# EXTRACTION OF IMPORTANT SENTENCES USING F0 INFORMATION FOR SPEECH SUMMARIZATION

*Yoichi Yamashita and Akira Inoue*

Dep. of Computer Science, Ritsumeikan University
1-1-1, Noji-Higashi, Kusatsu-shi, Shiga, 525-8577 Japan
{yama,pigman}@slp.cs.ritsumei.ac.jp

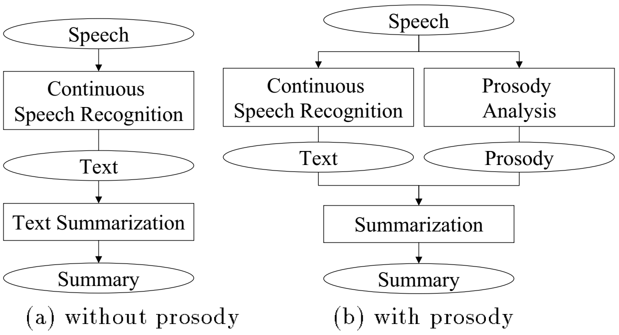
## ABSTRACT

This paper describes speech summarization using F0 information. The speech summarization in this work is realized by the extraction of important sentences from text data transcribed by hand. The important problem in this framework is automatic scoring of sentence importance based on prosodic information from speech wave as well as linguistic information from written text. Prosody conveys non-linguistic information such as speaker's intention and contributes to identify important parts of speech. The prosodic information is represented in terms of F0 parameters of Japanese *bunsetsu* unit, which is almost equivalent to a prosodic minor phrase. Six kinds of F0 parameters are compared in regard to correlation to the sentence importance and performance of extracting important sentences. Evaluation results show that introduction of F0 parameters is effective to the 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 the speech contents[1]. Many studies of the summarization have been tried for text[2].

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[3]. Introducing prosodic information into the speech summarization process, shown as Fig. 1 (b), is expected to improve the quality of summary[4]. This paper describes the relation between several F0 parameters and importance degree of sentences in lecture speech, and effectiveness of introduction of the F0 parameters in the speech summarization.



(a) without prosody (b) with prosody

**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 reconstruct the 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. In this paper, the speech summarization is also defined as extraction of important sentences from transcribed text. Lecture speech are 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 transcribed text.

### 2.2. Automatic Scoring of Sentence Importance

#### *2.2.1. F0 Parameters*

Prosodic variations due to speaker's intention or emphasis sometimes appear in a smaller unit, such as a word or a phrase, rather than a sentence. In this paper, F0 is analyzed *bunsetsu* by *bunsetsu* for a sentence, and several F0 parameters of the sentence are calculated. The *bunsetsu*, which is a syntactic unit of Japanese and consists of a content word followed by function words, is almost equivalent to a prosodic minor phrase.

The first type of F0 parameters of the sentence includes

- Fmin = $\min(f_1, f_2, \cdots, f_K)$,
- Fmax = $\max(f_1, f_2, \cdots, f_K)$, and
- Frange = Fmax - Fmin,

where $K$ is the number of *bunsetsus* in the sentence, and $f_i$ is the average F0 of $i$-th *bunsetsu* in the sentence.

These parameters absolutely evaluate F0 information. When intention or emphasis is prosodically expressed, however, prosodic parameters are relatively changed in comparison with ordinary utterance. The differences between observed F0 and ordinary F0 can be calculated by introducing an F0 model, supposing that the F0 model predicts an ordinary F0 pattern. The second type of F0 parameters is normalized by the predicted ordinary F0 and it includes

- NFmin = $\min(f_1 - \hat{f}_1, \cdots, f_K - \hat{f}_K)$,
- NFmax = $\max(f_1 - \hat{f}_1, \cdots, f_K - \hat{f}_K)$, and
- NFrange = NFmax - NFmin,

where $\hat{f}_i$ is the predicted average F0 of $i$-th *bunsetsu.*

*2.2.2. Linguistic Score*

A linguistic information is substantially important in identifying important sentences, while prosodic information may give supplementary cues for the summarization. A linguistic score predicting the sentence importance, which is noted as LING in this paper, is generated by a Japanese text summarization system, Posum[5]. The Posum system can calculate an importance score of each sentence for Japanese written text, based on linguistic information, such as word importance, relation of words, and so on.

*2.2.3. Combination of prosodic and linguistic information*

To predict the sentence importance, the prosodic and the linguistic information are combined as follows.

