Prediction of sentence importance for speech summarization using prosodic parameters

Akira Inoue, Takayoshi Mikami, Yoichi Yamashita

Department of Computer Science, Ritsumeikan University

{pigman,mikami,yama}@slp.cs.ritsumei.ac.jp

Abstract

Recent improvements in computer systems are increasing the amount of accessible speech data. Since speech media is not appropriate for quick scanning, the development of automatic summarization of lecture or meeting speech is expected. Spoken messages contain non-linguistic information, which is mainly expressed by prosody, while written text conveys only linguistic information. There are possibilities that the prosodic information can improve the quality of speech summarization. This paper describes a technique of using prosodic parameters as well as linguistic information to identify important sentences for speech summarization. Several prosodic parameters about F0, power and duration are extracted for each sentence in lecture speech. Importance of the sentence is predicted by the prosodic parameters and the linguistic information. We also tried to combine the prosodic parameters and the linguistic information 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. By combination of the prosodic parameters improves the quality of speech summarization.

1. INTRODUCTION

Recent improvement of the computer system is increasing amount of accessible speech data, such as news, lecture, public speech, and so on. This situation makes it much difficult to find out data which we want. Since speech media is not appropriate for quick scanning, it is not easy to understand the outline of the whole speech in a brief moment. One of techniques which overcome this disadvantage is speech summarization which extracts important parts from speech contents [1].

Many studies of the summarization have been tried for text. A speech summarization scheme can be realized by 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 the relation between several prosodic parameters and importance degree of sentences in lecture speech, and effectiveness to predict sentence importance by multiple regression analysis with prosodic parameters and linguistic score in the speech summarization.

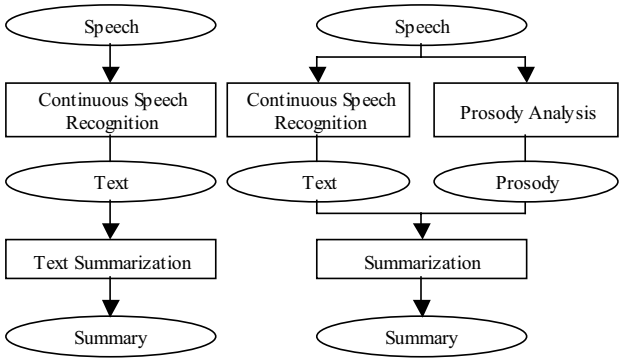

Fig. 1. Process of speech summarization

2. METHOD

2.1. Summarization

To produce a refined summary, in general, we need to understand contents of written text or spoken message, to extract its essential 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 [2][5][6]. 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 three F0 parameter parameters as follows.

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

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

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

N is a number of 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 the sentence, $\overline{D}(ph)$ is a duration average of the phoneme ph, $\sigma_D(ph)$ is a standard deviation of the phoneme ph and it was independently calculated for data-1,-2 and –3. D_i is determined by forced alignment of HTK.

We use four parameters of phoneme duration as follows.

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

n indicates the number of the phoneme in the 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 the sentence, $\overline{P}(ph)$ is a power average of the phoneme ph, $\sigma_P(ph)$ is a standard deviation of the phoneme ph and it was independently calculated for data-1,-2 and –3.

We use four phoneme power parameters as follows.

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

2.3. Linguistic Information.

In recent research, extraction of important sentence had been trying by using many methods as next subsection. 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 occurrence word*

  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. From these things, the frequency of the word occurrence is useful for summarization [3].

- *Cue word*

  In important sentences, cue keywords like “significant”, “impossible” or “hardly” often appear [4].

- *Title*

  The words appeared in a title are importance.

- *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 [7]. It reads input text and generates the importance score of each sentence. We use the Posum scores as linguistic information parameter for speech summarization and it is referred by $LING$ in this paper.

2.4. 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, $B(i)_j$ is *j*-th prosodic parameter in *i*-th sentence, M is the number of prosodic parameter to combine.

