Improvement of Speech Summarization Using Prosodic Information

Akira Inoue, Takayoshi Mikami and Yoichi Yamashita

Department of Computer Science, Ritsumeikan University
{pigman,mikami,yama}@slp.cs.ritsumei.ac.jp

Abstract

Speech summarization is a technique of extracting important sentences from spoken documents. It provides us useful information to looking for the spoken documents that we want. Spoken documents contain non-linguistic information, which is mainly expressed by prosody, while written text conveys only linguistic information. This paper describes a summarization method which uses prosodic information as well as linguistic information. The linguistic information is derived from text which is transcribed by a continuous speech recognition system. In this paper, the speech summarization is defined as extraction of important sentences from transcribed text. Importance of the sentence is predicted by the prosodic parameters and the linguistic information which are combined by multiple regression analysis. Proposed methods are evaluated both on the correlation between the predicted scores of sentence importance and the preference scores by subjects and on the accuracy of extraction of important sentences. Prosodic information improved the quality of speech summary, and it is more effective when the speech is transcribed by automatic speech recognition because speech recognition errors damage linguistic information.

1. Introduction

Stochastic methods based on corpora have improved the performance of continuous speech recognition (CSR). A CSR technique enables automatic summarization of spoken documents, such as news, lecture, public speech, and so on [2]. Since speech media is not appropriate to quick scanning, the automatic summarization for speech is more useful than for written text. Non-linguistic information is mainly expressed by prosody in speech unlike written text. Possibility of improving quality of speech summarization has been reported [1].

Many researchers have studied text summarization. A speech summarization scheme can be realized by the simple consecutive combination of two conventional techniques of the continuous speech recognition and the text summarization, shown as Fig.1 (a). This approach uses only a linguistic aspect of speech data and ignores non-linguistic information like prosody. The prosody plays important roles in speech communication to express non-linguistic information such as intension, topic change, emphasizing words or phrases, and so on. Introducing prosodic information into the speech summarization process, shown as Fig.1 (b), is expected to improve the quality of summary. This paper describes a method of the speech summarization and effects using several prosodic parameters as well as linguistic information when speech is transcribed by a continuous speech recognition system.

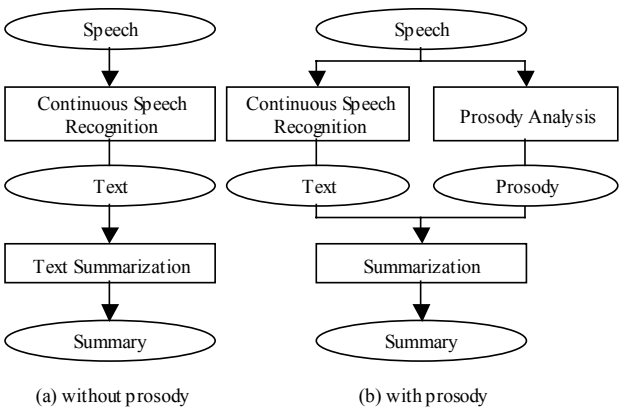

Figure 1: *Process of speech summarization.*

2. Method

2.1. Summarization

To produce a refined summary, in general, we need to understand contents of written text or a spoken message, to extract important parts, then to generate consistent sentences. The automatic understanding of meanings of the contents, however, is not easy task for computer. Many studies of the text summarization try to just extract important sentences or phrases from written text without deep understanding of the contents [3][6][7]. In this paper, the speech summarization is also defined as extraction of important sentences from transcribed text. Lecture speech is transcribed by hand and boundaries of the sentence are also manually defined. In this framework, the problem of speech summarization becomes automatic scoring of sentence importance for the transcribed text.

2.2. Prosodic parameters

Prosodic parameters of phoneme duration, power, and F0 are extracted for each sentence to predict importance of the sentences.