- LING + $w\times$ PROS. (1)

where PROS means one of above six F0 parameters, and weighting factor, $w$, is independently optimized for each F0 parameter.

### 2.3. F0 Model

An F0 model is necessary to extract differences between an observed F0 contour and an ordinary F0 contour of the same sentence. An F0 contour predicted by the F0 model is regarded as the ordinary F0 contour. The differences between two F0 contours are obtained by comparison of the *bunsetsu* F0 average.

The authors has proposed an F0 model in which the F0 contour of a sentence is represented by concatenation of F0 patterns of *bunsetsu*[6]. An *bunsetsu* F0 pattern is described both by the *bunsetsu* F0 average which is an average pitch value in the *bunsetsu* and the *bunsetsu* F0 shape which is a normalized F0 contour of the *bunsetsu* with its average zero. Fig. 2 shows a schematic F0 contour of a sentence based on the shape and the average of *bunsetsus* which are depicted by the solid and the dotted line, respectively.

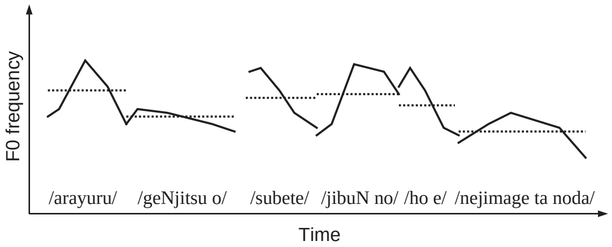


**Fig. 2.** F0 contour based on the shape and the average of bunsetsus.

The *bunsetsu* F0 average is predicted by the quantification theory (type I)[7], which estimates a numerical value from several categorical input features using the equation

$$\hat{y} = \bar{y} + \sum_{j=1}^{M} \sum_{v=1}^{V_j} \delta_{jv} x_{jv}, \tag{2}$$

where $\hat{y}$ is the estimated value, $\bar{y}$ is the mean value of training samples, $M$ is the number of input features, $V_j$ is the number of value types for the $j$-th feature, and $\delta_{jv}$ is defined as

$$\delta_{jv} = \begin{cases} 1 & \text{if the } j\text{-th feature takes the } v\text{-th value} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The weights, $x_{jv}$, are obtained by minimizing the total RMS error of estimated values.

The basis of this *bunsetsu* F0 average modeling is the same as the global model in Abe's model[8]. The linguistic features of the *bunsetsu* used for the quantification theory are listed as follows.

- The lexical accent type of the preceding, the current, and the following *bunsetsu.* (3 categories)
- The part-of-speech of the preceding, the current, and the following *bunsetsu.* (8 categories)
- The syllable number in the preceding, the current, and the following *bunsetsu.* (7 categories)
- The preceding and the following boundary types of the *bunsetsu.* (4 categories)

The F0 average is independently predicted for each *bunsetsu.*

## 3. EVALUATION

### 3.1. Speech Data

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

### 3.2. Sentence Importance

The summarization experiments were carried out to obtain the importance score of sentences. The number of the subject is 18 and 13 for data-1 and -2, respectively. The subjects watched the recorded video of the lecture to understand the contents. Then, they were asked to select both

**Table 1**. Speech Data

| data ID | data-1 | data-2 |
|---|---|---|
| contents | vitality of aged persons | regeneration of beach |
| speaker's gender | female | male |
| number of sentence | 68 | 71 |

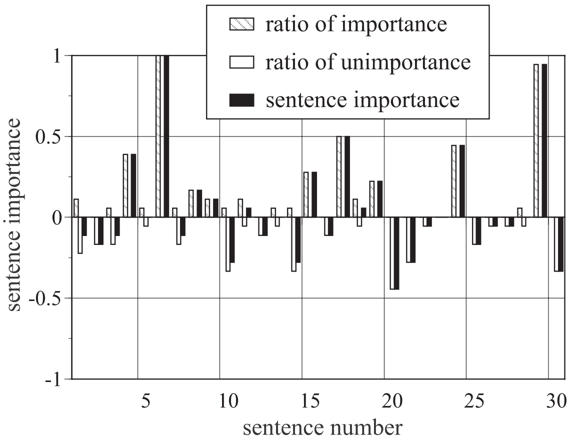


**Fig. 3**. Examples of sentence importance.

**Table 2**. Correlation coefficients with sentence importance.

| data | data-1 | | data-2 | |
|---|---|---|---|---|
| parameter | linguistic score | | linguistic score | |
| | without | with | without | with |
| Fmin | -0.315 | 0.498 | -0.289 | 0.582 |
| Fmax | 0.229 | 0.501 | 0.224 | 0.572 |
| Frange | 0.351 | 0.502 | 0.336 | 0.589 |
| NFmin | -0.396 | 0.527 | -0.272 | 0.566 |
| NFmax | 0.138 | 0.498 | 0.160 | 0.575 |
| NFrange | 0.351 | 0.505 | 0.237 | 0.581 |
| LING | - | 0.497 | - | 0.559 |

