Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Spontaneous Speech Recognition Taking Account of Characteristics of Speaker-Dependent Occurrence of Filled-Pauses

Yumi Shima (1), Mariko Koga(1), Masaru Yamashita(1), Katsuya Yamauchi (2) and Shoichi Matsunaga (2)

(1) Graduate School of Science and Technology, Nagasaki University, Nagasaki, Japan
(2) Department of Computer and Information Sciences, Nagasaki University, Nagasaki, Japan

PACS: 43.72.Ne

ABSTRACT

One of the characteristics of spontaneous speech is the occurrence of many types of filled-pauses that usually hamper the speech recognition accuracy considerably. In this study, we first investigated the occurrence frequency of filled-pauses in spontaneous speech by using a large corpus. The investigation results revealed that the cumulative occurrence frequency of filled-pauses reaches 0.8 with only four specific filled-pauses on an average; these frequent filled-pauses were differed among speakers. On the basis of these results, we propose a speech recognition procedure that employs a combination of two recognition processes; the first process involves the use of a common lexicon and the second involves the use of an individual lexicon. The filled-pause entries in the individual lexicon were estimated on the basis of their occurrence frequencies; these occurrence frequencies were observed from the preparatory results of the first recognition process. The proposed procedure demonstrated a statistically significant improvement in the word accuracy (1.1% word-error reduction) and indicated that the filled-pauses that are rarely used by speakers hinder improvements in word accuracy. We also showed that the use of an individual lexicon that was configured from a combination of the N-best results and word confidence score limitations induced a significant improvement in the word accuracy (1.3% word-error reduction). Furthermore, we examined the applicability of certain methods for reducing the processing time by implementing multiple candidates and confidence score limitations. Our procedure facilitated a significant improvement in the total processing amount (41% reduction in the number of recognition segments of the first recognition process) by using the N-best results and the word confidence score limitations.

INTRODUCTION

The aspects of spontaneous speech that do not appear in read speech decrease the recognition accuracy of spontaneous speech. One of the most prominent aspects of spontaneous speech is the presence of disfluencies such as filled-pauses, unwanted pauses, word fragments, self-corrections, and repeated words. In particular, the misidentification of filled-pauses affects the recognition results of the words concatenated to the filled-pauses and causes a decrease in the recognition accuracy.

In addition, there are individual variations in spontaneous speech with regard to the speakers' choice of filled-pause types and the occurrence frequency of the filled-pauses. Watanabe et al [1] revealed that the speakers' choice of filled-pause types is influenced by the speaking style (academic or casual) and the speakers' gender and age and that all relevant factors are in turn influenced by these filled-pause types.

Therefore, in this study, we focused on the individual variations in the type and occurrence frequency of the filled pauses. The speakers' choice of the types of filled-pauses and their occurrence frequencies were investigated using a large corpus of spontaneous speech. Based on the obtained results, we propose a two-step recognition procedure for spontaneous speech using an individual word lexicon that is augmented by the unsupervised estimation of filled-pauses uttered by each speaker. The aim of our proposed procedure is to achieve a more accurate recognition of spontaneous speech through the improvements in the recognition performance of not only the filled-pauses but also the words concatenated to the filled-pauses.

CHARACTERISTICS OF SPEAKER DEPENDENT OCCURRENCE OF FILLED-PAUSES

Thirty lectures uttered by 15 male and 15 female speakers were transcribed; these lectures were randomly selected from the Corpus of Spontaneous Japanese (CSJ) [2] except the evaluation set. The filled-pause occurrence frequency per word varied among the speakers (from 0.00 to 0.14), while the average occurrence frequency of filled-pauses for the 30 speakers was 0.06.

Table 1 shows the occurrence frequency of each filled-pause for two speakers whose filled-pause occurrence frequencies per word were equal. This table indicates that the speakers'

Table 1. Occurrence frequency of each filled-pause for two speakers

Speaker ID: A03F0003		Speaker ID: A01F0949	
Filled-pause type	Occurrence frequency	Filled-pause type	Occurrence frequency
a n o o	0.30	e e	0.49
a n o	0.23	e	0.39
m a	0.10	a	0.04
e e	0.08	a n o o	0.02
e e t o	0.07	a n o	0.01
m a a	0.03	ng	0.01
e	0.03	e e t o o	0.01

