NEWSCAST SPEECH SUMMARIZATION VIA SENTENCE SHORTENING BASED ON PROSODIC FEATURES

Kiyonori Ohtake[1], *Kazuhide Yamamoto*[1,2], *Yuji Toma* [3], *Shiro Sado*[3], *Shigeru Masuyama*[3], *and Seiichi Nakagawa*[4]

[1] ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288 Japan

[2] Department of Electrical Engineering, Nagaoka University of Technology, Niigata 940-2188 Japan

[3] Department of Knowledge-based Information Engineering, Toyohashi University of Technology, Aichi 441-8580 Japan

[4] Department of Information and Computer Sciences, Toyohashi University of Technology, Aichi 441-8580 Japan

E-mail: {otake,yamamoto}@fw.ipsj.or.jp, {nerusyu, masuyama}@smlab.tutkie.tut.ac.jp

ABSTRACT

This paper presents a speech summarizer that summarizes input speech via several prosodic features, unlike models that use a speech recognizer and conventional summarizing techniques proposed in natural language processing. Our approach analyzes the borders of summary units by employing prosodic features of pitch, power, and pause to summarize the input speech. Our summary generation trial implies robustness against noisy input compared with both a sequential connection model of a speech recognizer and a text summarizer.

1. INTRODUCTION

Speech is an effective communication medium for human-to-human communication and is thus used commonly in daily life. As interest in the study of automatic document summarization in the field of natural language processing continues to increase, more attention is being paid to speech summarization, where both the input and the output are speech, rather than text.

One may think that when we use a speech recognizer, the task of speech summarization is the same as written text summarization[1]. While this may be so, there may also be an alternative approach to realizing a speech summarizer that processes speech directly. Speech information involves not only linguistic information, but other types of information, such as emphasis and nuances. For example, conventional wisdom tells us that prosody must be somehow related, directly or indirectly, to syntactic structure. Thus, there is room to try to build a summarizer that is suitable for speech, using this information that text cannot effectively convey.

We present a new approach to summarize, or shorten, input newscast speech by utilizing prosodic information. Throughout the approach's implementation, we do not use a speech recognizer because we want to investigate the feasibility of a direct speech summarizer that is not connected with a speech recognizer. One advantage of this strategy is that the system is not affected by noise and errors due to the recognizer, which is the biggest headache in spoken language processing.

In newscast programs, the sentences that are read by broadcasters tend to be very long. If we simply select some important utterances as a summary of a news topic, the result of the summarization may not be acceptable because it may lack important information. Therefore, we try to summarize each utterance by selecting some of its important parts.

We use prosodic information to determine the unit for summarization, to analyze the dependency structures[2, 3] between the units, and to determine the importance of each unit. We will build a summarization method based on this idea, and apply it to Japanese broadcast news speech. In addition, we will compare the proposed method with a conventional method, i.e., using a speech recognizer and a text summarizer to investigate the effect of direct speech summarization.

2. NEWSCAST SPEECH SUMMARIZATION

An overview of the summarization method is described in the following five steps. **Step 1**: Extract pitch pattern with the method proposed by Hirose et al.[4]; **Step 2**: Extract accent phrases; **Step 3**: Detect summary units; **Step 4**: Analyze dependency structures of summary units; and **Step 5**: Select important summary units.

We describe the details contained in and after Step 2 in the following sub-sections.

2.1. Extracting accent phrases

An accent phrase, which is sometimes called a 'prosodic phrase,' can usually be observed as a *fall-rise pattern* of $F0$, and the $F0$ contour of a sentence uttered continuously is considered to be a connected series of these prosodic patterns. Based on this idea, we extract phrase boundaries by detecting each rise in the patterns of the $F0$ contour.

To pick up the rise in $F0$, we calculate the following formula:

$$diff(t) = \frac{1}{w^2} \int_{-w/2}^{w/2} \tau f(t+\tau) d\tau. \quad (1)$$

Here, w represents the window size, and $f(t)$ is $F0$ at time t. However, some t's do not have an $F0$ value, because $F0$ is not observed at unvoiced sounds. Although such times can usually be interpolated, we use the last value of $F0$ for an $f(t)$ that has no value. Moreover, this operation has little impact on the processes that use this $diff(t)$ value.

