Detection of Fillers Using Prosodic Features in Spontaneous Speech Recognition of Japanese

Keikichi Hirose[1] Yu Abe[2] & Nobuaki Minematsu[2]

[1]Dept. of Inf. and Commu. Engg, School of Inf. Science and Tech.
[2]Dept. of Frontier Informatics, School of Frontier Sciences
University of Tokyo, Tokyo, Japan
{hirose, yu-abe, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A new scheme of detecting fillers in spontaneous speech recognition process was developed. When a filler hypothesis appears during the 2nd pass decoding of a speech recognizer with two-pass configuration, a prosodic module checks the morpheme which is hypothesized as a filler and outputs the likelihood score of the morpheme being a filler. When the likelihood score exceeds a threshold, a prosodic score is added to the language score of the hypothesis as a bonus. The prosodic module is constructed using five-layered perceptron. With inputs on prosodic features of current, preceding and following morphemes, the perceptron calculates the filler likelihood. A comparative recognition experiment with and without the prosodic module was conducted for 100 utterances of spontaneous speech, which are included in the corpus of academic meeting presentations of the Corpus of Spontaneous Japanese. Seven fillers originally miss-recognized as non-fillers are correctly recognized as fillers when the prosodic module is used. No fillers originally recognized as fillers are wrongly recognized as non-fillers. Although a few non-filler morphemes are miss-recognized as other non-filler morphemes by the introduction of the prosodic module, they can be corrected by properly setting parameters of the 2nd pass search process. These results indicate the proposed scheme can improve the performance of spontaneous speech recognition.

1. Introduction

In view of the importance of prosodic features in human speech perception, a rather large number of research works have already been devoted for developing modules of prosodic event detection and for incorporating them into speech recognition process. The authors have been developing several methods for continuous speech recognition along this line, and realized certain improvements in the recognition rates [1-4]. However, in most of the works, including ours, recognition of text-reading style speech was addressed. In such cases, large amount of data are usually obtainable for training acoustic and language models, and a high recognition performance is obtainable without relying on prosodic features. Therefore the effect of the prosodic modules comes unclear in the total recognition process.

The situation may be different, when it comes difficult to obtain enough data, such as the case of spontaneous speech. Spontaneous speech may include number of irregularities, such as hesitations (fillers/pauses), re-statements, and so on, which may largely degrade speech recognition performance. Since these parts show prosodic features different from other parts (of normal utterance) [5], they may be detectable by viewing fundamental frequency (F_0) contours, power/amplitude contours, and segmental duration patterns, and their information may contribute to the final recognition results. The most naïve way of using filler information for speech recognition is to detect filler portions independently and skip those portions from the recognition process. However, this may not work well, because the filler detection with prosodic features may include a certain number of errors even with sophisticated schemes.

From this point of view, we have developed a new method of using filler information for continuous speech recognition: to calculate the likelihood of fillers appearing in the decoding process of speech recognition using prosodic features, and, if the likelihood is high, increase the score of the hypothesis with the fillers. As for the calculation of likelihood, a neural network was adopted, though other options were also possible.

The rest of the paper is organized as follows: The outline of the proposed method is explained in Section 2. After a short explanation on the speech material in Section 3, the neural network for the calculation of filler likelihood (prosodic module) is explained with experimental results in Section 4. Results of speech recognition experiments are shown in Section 5. Section 6 concludes the paper.

2. Configuration of the Method

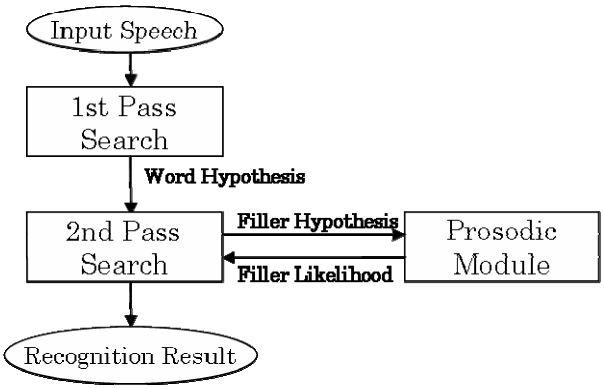

Figure 1: Total configuration of the proposed method.