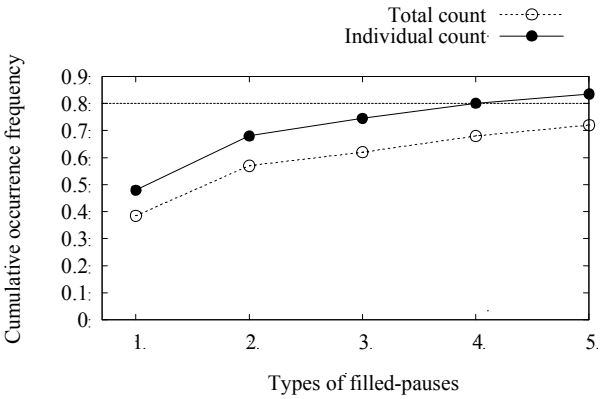

Figure 1. Cumulative occurrence frequency of filled-pauses for total and individual counting conditions

choice of filled-pause types and their occurrence frequencies vary among speakers.

Figure 1 shows the cumulative occurrence frequency of filled-pauses for two counting conditions: The "total count" denotes the frequency determined from the transcriptions of all the 30 speakers, while the "individual count" denotes the average value of each speakers' frequency determined individually.

Although the "total count" frequency does not reach 0.8 with five types of filled-pauses, the "individual count" frequency reaches 0.8 with only four types of filled-pauses.

These investigations reveal that the individual selection of each filled-pause entry in the word lexicon might be an effective way to improve word recognition performance.

A TWO-STEP SPONTANEOUS SPEECH RECOGNITION PROCEDURE

In this study, we propose a spontaneous speech recognition procedure constructed from two recognition processes; the first recognition process involves the use of a common lexicon and the second recognition process involves the use of an individual lexicon. The filled-pause entries in the individual lexicon are selected on the basis of the speakers' choice of filled-pause types and their

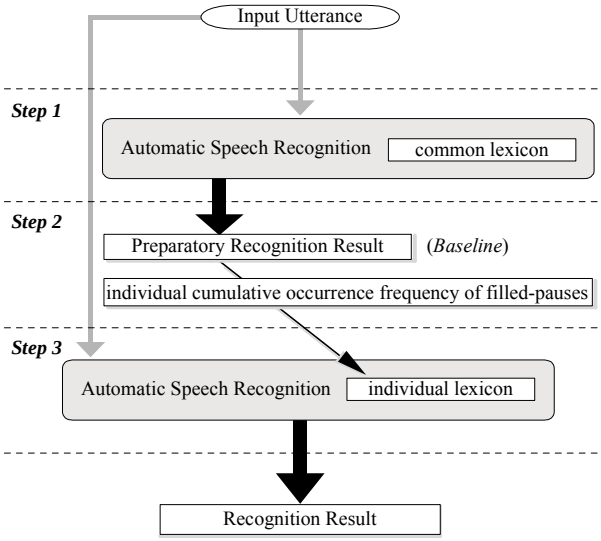

Figure 2. Procedure of the proposing concept which comprises two automatic speech recognition processes and individual lexicon adaptation

occurrence frequencies which are observed from the preparatory results derived from the first recognition process.

This concept is aimed to achieve a more accurate recognition of spontaneous speech through the improvements in the recognition performance of not only filled-pauses but also words concatenated to the filled-pauses. Our procedure only conducts filled-pause entry selection in the pronunciation lexicon without changing the occurrence probabilities of filled-pauses in the language model.

Our proposing procedure requires around twice the processing amount compared to that needed under the ordinal procedure which is constructed from a single recognition process. Thus, we also examine the ways for reducing processing amount using the word graph of N-best first recognition results and their word confidence scores.

Recognition procedure

The outline of our proposed procedure is illustrated in Fig. 2. At step 1, the preparatory recognition results are obtained from the first speech recognition process using the common lexicon. The 1-best preparatory recognition result, derived by using common lexicon, is called the baseline.

The individual lexicon is adapted on the basis of the individual cumulative occurrence frequency of filled-pauses which is counted from the preparatory recognition results for each speaker in step 2. The filled-pauses are sorted in order of their occurrence frequency, then the filled-pauses whose cumulative occurrence frequency fulfill the cut-off threshold are selected for the individual lexicon. Since the other filled-pauses —including the filled-pauses that have not appeared— are deleted from the individual lexicon, the individual lexicon with 100% cut-off threshold is different from the common lexicon. The word graph of the N-best first recognition results and their word confidence scores are also calculated to be used in the lexicon adaptation.

Finally, the recognition result is obtained from the second speech recognition process using the individual lexicon (step 3). Filled-pause occurrence frequency counting and steps regarding filled-pause entry limitations are carried out based on the phonetic transcription of CSJ [3].