The beginning of an accent phrase is picked up as follows: In an interval, where $diff(t) > 0$, the smallest t that maximizes $diff(t)$ is used, then the beginning of the accent phrase is given as $t' = t - w/2$. The end of one accent phrase is given as the beginning of the next accent phrase.

2.2. Detecting summary units

A summary unit consists of consecutive accent phrases. We extract unit boundaries by calculating regression lines that correspond to fall-patterns of the base-pitch (Bp) contour. The base-pitch (Bp) is a representative value of an accent phrase A, and is given by the following formula:

$$Bp(A) = 0.2F0_{max}(A) + 0.8F0_{min}(A), \quad (2)$$

where $F0_{max}(A)$ indicates the maximum value of $F0$ in phrase A, and $F0_{min}(A)$ indicates the minimum value of $F0$ in phrase A. We employ $Bp(A)$ instead of $F0_{min}(A)$ as the representative value of an accent phrase A to reduce some errors that are caused by the pitch extractor and other miscellaneous sources.

To approach the Bp contour with a regression line, we employ a mel scale, which is defined by the following formula:

$$m(f) = 2410 \log_{10}(1 + \frac{1.6f}{1000}), \quad (3)$$

where f is the frequency in Hz.

We calculate regression lines $l_{ij}(t)$ that minimize the mean squared error ϵ_{ij} for all combinations of any two accent phrases $(A_i, A_j, i < j)$.

$$l_{ij}(t) = a_{ij}t + b_{ij} \quad (4)$$

$$\epsilon_{ij} = \frac{1}{j-i+1} \sum_{k=i}^{j} z(A_k)(m(Bp(A_k)) - l_{ij}(x(A_k)))^2, \quad (5)$$

where, $z(A)$ indicates the interval length that is not silent in phrase A, and $x(A)$ indicates the center time point of phrase A.

We preserve the regression lines that fulfill the two conditions, $\epsilon_{ij} \leq \theta_\epsilon$, $a_{ij} < 0$, and eliminate the remaining regression lines.

Regression lines that fulfill the following two conditions are also eliminated.

- $a_{ij} < -\theta_s$
- In the accent phrases $(A_i, A_{i+1}, \cdots, A_j)$ corresponding to l_{ij}, there is a silent pause between any two adjoining phrases, given as $(A_k, A_{k+1}, i \leq k < j)$.

In addition, the regression line l_{ij} is eliminated, if $l_{i'j'}$ exists, where $i' \leq i, j \leq j'$.

The remaining regression lines basically correspond to summary units. Moreover, each accent phrase that does not correspond to a regression line is considered as a summary unit.

We introduce heuristics to adjust the boundaries of the summary unit, because there are mis-extractions of boundaries caused by pitch that rises to express emphasis or paraphrasing. We connect two summary units that sandwich such a boundary and unite them as one unit.

The processing of connection is as follows:

If there is a silent pause whose length is less than or equal to θ_p between two consecutive summary units (C_k, C_l), and the beginning accent phrase A_i of C_l satisfies the following formula, then C_k and C_l will be connected.

$$Bp(A_{i-1}) \leq Bp(A_i) \leq Up(A_{i-1}), \quad (6)$$

where the function $Up(A_i)$ expresses the upper 20% of the distribution of $F0$ in the accent phrase A_i, as defined by the following formula:

$$Up(A_i) = 0.8F0_{max}(A_i) + 0.2F0_{min}(A_i). \quad (7)$$

On the other hand, if there is a silent pause whose length is more than θ_p between two consecutive summary units (C_k, C_l), and the beginning accent phrase A_i of C_l satisfies the following formula, then C_k and C_l will be also connected.

$$Up(A_i) - Bp(A_i) < Up(A_{i-1}) - Bp(A_{i-1}) \quad (8)$$

2.3. Analyzing dependency structure

In this paper, we also introduce an easy dependency analysis method. This method analyzes the dependency structures among summary units of the input speech.