2.2.1. F0 parameters

We use four F0 parameter parameters as follows.

$$F_{\min} = \min(f_1, f_2, ..., f_N)$$

$$F_{\max} = \max(f_1, f_2, ..., f_N)$$

$$F_{range} = F_{\max} - F_{\min}$$

$$F_{mean} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

N is a number of the frame in a sentence, f_i is an F0 of i-th frame in the sentence. F0 is computed by ESPS.

2.2.2. Phoneme duration

Observed phoneme duration D_i is normalized by the following equation (2).

$$d_i = \frac{D_i - \overline{D}(ph_i)}{\sigma_D(ph_i)} \quad \cdots(2)$$

In this equation, D_i is the duration of *i*-th phoneme ph_i in a sentence, $\overline{D}(ph)$ and $\sigma_D(ph)$ are a mean and a standard deviation of the duration of the phoneme ph, respectively. $\overline{D}(ph)$ and $\sigma_D(ph)$ were independently calculated for each data.

We use four parameters of phoneme duration as follows.

$$DUR_{mean} = \frac{1}{n}\sum_{i=1}^{n} d_i$$
$$DUR_{\min} = \min(d_1, d_2, ..., d_n)$$
$$DUR_{\max} = \max(d_1, d_2, ..., d_n)$$
$$DUR_{range} = DUR_{\max} - DUR_{\min}$$

n indicates the number of the phoneme in a sentence.

2.2.3. Sentence length

The duration of a sentence, *LEN* , is used. *LEN* includes pause time in the sentence.

2.2.4. Power

Observed phoneme power P_i is normalized by equation (3).

$$p_i = \frac{P_i - \overline{P}(ph_i)}{\sigma_P(ph_i)} \quad \cdots(3)$$

In this equation, P_i is the power of *i*-th phoneme ph_i in a sentence, $\overline{P}(ph)$ and $\sigma_P(ph)$ are a mean and a standard deviation of the power of the phoneme *ph* , respectively. $\overline{P}(ph)$ and $\sigma_P(ph)$ were independently calculated for each data.

We use four phoneme power parameters as follows.

$$POW_{mean} = \frac{1}{n}\sum_{i=1}^{n} p_i$$
$$POW_{\min} = \min(p_1, p_2, ..., p_n)$$
$$POW_{\max} = \max(p_1, p_2, ..., p_n)$$
$$POW_{range} = POW_{\max} - POW_{\min}$$

2.3. Linguistic information

In recent research, many methods have been proposed to extract important sentences. Since identification of linguistic information which is useful to summarization is out of scope of this paper, we introduce the linguistic information which is employed in conventional text summarization.

- *Frequency of word occurrence*

  Research in natural language study shows that words whose frequency of occurrence is intermediate are important. A sentence in which important words often appear has a high probability that it is an important sentence. Therefore, the frequency of the word occurrence is useful for summarization [4].

- *Cue word*

  In important sentences, cue keywords like “significant”, “impossible” or “hardly” often appear [5].

- *Title*

  The words appeared in a title are important.

- *Location*

  Important sentences sometimes appear after a title, a head or an end of text or paragraph. This indicates that the sentence importance depends on the location in text.

This study uses a summarization engine for Japanese written text, Posum [8]. It reads input text and generates the importance score of each sentence. We use the Posum score as linguistic information parameter for speech summarization and it is denoted by *LING* in this paper.

3. Evaluation

3.1. Speech Data

Recorded video data of five lecture talks, denoted as data-1, -2, -3, -4 and -5, from TV program is employed for experiments. The details of data are shown as Table 1. Sentence segmentation is carried out by hand.

3.2. Continuous Speech Recognition

Each speech data is transcribed by a Japanese continuous speech recognition (CSR) system, Julius 3.3p3 [9], to get linguistic information. The average word recognition accuracy of five speech data is 64.6%. The speech data is also transcribed by hand to investigate effects of speech recognition errors.