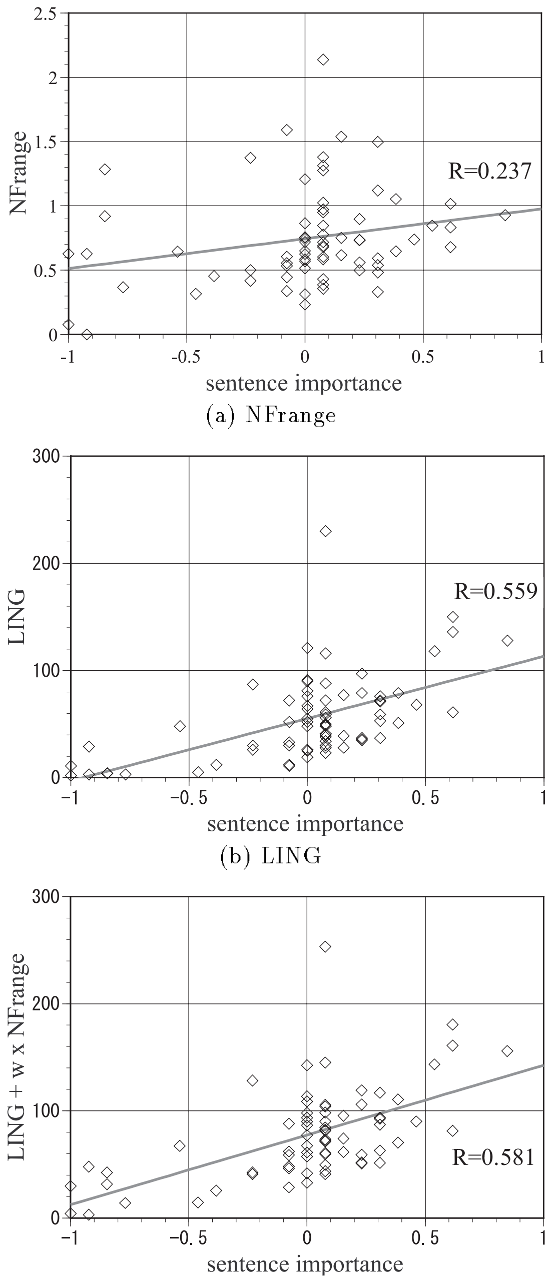
about 10 important sentences and about 10 unimportant sentences from all sentences in the lecture using its transcription, during listening the speech without visual information.

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

$$SI(i) = R(i)_{imp} - R(i)_{unimp} \qquad (4)$$

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. 3.

To investigate reliability of the sentence importance, the correlation coefficients among subjects are calculated in terms of average of correlation coefficients between one subject and the other subjects. They are 0.58 and 0.67 for data-1 and -2, respectively.



(a) NFrange

(b) LING

(c) LING + $w\times$ NFrange

**Fig. 4**. A scattering plot of sentence importance and the NFrange parameter combined with linguistic information.

### 3.3. Evaluation Results

*3.3.1. Correlation coefficients*

Table 2 shows correlation coefficients between the sentence importance and the F0 parameters with and without the linguistic score. The correlation coefficients with only linguistic score, LING, is listed in this Table for the baseline,

**Table 3**. Automatic extraction of important sentences.
(a) agreement to the $N$ most important sentences

| data ID | data-1 | | | | data-2 | | | |
|---|---|---|---|---|---|---|---|---|
| ling. score | without | | with | | without | | with | |
| $N$ | 10 | 15 | 10 | 15 | 10 | 15 | 10 | 15 |
| Fmin | 3 | 6 | 4 | 7 | 6 | 8 | 4 | 7 |
| Fmax | 2 | 4 | 4 | 6 | 2 | 6 | 5 | 8 |
| Frange | 3 | 6 | 4 | 7 | 5 | 7 | 5 | 8 |
| NFmin | 2 | 4 | 4 | 7 | 3 | 6 | 4 | 8 |
| NFmax | 2 | 4 | 4 | 6 | 2 | 4 | 5 | 8 |
| NFrange | 2 | 4 | 4 | 6 | 1 | 4 | 5 | 8 |
| LING | - | - | 4 | 6 | - | - | 4 | 7 |