3. EVALUATION

3.1. Speech Data

Recorded video data of three lecture talks, referred as data-1, -2 and -3, from TV program is employed for experiments. The details of data are shown as Table 1. Sentences in the lecture talks are manually identified and speech is transcribed by hand.

Table 1. Speech Data

data ID	data-1	data-2	data-3
contents	vitality of aged persons	regeneration of beach	nuclear flash criticality accident
speaker	female F1	male M1	male M2
number of sentence	68	71	65

3.2. Sentence Importance

Summarization experiments were carried out to obtain the importance score of sentences. The number of the subject is 18, 13 and 14 for data-1, -2 and -3, 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 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-1 is shown in Fig. 2.

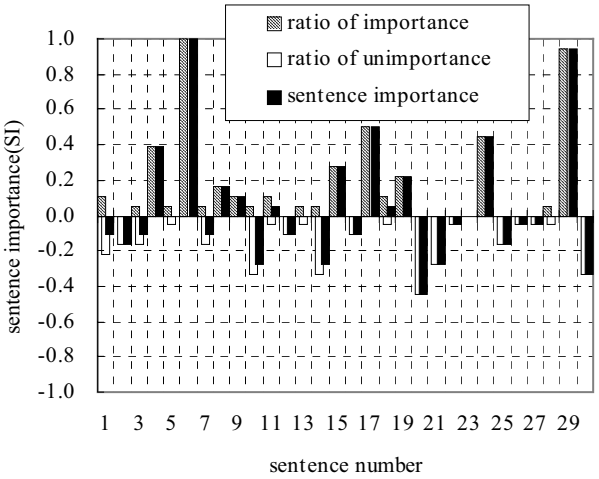

Fig. 2 Examples of sentence importance

3.3. Correlation between Sentence Importance and Each Parameter

Table 2 shows correlation coefficients between sentence importance $SI(i)$ and each parameter. $F_{\min}$, DUR_{range}, POW_{avg} in this table shows high correlation in the F0, the duration and the power parameter, respectively.

3.4. Combination of Prosodic Parameters

We combine linguistic information, *LING* , the sentence length, *LEN* , and some prosodic parameters to predict sentence importance using the multiple regression model mentioned in 2.4. We selected following three parameters from parameter sets for each type of the prosodic parameter based on the results in 3.3.

- $F_{\min}$
- DUR_{range}
- POW_{avg}

These three parameters gave the largest correlation with the sentence importance *SI* in F0, duration and power, respectively. Eight combination patterns shown in Table 3 are tried to combine parameters.

Table 4 shows multiple regression coefficients for each combination pattern. The combination pattern C3, C7, C8 using POW_{avg} shows higher multiple regression coefficients than other combination patterns.

Table 2. Correlation coefficients with sentence importance $SI(i)$

Parameters	data-1	data-2	data-3	average
LING	0.497	0.559	0.482	0.513
Fmin	0.325	0.278	0.251	0.285
Fmax	0.209	0.193	0.023	0.142
Frange	0.344	0.240	0.098	0.227
DURavg	0.198	0.270	0.006	0.158
DURmin	0.051	0.203	0.173	0.108
DURmax	0.215	0.244	0.367	0.275
DURrange	0.205	0.254	0.376	0.278
POWavg	0.459	0.584	0.390	0.477
POWmin	0.001	0.009	0.101	0.030
POWmax	0.259	0.600	0.321	0.393
POWrange	0.256	0.579	0.323	0.386
LEN	0.492	0.554	0.394	0.480

Table 3. Combination pattern

Combination pattern	C0	C1	C2	C3	C4	C5	C6	C7	C8
LING	O	O	O	O	O	O	O	O	O
LEN	×	O	O	O	O	×	×	×	O
Fmin	×	O	×	×	×	O	×	×	O
DURrange	×	×	O	×	×	×	O	×	O
POWavg	×	×	×	O	×	×	×	O	O

