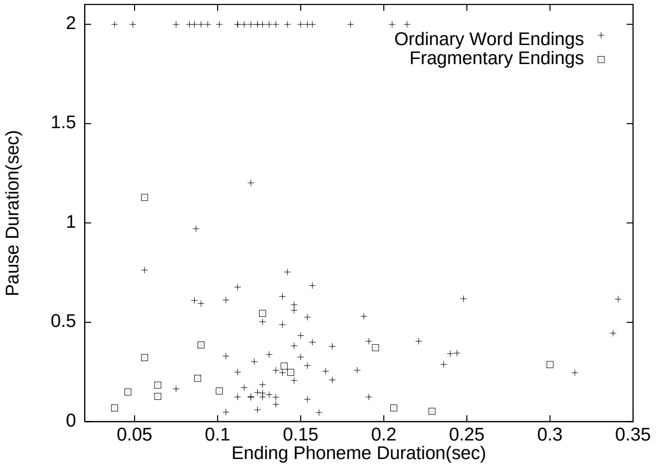

Figure 6: Duration of ending phoneme vs. pause length for ordinary words and fragments

4. CONCLUSION

We have investigated the prosodic features of fillers and self-repair. Although the number of samples is not large enough, the obtained result gives perspective to implement a stable discriminant mechanism that phonetic information alone cannot realize. They can be used to detect the disfluencies in early stage of the recognition or to give confidence to the recognized results.

References

[1] M. Kawamori, T. Kawabata, and A. Shimazu : A Phonological Study on Japanese Discourse Markers, *IEICE Tech. Report NLC95-26* (1995).

[2] H. Fujisaki : From Read Speech to Spontaneous Speech: Problems and Approaches in the Processing of Prosody, *ATR Intl. Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing* (1995).

[3] `http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus`

[4] C. Nakatani and J. Hirschberg : A Corpus-Based Study of Repair Cues in Spontaneous Speech, *Journal of Acoustical Society of America*, pp.1603-1616 (March 1994).

[5] D. Hindle : Deterministic Parsing of Syntactic Non-fluencies, *ACL* (1983).