Figure 1 shows the total configuration of the proposed method. As for the speech recognition engine, Julius developed as an open-software for continuous speech recognition is used. The engine conducts quick coarse search (1st pass search) first and then conducts detailed search backwoods (2nd pass search) [6]. The 1st pass is the frame synchronous beam search with

(morpheme) bi-gram language model and the 2nd one is N-best stack decoding search with (backward) tri-gram language model. When calculating the likelihood of hypotheses, the weight of the language score to the acoustic score was set to 8.0 throughout the current experiment. The prosodic module calculates probability of a morpheme being a filler morpheme (henceforth, filler likelihood score). Although the module can calculate the filler likelihood scores for all the morphemes included in the input utterance, in the current method, it needs to calculate only for those hypothesized to be fillers in the 2nd pass search process. The language score is changed depending on the result of the prosodic module. Our preliminary experiment showed that reducing the language score when the likelihood score being low degraded the final recognition rates. Taking this into account, a certain value (bonus) is added to the language score only when the filler likelihood score exceeds a threshold. Henceforth we call this value as the prosodic score. Since there is no clear difference in the recognition performance, whether the prosodic score is changed according to the filler likelihood score or is kept constant, we set it to a constant value in the current paper. The threshold and the prosodic score are respectively set to 0.5 and 5 in the experiments shown in section 5. Surely, if we reduce the prosodic score, the number of false filler detection may decrease, but the number of filler recovery by the prosodic module may also decrease.

3. Speech Material

The speech material used for the experiments is 100 utterances (including one or more fillers) by 7 males and 6 females, which are selected from the corpus of academic meeting presentations included in the Corpus of Spontaneous Japanese (CSJ) prepared under a national project [7]:

http://www2.kokken.go.jp/~csj/public/index.html

In the original corpus, all the utterances of each speaker are recorded in a file. So, we first segmented it into utterances and then selected 100 utterances so that each of them includes one or more fillers, and does not include any restatements or coughs. In the entire CSJ corpus, 160 filler variations are included, while 17 variations are included in the selected 100 utterances. The numbers of fillers in the 100 utterances sorted in the order of frequency are, 185 /eH/, 82 /e/, 16 /sonoH/, 14 /ma/, 13 /maH/, 12 /eQto/, 11 /ano/, etc. (Symbols "H" and "Q" mean elongation of previous vowel and gemination, respectively.)

4. Prosodic Module

The prosodic module is constructed as a 5-layered perceptron with 3 middle layers, each of which has 20 units. These numbers were decided through some preliminary experiments. The input and output layers have 10 and 1 units, respectively. One unit of input layer accepts each of 10 input parameters listed in Table 1. The output layer unit outputs the filler likelihood in the range between 0 and 1.

Figure 2 shows an example on how fillers appear in the F_0 contour of utterance. It is clear that they have low and level contours. Taking this feature into account, four F_0-related parameters in Table 1 are included into the input parameters. Lengths of immediately preceding and following silences are included in the input parameters, because they frequently co-occur with fillers as shown again in Fig. 2. In the current method, silences are detected simply searching periods whose waveform amplitudes do not exceeds a threshold.

An experiment of filler detection was conducted for the 100 utterances. First, all the utterances are segmented into phonemes by the forced alignment, and then their F_0's were extracted in order to calculate the input parameters. Twelve utterances were discarded where the input parameters were not properly extracted because of errors in segmentation and/or pitch extraction. Then, the rest 88 utterances (of 6 male and 6 female speakers) were divided into 76 utterances for training and 12 utterances (one utterance from each of 6 male and 6 female speakers) for testing. They include 306 fillers (in total of 2846 morphemes) and 39 fillers (in total of 420 morphemes), respectively. Figure 3 shows the error convergence according to the number of training cycles. From this result, the training cycle 50 was selected for the experiments. Table 2 shows the filler detection rates for each speaker/utterance, when morphemes with filler likelihood scores larger than 0.5 are assumed to be fillers. It also shows the filler/non-filler identification rates for all 420 morphemes of the testing utterances. As a whole, 29 fillers are correctly detected out of 39 fillers, while 13 fillers are incorrectly detected out of 381 non-filler morphemes.

Table 1: Input parameters for filler identification. The F_0's and amplitudes are those for the (current) morpheme in question other than specified. All the F_0 values are processed in a logarithmic scale.