3.3. Sentence Importance

Summarization experiments were carried out to obtain the importance score of sentences. The number of the subject is 14,18,13,14 and 15 for data-1, -2, -3, -4 and -5, respectively. The subjects watched the recorded video of the lecture to understand the contents. Then, they were asked to select both about 10 important sentences and about 10 unimportant

Table 1: *Speech Data*.

data ID	data-1	data-2	data-3	data-4	data-5
contents	nuclear flash criticality accident	vitality of aged persons	regeneration of beach	decrepit nuclear power plant	reconciliation of the lawsuit of Creutzfeldt-Jakob disease
speaker	male M1	female F1	male M2	male M3	female F2
number of sentence	65	68	71	76	71

sentences from all sentences in the lecture using its transcription, during listening the speech without image information.

The sentence important of the *i*-th sentence, $SI(i)$, is defined as follows.

$$SI(i) = R(i)_{imp} - R(i)_{unimp}$$

In this equation, $R(i)_{imp}$ and $R(i)_{unimp}$ is ratio of the subjects who selected the *i*-th sentence as an important and an unimportant sentence, respectively. The importance of the first 30 sentences for data-2 is shown in Fig. 2. In many studies of speech summarization, the sentence importance is measured by a binary decision which selects a set of important sentences. In this paper, the importance of each sentence is continuously scaled between –1 and 1 by averaging the judgments by several subjects.

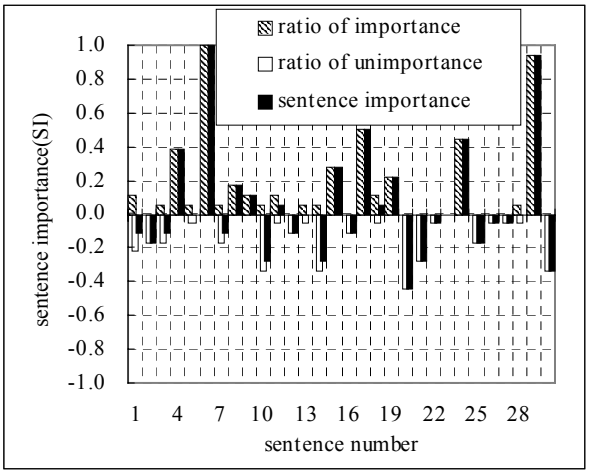

Figure 2: *Examples of sentence.*

3.4. Correlation between Sentence Importance and Each Parameter

Figure 3 shows means of the correlation coefficient between sentence importance $SI(i)$ and each parameter. F_{range}, DUR_{range}, POW_{mean} in this figure shows the highest correlation in the F0, the duration and the power parameters, respectively.

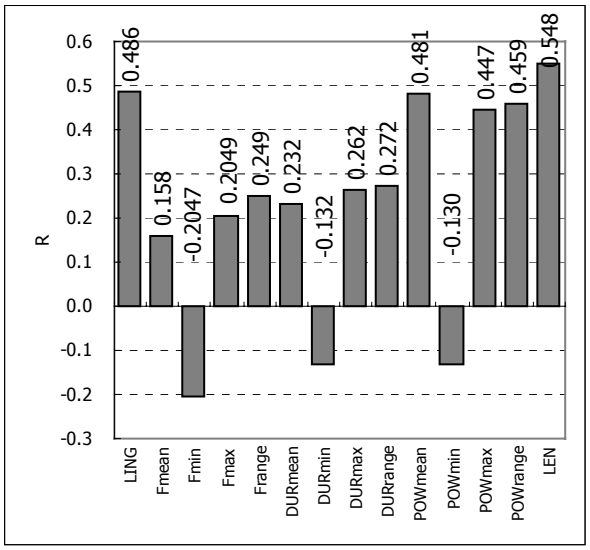

Figure 3: *Correlation coefficient with sentence importance* $SI(i)$.