If there is a silent pause whose length is less than or equal to θ_p in any two consecutive summary units (C_k, C_l), then C_k depends on (modifies) C_l. Otherwise, the summary unit depends on the last summary unit in the utterance.

2.4. Selecting summary units

The method for generating the summary is as follows: (1) Select the summary units, then (2) concatenate the selected units.

We introduced three selection methods. Method 1 is based on common heuristics for broadcast news texts, method 2 is based on the heuristics of prosodic features derived from the observation of manually generated summaries, and method 3 is similar to method 2, but we free it from parameters.

The details of each method are as follows:

[method 1]

(1) The first sentence: select all units. (2) The second sentence: eliminate the first summary unit. (3) After the third sentence: select the last two summary units for each sentence.

[method 2]

(1) Select all summary units in the first sentence. For all sentences after the second sentence, the following items are applied: (2) Eliminate summary units that hold the mean value of Bp for all accent phrases in units more than or equal to a speaker-dependent threshold, e.g., 150 Hz. (3) Eliminate summary units that have a mean power less than or equal to 0.01 dB. (4) Eliminate summary units that depend on other units eliminated by the above items.

The values 150 Hz and 0.01 dB, were determined by a preliminary parameter estimation experiment.

[method 3]

(1) Select all summary units in the first sentence. For all sentences after the second sentence, following items are applied: (2) Eliminate a summary unit if the mean value of Bp in the summary unit is greater than or equal to the top 30% of the range of the mean value of Bp in the sentence. (3) Eliminate summary units that depend on other units eliminated by the above items.

3. EXPERIMENTS

We carried out several experiments to investigate the following points: (a) Whether the summary unit is an effective unit in speech summarizing, and (b) whether the speech summarized by the proposed method is natural.

We employed an NHK news speech database (16 kHz, 16 bit) for these experiments.

3.1. Estimating parameters

The summarizing method requires some parameters, which we determined in a preliminary parameter estimation experiment. Three

Table 1. Detecting summary units with/without heuristics

	(a)	(b)	(c)	P	R	F
with	123	87	112	70.7%	77.7%	74.0%
without	213	102	112	47.9%	91.1%	62.8%

(a): # of detected boundaries P: Precision
(b): # of hit boundaries R: Recall
(c): # of correct boundaries F: F-measure

Table 2. Speech summarization ratio (SS-ratio) by three methods

	Avg	Max	Min	SD
Method 1	72.4%	98.7%	53.2%	8.25%
Method 2	70.1%	96.1%	53.3%	7.55%
Method 3	80.7%	91.2%	74.7%	6.63%

articles, containing 15 sentences, were randomly selected from the NHK news speech database. These articles were all uttered by a male broadcaster.

First, we estimated the following parameters by trial and error. The estimated parameters were as follows: $\theta_e = 1000\,\text{mel}^2$, $\theta_s = 0.1$ mel/s, $\theta_p = 0.5$, and $w = 500$ ms.

Second, we estimated some parameters for the selection method. To estimate these parameters, we had eight subjects summarize three transcribed articles by selecting the most important parts. We created good examples of summaries by selecting parts chosen by more than three subjects. We then estimated the parameters required for selecting methods based on these good examples.

3.2. Detecting summary units

Ten new articles, containing 52 sentences, were randomly selected from the NHK news speech database. These articles were all uttered by a male broadcaster. We also prepared the correct summary unit boundaries for these ten articles by hand. We applied our method to these articles and evaluated them by comparing the results with the manually prepared answers. In addition, to evaluate the effectiveness of the heuristics introduced in Section 2.2, we applied our method without the heuristics. The evaluation results are shown in Table 1.

3.3. Evaluating speech summarization

We summarized ten new articles randomly selected from the NHK news speech database with the three selection methods introduced in Section 2.4.