RECOGNITION EXPERIMENTS

Experimental condition

In this section, the general experimental conditions for the following experiments are described.

Ten male speakers' academic presentation speeches containing approximately 27k words from the CSJ were used to form the evaluation set. The average and the standard deviation (s.d.) of the independent types of filled-pauses were 14.2 and 6.2. The speech sound data of the evaluation set had been divided into segments (physical utterances) whereas pauses that were longer than 200 ms were located [2].

The training data used for acoustic modeling is a part of CSJ, which comprises 819 academic presentation speeches (APS) with a combined length of 274.4 hours. The conventional 3000 states 16-mixture hidden Markov models (HMMs) were constructed with the set of 27 Japanese phonemes, while the feature parameters were 12 mfcc, Δmfcc, power, Δpower. The language model was trained from the transcriptions of APS and simulated public speeches (SPS) in CSJ. The source counts are 2668, while the total words of the training data set are 7410 k. The word entry size of the common lexicon is 40 k. The number of independent types of phonetic transcription of filled-pause entries in the individual lexicon is 284.

The recognition process is performed using the Julius version 4.1.2 [4]. The recognition results are estimated with the word accuracy (Acc). The correctness of the filled-pauses is usually denoted by the parse according to the filled-pause tag in the transcription of the CSJ; however, the correctness of filled-pauses is gauged in this study by checking them alongside the phonetic transcriptions and recognition results used in this study.

Individual lexicon adaptation (Experiment I)

In order to examine the validity of the proposed procedure using the individual lexicon adaptation, one supervised and one unsupervised recognition experiment are conducted. Here, the supervised condition uses the individual cumulative occurrence frequency of filled-pauses counted from the transcription of the input utterances. The unsupervised condition uses the individual cumulative occurrence frequency of filled-pauses counted from the 1-best preparatory recognition results derived by using common lexicon (baseline).

Figure 3 shows the Acc values under the supervised and unsupervised conditions as compared with the baseline. The Acc value of the baseline condition was 69.99%. Both the supervised and unsupervised conditions effect a significant increase over the baseline word accuracy.

With respect to the supervised condition, the Acc achieved 70.42% at a 91% filled-pause cut-off threshold. The Acc difference (+0.43%) between the baseline and the 91% cut-off threshold condition is statistically significant ($p < 0.05$) according to the matched-pair test. The average of the independent types of phonetic transcription of filled-pause entries in the individual lexicon was 6.1 (s.d. 3.4). With respect to the unsupervised condition, the Acc was 70.31% under the 91% cut-off threshold. This Acc difference (+0.32%) is also statistically significant ($p < 0.05$). This improvement means 1.1% word-error reduction. The average type of phonetic transcription of filled-pause entries in the individual lexicon was 9.4 (s.d. 2.8). Efficient filled-pause entry selection has led to a significant improvement in word accuracy values.

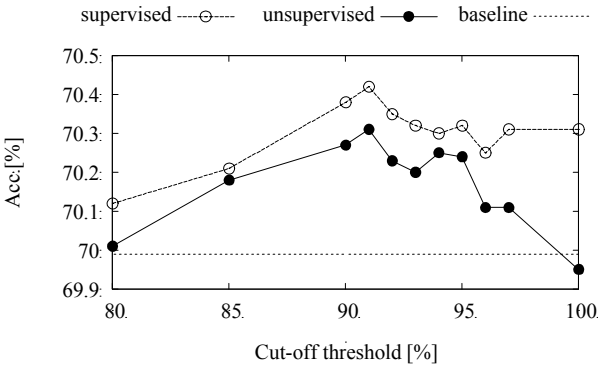

Figure 3.Word accuracies for experiment I

According to the comparative observations of the recognition result and the transcription, the insertion errors of filled-pauses and replacement error of words seem to decrease. The cut-off threshold conditions of around 91% show a greater increase than the 100% cut-off threshold. This tendency suggest that the filled-pauses which have low occurrence frequency cause relatively less improvement in terms of Acc.

Employing N-best results and word confidence score (Experiment II)

We conducted an unsupervised recognition experiment, using multi-candidates results in the individual lexicon adaptation. Although the 1-best results were used to configure the individual lexicon in experiment I, generally, not all of the correct words (filled-pauses) were included in the 1-best recognition results. In this experiment, the individual cumulative occurrence frequency of filled-pauses is counted from the filled-pause nodes in the word graph that is generated from N-best recognition results. Nine conditions on the number of candidates (N =1, 5, 10, 15, 20, 25, 30, 35, and 40) were set.