3.5. Multiple regression analysis

The sentence importance is predicted by a multiple regression model. The multiple regression is formulated by

$$SI(i) = a_0 \times LING(i) + \sum_{j=1}^{M} [a_j \times B(i)_j], \quad \cdots(4)$$

where $LING(i)$ is the sentence importance score from linguistic information, and $B(i)_j$ is *j*-th prosodic parameter in *i*-th sentence. M is the number of prosodic parameter to be combined. We combine linguistic information, $LING$, the sentence length, LEN, and some prosodic parameters to predict sentence importance using the multiple regression model mentioned in 3.5. We tried following three combinations of the parameters.

- C0
  $LING$
- C1
  $LING$, LEN, F_{range}, DUR_{range}, POW_{mean}
- C2
  $LING$, LEN, $F_{\min}$, F_{range}, $DUR_{\max}$, DUR_{range}, POW_{mean}, POW_{range}

C0 is a parameter set which uses only linguistic information, and is gives a baseline performance. C1 and C2 include parameters which has high correlation with the sentence importance SI in F0, duration and power. Combination C1 uses parameters which has the highest correlation in each parameter category, F0, duration, and power, shown in 3.4, and C2 also uses parameters with the two highest correlation.

To evaluate the prediction for unseen speech data by the multiple regression model, the model should apply for open data which is not used for the model training. There are five spoken lectures as mentioned 3.1. The multiple regression model is trained with four lectures, and the other one is evaluated. This open evaluation is repeated five times replacing the evaluation data. In the closed evaluation, the model is trained with a lecture and is applied to the same lecture. The closed evaluation is also repeated five times.

3.5.1. Evaluation by multiple regression coefficient

Figure 4 shows multiple regression coefficients for each combination pattern of the parameter. In this figure, "trans-" and "CSR-" indicate the result for transcribed text by hand and CSR, respectively, and "-closed" and "-open" indicate open and closed evaluation, respectively.

In figure 4, C1 and C2, which use prosodic parameters, take higher correlation coefficients with the sentence importance than the baseline C0. However, in the case of the open evaluation, correlation coefficients take lower values by increasing parameters from C1 to C2. It seems that four data is not enough to train the multiple regression model. C2 may get higher prediction accuracy if it can learn more data set.

In the parameter combination C0, the multiple correlation coefficients are decreased for the cases using the text transcribed by CSR due to word recognition errors. However, in the prediction by C1 and C2 which use prosodic parameters, the multiple correlation coefficients are not degraded for the

CSR text. The introduction of prosodic parameters are more effective for the text transcribed by CSR.

3.5.2. Evaluation by Identification Rate of Important Sentences

The quality of the summary is evaluated by another measure, identification rate of important sentences, IR , which is defined by following equations.

$$IR = (IR_5 + IR_{10} + IR_{15} + IR_{20})/4$$

$$IR_r = (C(r)_{imp} - C(r)_{unimp})/r$$

In this equation, $C(r)_{imp}$ is the number of sentences which match with one of the r most important sentences, when r sentences are automatically extracted, and $C(r)_{unimp}$ is the number of matched unimportant sentences in the same manner. The IR score indicates an expectation rate that an important sentence is correctly detected at r =5,10,15,20. IR will be 1 if an extracted summary is completely the same as a summary by hand. On the other hand, IR will be 0 if a summary is randomly generated.