First, we calculated the speech summarization ratio (SS-ratio), which is defined by the following formula for each article.

$$\text{SS-ratio} = \frac{\text{playback time of summary}}{\text{playback time of original}} \times 100(\%) \qquad (9)$$

Table 2 shows the results of the calculation for the average, maximum value, minimum value, and standard deviation (SD).

Second, we conducted a questionnaire-based survey with eight subjects on the preservation ratio of important information and phonological naturalness of the summarized speech, and the original speech for the ten articles. The subjects evaluated each article using a scale of 1 to 5 (best: 5, worst: 1) with respect to the two points above. We took the average of eight evaluations for each article, and computed the average, maximum value, minimum value, and standard deviation (SD) for all ten articles. Table 3 presents the results for the preservation ratio of important information, and Table 4 shows the results for phonological naturalness.

Table 3. Results for preservation ratio of important information

	Avg	Max	Min	SD
Original	4.9	5.0	4.5	0.17
Method 1	3.9	4.5	2.5	0.69
Method 2	4.3	4.8	3.0	0.60
Method 3	3.7	5.0	2.6	0.69

Table 4. Results for phonological naturalness

	Avg	Max	Min	SD
Original	4.9	5.0	4.3	0.24
Method 1	4.0	4.5	3.5	0.50
Method 2	4.4	4.8	3.9	0.33
Method 3	3.6	4.3	2.4	0.58

We also compared our method with a sequential connection model of a speech recognizer and a text summarizer. This experiment was carried out as follows: Ten articles randomly selected from the speech database were transcribed by a speech recognizer. We then applied a text summarization method developed by Mikami et al.[5] to the transcribed article. Next, we summarized the article with our speech summarizer using selection method 2, and mapped the result onto the transcribed article. Finally, we compared the two summarization results.

To compare the results, we had six subjects summarize the ten articles by selecting the important parts, with the condition that the text-based summarization ratio is 70%. We can create good examples of summaries by selecting the parts chosen by more than four subjects; the average summarization ratio of these good examples is 56.8%.

For those good examples, we calculated recall and precision of two summaries. The results are shown in Table 5 with the summarization ratio (S-ratio).

4. DISCUSSIONS

4.1. Analysis of summary unit boundaries

In our approach we have introduced a heuristic, wherein base pitch increases caused by a paraphrase or an emphasis expression are not regarded as boundaries. This heuristic results in lower recalls, and thus changes of summarization ratios, by changes of articles. Eight instances related to this reason are observed among the 25 instances in which summary units were not recognized.

Some pauses tend to be inserted when parallel clauses are segmented in utterance. These pauses are mis-recognized by our approach as boundaries of summary units. Precision decreases down for this reason, and it mainly affects phonemic naturalness. There were 12 incidents observed among 36 analysis errors.

The parameters for summary unit boundary detection, fixed by the preliminary experiments, depended on each feature of the article in general. There were certainly some errors due to inap-

Table 5. Comparing two summaries

	S-ratio	SD of S-ratio	P	R
Ours	73.7%	12.7%	69.1%	48.3%
Mikami's	51.0%	20.8%	66.8%	36.5%

propriate parameter values. This static parameter determination causes both summarization ratio changes and errors in selection of summary units.

Some analysis errors due to the pitch extractor were also observed. For instance, voiceless sounds, such as a fricative /s/, are regarded as being spoken with no vibration of the vocal cords. It therefore does not have a pitch and no detection is possible by a pitch extractor. Consequently, in a part where no pitch information is observed, a distinction between fricatives and silent parts is required.

The dependency analyzer also made some errors: it judged that the dependent part was the predicate even when it was in fact following part. This type of errors causes the analyzer to fail in deletion. Phonemic naturalness is affected by these errors.

The errors listed here were the major errors observed. Although many errors were observed in this experiment, the evaluations shown in Tables 3 and 4 illustrate a fine level of performance.

4.2. Comparison of Strategies

Tables 3 and 4 show that method 2 is the best method among the three, considering both the preservation ratio of important information and phonemic naturalness, indicating that the thresholds we have determined fit the inputs for the experiment. However, it is known that the base pitch greatly differs between male and female newscasters. Moreover, the degree of power depends on the article, even when the speaker is the same. These two factors may cause numerous performance changes.