Word confidence scores were employed to select highly confident filled-pauses for the configuration because a greater number of candidates usually implied a greater number of misidentified filled-pauses. Word confidence was then computed by using estimated posterior probability while decoding [5]. Five conditions on the confidence threshold (C = 0.0, 0.1, 0.2, 0.3, 0.4) were set. The node filled-pauses that had higher word confidence scores than the threshold were counted; these represent the individual cumulative occurrence frequency. The filled-pause cut-off threshold was set to 95%.

Figure 4 shows the Acc for each word confidence threshold condition. The increase of Acc for the conditions C = 0.1, 0.2, 0.3 are greater than that for the C = 0.0 condition. The confidence score limitation leads significant improvement in the Acc. With respect to the C = 0.2, 0.3 conditions, the Acc with N-best recognition result is higher than that arrived at by using 1-best recognition results. The combination of N-best results and word confidence scores was found to be effective in improving the Acc values.

The estimate accuracy of the individual lexicon by resorting to N-best result and word confidence score limitations was evaluated by the F-measure that was calculated from the recall and precision rates of the filled-pause entries in the individual lexicon for each condition on the basis of the supervised condition at a 91% filled-pause cut-off threshold. The F-measure at the 20-best preparatory results and 0.2 confidence threshold was 0.76, which was 0.01 higher than the unsupervised condition at a 91% filled-pause cutoff threshold.

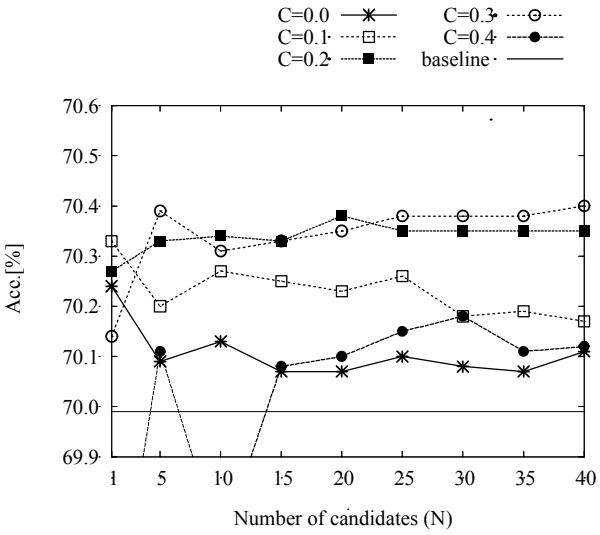

Figure 4. Word accuracies for experiment II

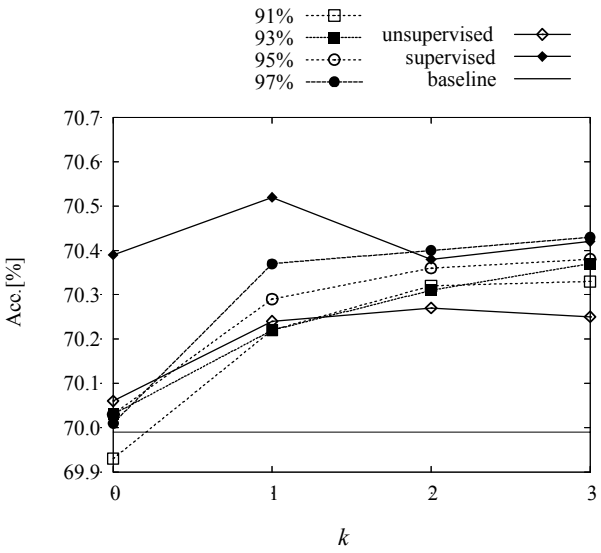

Figure 5. Word accuracies for experiment III

Reducing processing time (Experiment III)

The recognition procedures in the experiments discussed above require twice (or more) the processing time compared with the baseline; this is because the proposed procedure includes two recognition processes. Thus, in this experiment, the segments the first recognition process are reduced.

The number of segments (s) is denoted by the equation $s = round(\mu + k\sigma)$, where the k = 0, 1, 2, 3. The μ and σ are the average and s.d. of segments which fulfill the 91% cumulative occurrence frequency of filled-pauses for each speaker. The average and s.d. measured from the transcription of 50 speakers randomly selected from the CSJ were 31.7 and 40.0. Thereafter, four conditions with respect to the segments (s = 32, 72, 112, and 152) were established.