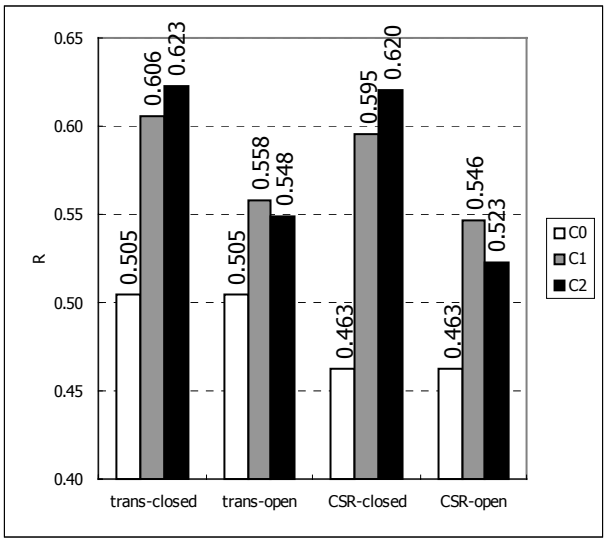

Figure 4: *Multiple correlation coefficient with sentence importance $SI(i)$.*

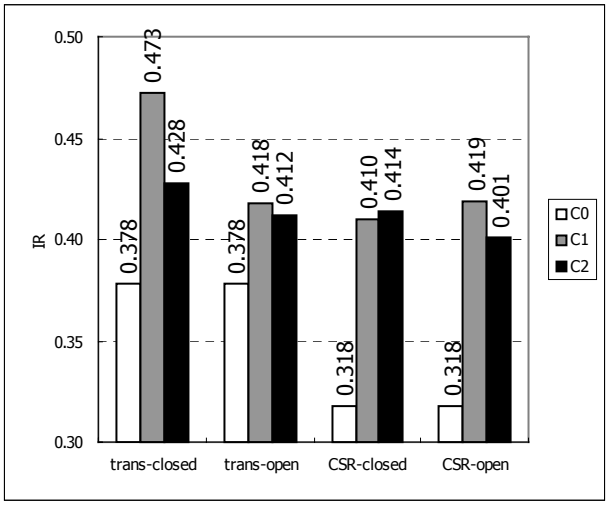

Figure 5: *Mean of Identification rate of important sentences.*

Figure 5 shows evaluation of the automatic summarization by the identification rate IR . In this figure, C1 and C2 take higher IR values than the baseline C0. This indicates that summarization quality is improved by using prosodic information. For the evaluation by the IR score, the introduction of prosodic parameters is more effective for the text by CSR.

4. Conclusions

This paper describes a method of the speech summarization and effects using several prosodic parameters as well as linguistic information when speech is transcribed by continuous speech recognition. Quality of speech summarization is improved by introduction of prosodic parameters, especially for the text transcribed by continuous speech recognition because speech recognition errors damage linguistic information. Prosodic information potentially improves the performance of other speech recognition applications because prosody may include not only useful information to summarization but also various non-linguistic information.

In order to obtain further improvement, large speech data sets are necessary to train a multiple regression model, since speakers prosodically emphasize sentences in different manners, it is necessary to classify types of the speakers and to model speakers' characteristics. To find other prosodic parameters which are more effective for summarization will be also a future work.

5. Acknowledgement

The present research was partly supported by Grant-in-Aid for scientific Research on priority Area (B) "Prosody and Speech Processing" from the Ministry of Education, Culture, Sports, Science and Technology.

6. References

[1] A.Inoue, T.Mikami and Y.Yamashita, 2003. Prediction of sentence importance for speech summarization using prosodic parameters. *Eurospeech 2003*, vol 1, 1193-1196.

[2] C.Hori and S.Furui, 2001. Advances in automatic speech summarization. *Eurospeech 2001*, vol 3, 1771-1774.

[3] I.Mani and M.Maybury, 1999. Advances in Automatic Text Summarization, *The MIT Press*.

[4] Luhn, H.P., 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2, 2, 159-165.

[5] Edmundson, H.P., 1969. New Methods in Automatic Extracting. *Jurnal of the Association for Computing Machinery*, 16, 2, 264-285.

[6] M.Okumura, T.Hisamitsu and S.Masuyama, 2002. Special edition: Text automatic summary. *IPSJ Magazine* 43, 12, 1285-1316.

[7] S.Satoh and M.Okumura, 1999. How does a computer summarize a text. *IPSJ Magazine* 40, 2, 157-161.

[8] http://www.tufs.ac.jp/ts/personal/motizuki/software/posumcl

[9] http://julius.sourceforge.jp/