Table 4. Multiple regression coefficient with sentence importance

Combination pattern	data-1	data-2	data-3	average
C0	0.497	0.559	0.482	0.513
C1	0.552	0.586	0.490	0.543
C2	0.547	0.587	0.549	0.561
C3	0.573	0.636	0.502	0.570
C4	0.547	0.583	0.484	0.538
C5	0.509	0.562	0.490	0.520
C6	0.508	0.559	0.542	0.536
C7	0.570	0.605	0.498	0.558
C8	0.579	0.656	0.565	0.600

3.5. 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, 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 complete the same as a summary by hand. On the other hand, IR will be 0 if a summary is randomly generated.

Table 5 compares the parameter combination patterns in terms of the IR score. Combination patterns using prosodic parameters C1~C8 improve the extraction accuracy of important sentences. It is shown that using prosodic parameters for summarization gave higher quality of speech summarization than using only linguistic information. The effect of using prosody for summarization is clear. However, a combination pattern which dramatically improves quality of summarization was not observed.

3.6. Evaluation of model applying to open data

In 3.4, multiple regression model was evaluated for closed speech data. We try to predict the sentence importance of unseen speech data by applying the multiple regression model which was trained with other speech data. There is three spoken lectures as mentioned 3.1. The multiple regression model is trained with two lecture data, and it is evaluated for another. This open evaluation is repeated three times replacing the evaluation data.

Table 6 shows results of the open evaluation. In the table, figures in the "closed" column are the same value as the average in Table 5. Although the identification rates for the open evaluation are a little lower than for the closed evaluation, combination pattern C1~C8 improves extraction accuracy of important sentences than C0 which uses only linguistic information. Applying the multiple regression model to unseen speech data also improves the quality of speech summarization.

Table 5. Identification rate of important sentences

Combination pattern	data-1	data-2	data-3	average
C0	0.304	0.392	0.358	0.351
C1	0.346	0.542	0.358	0.415
C2	0.333	0.538	0.404	0.425
C3	0.363	0.450	0.383	0.399
C4	0.350	0.529	0.325	0.401
C5	0.358	0.467	0.421	0.415
C6	0.333	0.392	0.363	0.363
C7	0.350	0.554	0.275	0.393
C8	0.388	0.496	0.417	0.433

Table 6. Average of identification rate of important sentences

Combination pattern	closed	open
C0	0.351	0.351
C1	0.415	0.407
C2	0.425	0.358
C3	0.399	0.372
C4	0.401	0.397
C5	0.415	0.390
C6	0.363	0.344
C7	0.393	0.385
C8	0.433	0.361

4. CONCLUSIONS

This paper describes a technique which combine prosodic information not only linguistic information for summarization of speech lecture. Combination of prosodic parameters and linguistic information improves quality of summary than using only linguistic information. Combination of prosodic parameters and linguistic score by a multiple regression model is also effective to unseen speech data. 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. REFFERENCES

[1] C.Hori and S.Furui, "Advances in automatic speech summarization" in Proc. of Eurospeech 2001, vol. 3, pp.1771-1774 (2001)

[2] I.Mani and M.Maybury, "Advances in Automatic Text Summarization", The MIT Press (1999)

[3] Luhn, H.P., "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, Vol.2, No2, pp.159-165 (1958)

[4] Edmundson, H.P., "New Methods in Automatic Extracting", Jurnal of the Association for Computing Machinery,Vol 16,No.2, pp.264-285 (1969)

[5] M.Okumura, T.Hisamitsu and S.Masuyama, "Special edition: Text automatic summary", IPSJ Magazine Vol.43, No.12, pp.1285-1316 (2002).

[6] S.Satoh and M.Okumura, "How does a computer summarize a text", IPSJ Magazine Vol.40, No.2, pp.157-161 (1999)

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