(b) agreement to the $N$ least important sentences

| data ID | data-1 | | | | data-2 | | | |
|---|---|---|---|---|---|---|---|---|
| ling. score | without | | with | | without | | with | |
| $N$ | 10 | 15 | 10 | 15 | 10 | 15 | 10 | 15 |
| Fmin | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 |
| Fmax | 0 | 3 | 0 | 0 | 1 | 2 | 0 | 1 |
| Frange | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| NFmin | 0 | 1 | 0 | 0 | 1 | 2 | 0 | 1 |
| NFmax | 1 | 2 | 0 | 1 | 1 | 2 | 0 | 1 |
| NFrange | 1 | 1 | 0 | 0 | 1 | 2 | 0 | 1 |
| LING | - | - | 0 | 1 | - | - | 0 | 1 |

and they are 0.497 and 0.559 for data-1, and -2, respectively. Introduction of the F0 parameters increases the correlation. By comparing the F0 parameters without the linguistic score, Fmin and Frange have larger correlation with the sentence importance. However, it is not clear for the case with the linguistic score, The F0 parameters based on the F0 model, NF$x$, give almost the same performance as the simple F0 parameter, F$x$, except for NFmin for data-1. Since the F0 models are trained with the closed data of each lecture speech, the amount of data is not sufficient to get an accurate F0 model. Performance of normalized NF$x$ parameters is potentially improved if the ordinary F0 is predicted by a more accurate F0 model.

Fig. 4 (a), (b) and (c) depict scattering plots of the sentence importance and measures for predicting the sentence importance, which are NFrange, LING, and NFrange combined with LING for data-2, respectively. Comparison of (a) and (c) shows that the linguistic information increases correlation coefficients drastically. Fig. 4 (b) and (c) are almost the same plot though the correlation coefficient in (c) is slightly large than in (b).

### *3.3.2. Extraction of important sentences*

Table 3 shows results of the sentence extraction based on the F0 parameters and the linguistic score. Table 3 (a) indicates the number of sentences which match with one of the $N$ most important sentences, when $N$ sentences are automatically extracted for $N$=10 and $N$=15. Table 3 (b) indicates the number of matched unimportance sentences in the same manner. Figures in this Table (a) and (b) should be $N$ and 0, respectively, when important sentences are perfectly extracted. For several cases, the number of the matched sentences in Table 3 (a) and (b) successfully increases and decreases by utilizing F0 parameters, respectively.

## 4. CONCLUSIONS

This paper discusses the speech summarization using F0 information. Introduction of F0 information improves both correlation of the sentence importance and measures for predicting the sentence importance and extraction of important sentences, in comparison with only linguistic information. In order to obtain further improvement, it is necessary to investigate how to extract F0 parameters and other prosodic features and how to combine the prosodic and linguistic information. People possibly express intention and emphasis in a spoken message in different prosodic manners. The analysis of various types of spoken monologue by many speakers is also necessary.

## 5. ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for Scientific Research of Priority Areas (B) "realization of advanced spoken language information processing from prosodic features" (No. 12132203) from Ministry of Science, Culture, Sports, Education.

## 6. REFERENCES

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

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

[3] J. Hirschberg and C. Nakatani, "Acoustic indicators of topic segmentation," in *Proc. of ICSLP '98*, 1998, vol. 4, pp. 1255–1258.

[4] K. Koumpis, S. Renals, and M. Niranjan, "Extractive summarization of voicemail using lexical and prosodic feature subset selection," in *Proc. of Eurospeech 2001*, 2001, vol. 4, pp. 2377–2380.

[5] H. Mochizuki, "Posum," http://galaga.jaist.ac.jp:8000/˜motizuki/software/posumcl/.

[6] Y. Yamashita and T. Ishida, "Stochastic f0 contour model based on the clustering of f0 shapes of a syntactic unit," in *Proc. of Eurospeech 2001*, 2001, vol. 1, pp. 533–536.

[7] C. Hayashi, "On the quantification of qualitative data from the mathematicostatistical point of view," *Ann. Inst. Statist. Math.*, vol. 2, 1950.

[8] M. Abe and H. Sato, "Two stage f0 control model using syllable based f0 units," in *Proc. of ICASSP '92*, 1992, pp. 53–56.