The multiple candidates condition was set, wherein the 20-best preparatory results and 0.2 confidence threshold were used in the individual lexicon adaptation. The filled-pause cut-off threshold was set to four conditions (91, 93, 95, and 97%). In contrast to the multi-candidates conditions, the supervised condition and unsupervised conditions, wherein the 1-best preparatory recognition results were used, were set separately. The filled-pause cut-off threshold was set at 91% for these conditions.

Figure 5 shows the Acc for each condition compared with the baseline value. With respect to the supervised 1-best condition, the Acc differences from the baseline were statistically significant ($p < 0.05$), even for the k = 0 condition (s = 32); this was true for 26% of the average number of segments in the input. With regard to the unsupervised 1-best condition, the Acc differences between the baseline value and that observed for the k = 2, 3 conditions (s = 112, 152) were statistically significant ($p < 0.05$). At these conditions, however, we exceed the average number of segments for 10 speakers in the evaluation set (121.9). The unsupervised condition requires a larger number of segments that use the first recognition than the supervised condition in order to estimate the correct cumulative occurrence frequency of filled-pauses.

On the subject of the multi-candidates condition, the Acc value touched 70.37% at k=1 with a 97% cut-off threshold. The Acc difference (+0.38%) from the baseline is statistically significant ($p < 0.05$). This improvement means 1.3% word-error reduction. The number of segments used in the first recognition process at k = 1 condition was 72, which was 59% of the all segments of input. In comparison with the unsupervised 1-best condition, the multiple candidates condition shows a significant improvement in Acc, even for the lesser segments used in the first recognition process. It is therefore suggested that a more correct estimation of the cumulative occurrence frequency of filled-pauses leads to improvements in the word accuracy.

The occurrence frequency of filled-pauses is estimated more correctly if low confidence filled-pauses are deleted on the basis of confidence scores. The filled-pauses that rarely occur are deleted by using the filled-pause cut-off threshold. Setting up of these two limitations — confidence scores and filled-pause cut-off thresholds—resulted in the creation of a higher cut-off threshold condition than those in the former experiments which in turn helped achieve a higher Acc.

SUMMARY

In this study, we investigate the occurrence rate of filled-pauses in spontaneous speech by using the Corpus of Spontaneous Japanese. Our investigation reveals that the speakers' choice of filled-pauses and their occurrence frequencies vary among speakers. The investigation also suggests that the individual limitation of filled-pause entries in the individual lexicon could be effective in improving the recognition accuracy.

On the basis of these characteristics of the filled-pauses, we propose a two-step spontaneous speech recognition procedure that consists of two recognition processes; the first recognition process involves the use of a common lexicon and the second recognition process involves the use of an individual lexicon. The aim of this procedure is achieve a more accurate recognition of spontaneous speech. The filled-pause entries in the individual lexicon are selected on the basis of their occurrence frequency, which is estimated by using the first recognition results.

The validity of our proposed procedure was demonstrated in experiment I. Both the supervised and unsupervised conditions showed a statistically significant improvement in the word accuracy. The recognition results also indicated that the filled-pauses that are rarely used by speakers may nevertheless hinder improvements in the word accuracy.

Experiment II was an unsupervised recognition experiment that involved the use of multiple candidate results and word confidence scores in the individual lexicon adaptation. The experimental result showed that the use of an individual lexicon that was configured from a combination of the N-best results and word confidence scores provided more accurate recognition results. It was observed that the imposition of the filled-pause entry limitation with regard to establishing the correct occurrence frequency of filled-pauses induced significant statistical improvements in recognition accuracy.

The required processing amount of the proposed procedure is twice (or more than twice) the processing amount of the baseline procedure; this is because the former involves two recognition processes. In experiment III, we showed the possibility of reducing processing amount by using the N-best results and word confidence score limitations.

REFERENCES

1 Watanabe, M., et al. “Factors Affecting Speakers’ Choice of Fillers in Japanese Presentations”, *IINTERSPEECH 2006*, 1498-Tue3A3O.3 (2006).
2 Maekawa, K., “Corpus of Spontaneous Japanese: Its design and evaluation”, *SSPR-2003*, MMO2 (2003).
3 Koiso, H., et al. “Transcription criteria for the Corpus of Spontaneous Japanese”, *Japanese Linguistics*, Nat’l Inst. for Japan. Lang., 9:43-58 (2001).
4 Lee, A., et al. “Julius – an open source real-time large vocabulary recognition engine”, *EUROSPEECH*, 1691-1694 (2001).
5 Lee, A, et al. “Real-time word confidence scoring using local posterior probabilities on tree trellis search”, *ICASSP ’04*, Vol. I, 793-796 (2004).