The performance of method 1 was inferior to that of method 2, but better than that of method 3 because method 1 is less likely to reduce the amount of important information or break phonemic naturalness than the others, since it employs features that depend on newscast speech. However, it does not outperform method 2 because the inputs did not necessarily contain the features we had expected.

As the maximum preservation ratio of important information illustrates, method 3 works very well for some articles. However, the method is the worst among the three. This result indicates that the method, in which the threshold for selecting summary units is relatively determined by the pitch range of the input, did not work for most of the articles.

4.3. Summarized Speech

We have obtained fine evaluations of summarized speech, for both naturalness of phonemes and in the preservation ratio of important information. We assume that these evaluations originate in the character of news speech, which is clearer and more well-formed than ordinary daily speech. However, some speech in the NHK news speech database included background sounds such as music, which may decrease the performance. Thus, it can be said that our approach works well, even if there is some background noise in the input. We expect that the proposed approach may be applicable not only to broadcast speech, but also to more difficult tasks, such as monologues spoken in lectures and meetings.

Moreover, in a situation where a speech recognizer may make mistakes in recognizing particles, better performance is expected when using the parsing approach proposed in Section 2, as compared with that by Mikami et al.[5]. Unfortunately, Table 5 does not show us the striking difference between the performances, measured by recall and precision, of the two approaches.

However, Table 5 does tell us that the result for the sequential connection model was very unstable. Due to mistakes made by the speech recognizer, it is difficult to determine the correct parts that should be deleted. The summarizer tends to delete long parts because it is difficult to determine the correct boundaries suited for summary. On the other hand, our method achieved stable performance, and the summarization results were more natural.

5. CONCLUSION

This paper describes our attempt to build a speech summarizer that summarizes input speech via several prosodic features, unlike models that use a speech recognizer and the conventional summarizing techniques proposed in natural language processing.

Our experiments have proved that a summarizing strategy based on the base pitch and power of the target summary units attains the best performance among the three strategies that we compared. However, to summarize a variety of speeches, we need to determine the parameters dynamically, and that depends on the characteristics of the target input.

Our summary generation trial has proved some advantages of the proposed method in terms of robustness against noisy input, compared with a sequential connection model of a speech recognizer and a text summarizer.

Further investigations are necessary to (1) determine parameters, (2) recognize parallel structure against clauses, and (3) consider other prosodic features, such as speed.

6. ACKNOWLEDGMENT

This research was supported in part by the Telecommunications Advancement Organization of Japan, Grants-in-Aid and the 21st Century COE Program from the Ministry of Education, Science and Culture of Japan. The authors would like to thank NHK (NIPPON HOSO KYOKAI) for providing Toyohashi University of Technology with the NHK news speech database.

7. REFERENCES

[1] Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex Waibel, "Automatic speech summarization applied to English broadcast news speech," in *Proceedings of ICASSP 2002*, 2002, vol. I, pp. 9–12.

[2] Yasuo Horiuchi and Akira Ichikawa, "Prosodic structure in Japanese spontaneous speech," in *Proceedings of ICSLP 98*, 1998, pp. 591–594.

[3] Norihiro Eguchi and Kazuhiko Ozeki, "Dependency analysis of Japanese sentences using prosodic information," *The Journal of The Acoustical Society of Japan*, vol. 52, no. 12, pp. 973–978, 1996, (in Japanese).

[4] Keikichi Hirose, Hiroya Fujisaki, and Shigenobu Seto, "A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag," in *Proceedings of ICASSP-92*, 1992, vol. I, pp. 149–152.

[5] Makoto Mikami, Yuko Ishizako, Hirotaka Akamatsu, Shigeru Masuyama, and Seiichi Nakagawa, "An experiment on generating captions by summarizing using speech recognition results of broadcast news," in *Proceedings of the 58th Meeting of the Information Processing Society of Japan*, 1999, vol. 3W-5, pp. 275–276, (in Japanese).