Number of phonemes
F_0 range (Maximum F_0 minus minimum F_0)
Gradient of F_0 contour when approximated with a line
F_0 average divided by F_0 average of the utterance
Difference in F_0 between the last vowel of current morpheme and the first vowel of following morpheme
Length of immediately preceding silence
Length of immediately following silence
Gradient of amplitude pattern of the last vowel when approximated with a line
Average amplitude of vowel parts
Duration of the last vowel of current morpheme divided by that of average phoneme length of the utterance

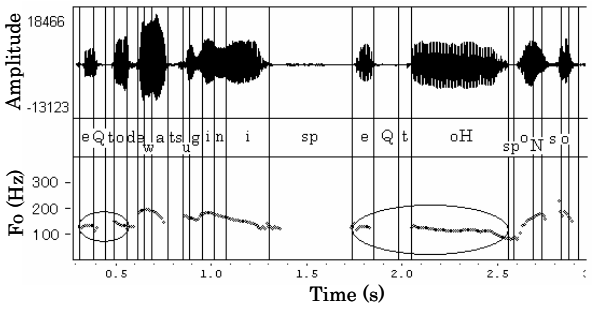

Figure 2: Waveform (upper panel) and F_0 contour (lower panel) for the utterance "eQto dewa tsugi ni eQtoH oNso ([Filler] Then, next [Filler] a phoneme…)" by a male speaker. The underlined morphemes are fillers. The circled parts of F_0 contour are those corresponding to the fillers. "sp" means a short pause.

Since the proposed method only checks the morphemes, which are selected as filler candidates in the 2nd pass search of Julius, the prosodic module does not work on the morphemes with no filler possibilities in the 1st pass. Therefore, it is of interest to compare the fillers detectable by the prosodic module and those included in the recognition hypotheses. Table 3 shows such data. The acoustic and language models used for the speech recognition are those included in the CSJ corpus [8]. Ten fillers out of 39 fillers included in the test utterances are included in the recognition hypotheses but not detectable by the prosodic module. All of these fillers are correctly included in the final recognition result after the 2nd pass. Taking this situation into account, we decided not to decrease the likelihood of the recognition hypothesis with filler(s), even if the prosodic module did non-filler judgments.

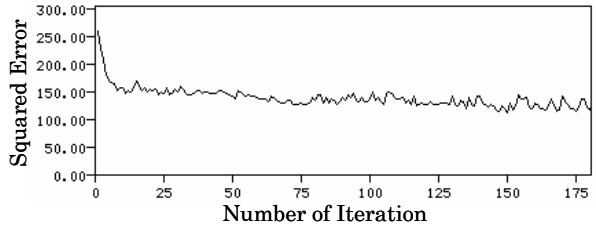

Figure 3: Total sum of the squared error versus number of training cycles (iterations).

Table 2: The numbers of fillers correctly detected and total numbers of fillers and non-fillers correctly identified. These are listed before "/" while the numbers after "/" indicate the number of samples. The numbers in parentheses are filler detection rates (%) and filler/non-filler identification rates (%). Utterances by Males 1 and 2, and females 3 and 6 are not included in the training corpus. So the results indicated in italic are speaker open.

Utterance	Filer Detection	Filler/Non-filler Identification
Male 1	*3/5 (60)*	*43/48 (90)*
Male 2	*4/4 (100)*	*37/37 (100)*
Male 3	1/2 (50)	21/23 (91)
Male 4	2/3 (67)	30/31 (97)
Male 5	5/6 (83)	38/41 (93)
Male 6	1/2 (50)	21/23 (91)
Female 1	2/3 (67)	23/25 (92)
Female 2	2/3 (67)	38/40 (95)
Female 3	*3/3 (100)*	*30/31 (97)*
Female 4	1/3 (33)	29/31 (94)
Female 5	4/4 (100)	58/59 (98)
Female 6	*1/1 (100)*	*29/31 (94)*
Total	29/39 (74)	397/420 (95)

Table 3: Numbers of fillers in the training corpus sorted from the viewpoints if they are detected by the prosodic module and by the speech recognizer (Julius). Symbols "O" and "X" mean detected and not detected, respectively.

Prosodic Module	Julius		Number
	Final Recognition Result	Included in Recognition Hypotheses of 1st Pass	
X	X	X	0
X	X	O	0
X	O	O	10
O	X	X	0
O	X	O	6
O	O	O	23

5. Experiments

Speech recognition experiments were carried out for the 100 utterances using two versions of recognizer: one with prosodic module (proposed recognizer/method) and the other not (baseline recognizer/method). As explained already, the baseline recognizer is Julius for the spontaneous speech provided by the CSJ project [8]. The acoustical (phone hidden Markov) models were trained using 486 hours of academic meeting presentations by 2496 people included in the CSJ corpus. The 100 utterances are included in these training speech samples. The language models were trained using transcriptions of 2592 lectures, which include 6.6×10^6 morphemes. Table 4 shows the conditions of acoustic analysis.

Table 4: Conditions of acoustic analysis.

Sampling frequency	16 kHz
Pre-emphasis	$1 - 0.97\,z^{-1}$
Window	25 ms Hamming
Frame shift	10 ms
Feature vector	12 MFCC + 12 ΔMFCC + Δpower

The utterance "kasetsu ga e shiji sa re mashi ta (The hypothesis was accepted.)," by Female 5 (in Table 2) was recognized as "kasetsu ga ninshiki (recognize) sa re mashi ta." by the baseline recognizer, while it was recognized as "kasetsu ga e shi (do) sa re mashi ta" by the proposed recognizer. It is clearly shown filler /e/ (underlined in the example) is correctly recognized in the version with the prosodic module. Improvements at non-filler morphemes are also observable in the utterance "e kochira ga eH hana no aru (This one is with a nose…)" by Male 4, which was miss-recognized as "e kochiragawa (this side) eH hana no aru" by the baseline recognizer. It was correctly recognized when the prosodic module was introduced.

Table 5 summarizes changes in the recognition results caused by the introduction of the prosodic module. Seven fillers, miss-recognized by the baseline method as non-filler morphemes, are correctly recognized by the proposed method,

while no fillers correctly recognized by the baseline method are miss-recognized by the proposed method. In the 100 utterances, a total of 389 fillers are included and 349 of them are detected by the baseline method. Therefore, 356 fillers are detected by the proposed method. Three non-filler morphemes correctly recognized by the baseline recognizer are miss-recognized by the introduction of the prosodic module. These errors can be avoided by decreasing the prosodic score, but improvement in filler detection also degraded. This type of miss-recognition is tightly related to the (sophisticated) search algorithms of the 2nd pass, such as: when a hypothesis survives beyond a threshold, hypotheses with shorter lengths are terminated. Because of these algorithms, the best hypothesis selected by the 2nd pass is not guaranteed to be really the best one. It is confirmed that all the three morphemes miss-recognized by the introduction of the prosodic module are correctly recognized in the "really" best hypotheses.

Table 5: Numbers of morphemes where the recognition results are changed by the introduction of the prosodic module. "Baseline" and "Proposed" indicate speech recognizers without and with prosodic module, respectively.

(Baseline → Proposed)	Filler	Non-filler
Incorrect → Correct	7	4
Correct → Incorrect	0	3

6. Conclusion

A new method of detecting fillers in spontaneous speech during the speech recognition process was developed. It checks the feasibility of filler hypothesis by viewing the prosodic features of current and surrounding morphemes, and adds a bonus to the hypothesis if the feasibility is high enough. Experiments on the utterances selected from the corpus of academic meeting presentations in CSJ showed that some errors in filler detection in the baseline method were recovered by the proposed method with no co-occurring degradation. Although some errors arose for non-filler morphemes, they were due to the search algorithm of the 2nd pass of the baseline recognizer Julian, and could be recovered by changing the algorithm. Further experiments are planned for increased number of utterances. It is known that speakers use fillers rather differently in their spontaneous utterances. Adaptation methods to cope with this variation are also in the scope of our future work.

The work is partly supported by Grant in Aid for Scientific Research (16650034).

7. References

[1] Hirose, K. and Iwano, K., "Detection of prosodic word boundaries by statistical modeling of *mora* transitions of fundamental frequency contours and its use for continuous speech recognition," *Proc. IEEE ICASSP, Istanbul*, 3, 1763-1766, 2000.

[2] Lee, S., Hirose, K., and Minematsu, N., "Incorporation of prosodic module for large vocabulary continuous speech recognition," *Proc. ISCA Tutorial and Research Workshop on: Prosody in Speech Recognition and Understanding, Red Bank*, 97-101, 2001.

[3] Hirose, K., Minematsu, N. and M. Terao, "Statistical language modeling with prosodic boundaries and its use for continuous speech recognition," *Proc. ICSLP, Denver*, 2, 937-940, 2002.

[4] Hirose, K., and Minematsu, N., "Use of prosodic features for speech recognition," *Proc. ICSLP, Jeju*, 2, 1445-1448, 2004.

[5] Quinbo, F., Kawahara, T., and Doshita, T., "Prosodic analysis of fillers and self-repair in Japanese speech," *Proc. ICSLP, Sydney*, 3313-3316, 1998.

[6] Lee, A., Kawahara, T., and Shikano, K., "Julius – an open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH, Aalborg*, 1691-1694, 2001.

[7] Maekawa, K., "Corpus of Spontaneous Japanese: Its design and evaluation." *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo*, 7-12, 2003.

[8] Kawahara, T., Nanjo, H., Shinozaki, T., and Furui, S., "Benchmark test for speech recognition using the Corpus of Spontaneous Speech," *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo*, 135-138, 2003.