INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *23*(1–2), 115–129

A New Multimedia Content Skimming Technique at Arbitrary User-Set Rate Based on Automatic Speech Emphasis Extraction

Kota Hidaka
Shinya Nakajima
NTT Cyber Solutions Laboratories, NTT Corporation, Japan

This article proposes a new technique for skimming multimedia content such as video mail, audio/visual data in blog sites, and other consumer-generated media. The proposed method, which is based on the automatic extraction of emphasized speech, locates emphasized portions of speech with high accuracy by using prosodic parameters such as pitch, power, and speaking rate. As the method does not employ any speech recognition technique, it enables a highly robust estimation in noisy environments. To extract emphasized portions of speech, the method introduces a metric, "degree of emphasis," which indicates the degree of emphasis of each speech segment. Given an article, the method computes the degree of emphasis for each speech segment in it. When a user requests a skimming of the article's content, the method refers to the user-specified "skimming rate" to collect the emphasized segments. Preference experiments were performed in which participants were asked to select either the skimmed contents created by our method or those created using a fixed interval approach. The preference rate of our method was about 80%, which suggests that the proposed method can generate proper content skimming.

1. INTRODUCTION

Advances in broadband network technologies offer the opportunity to distribute multimedia contents. For instance, video mail exchange is now getting more popular through the broadband network. With its highly intuitive communication ability, video has become increasingly important, and video mail is very useful in this respect. Another significant aspect of video mail is that neither the sender nor

We thank Katsuhiko Ogawa, Director of NTT Cyber Solutions Laboratories, for encouraging us to write this article. We have had helpful discussions with Osamu Mizuno, Broadband Business Development Division, Corporate Business Headquarters, NTT EAST, Junji Takeuchi, Visual Communications Division, NTT Bizlink, and the office support staff of Megumi Machiguchi, Kaori Takei, and Harumi Matsuura.

Correspondence should be addressed to Kota Hidaka, 1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847, Japan. E-mail: hidaka.kota@lab.ntt.co.jp

the receiver is tied down, in contrast with real-time video communication such as video telephony. The question we have to ask here is whether an efficient method can be provided to enable users to grasp video content given that the amount of video mail will increase rapidly. It is important to view each video mail in the shortest possible time. Another aspect of the broadband network is that sites on the Internet to which anyone can contribute personal audio-visual data are getting more prevalent. Furthermore, blogs have become popular for disseminating information, and many more sites now include audio-visual content. Consumer-generated media (CGM) also has infiltrated multimedia communication. Although multimedia content distribution across the network is increasing, the time one person can spend watching content has remained unchanged. (It is said to be around 4 hr a day on average in Japan.) The goal of our work is to provide an efficient interface by which people can more rapidly browse or search multimedia content without losing too much information. Given this goal, this article proposes a new multimedia content skimming technique for applications such as video mail, blogs that included audiovisual content, and CGM.

The following section describes related work and our objectives. The subsequent section gives the proposed multimedia content skimming method. In section 3, we also describe the results of evaluation experiments in which we assess the quality and adequacy of the skimmed content generated by the proposed method. The final section gives our conclusions and future work.

2. RELATED WORK AND OBJECTIVES

Before we examine multimedia content skimming techniques, it will be useful to discuss the content on which this article focuses attention in more detail. In this article, multimedia content is assumed to comprise unedited, natural speech, including shouts, cheers, and the like, captured in a variety of environments including noisy ones. Various studies have demonstrated a number of ways through which users can confirm contents effectively. Lienhart, Pfeiffer, and Effelsberg (1997) proposed a drama content summarization technique based on recognizing actors' faces, extracting explosions or gunfire on audio, and giving priority to scenes that include them. Moriyama and Sakauchi (2001) employed a technique that utilizes the frequency of scene changes, the ratio of speech or background music, and heuristic knowledge of sound to summarize video dramas statistically. Aoyagi, Sato, Takada, Sugawara, and Onai (2005) stated that an effective approach for summarizing edited video with educational content is to give priority to portions containing loudness and use the nearest scene changes occurring before or after them. He, Sanocki, Gupta, and Grudin (1999) focused on the time needed to change presentation slides, portions of silence, and portions of high-pitched speech to summarize presentation contents.

Such information is seen to be of value in structured contents such as news and dramas, becausethe video characteristics useful for skimming these contents can be extracted and represented. However, we are concerned with handling unedited multimedia contents having neither scene changes nor camera parameters.

This question could be addressed with an audio-based approach. Some studies based on speech recognition technology appear to be of value in the skimming of clear speech (Hori & Furui, 2003) or natural speech on a definite theme. However, as mentioned in the previous section, we are targeting content, characterized in the following list, made by nonprofessional network users:

- noisy environment
- inarticulate speech
- low semantics
- high verbiage
- indefinite themes
- existence of more than one speaker

Although a number of speech-recognition-based studies on content skimming have been carried out, little is known about their feasibility when applied to contents such as video mail, blogs that include audio-visual data, and CGM. In addition, the speech-recognition-based approaches normally employ text summarization techniques and so are unable to capture paralinguistic information such as emphasis and other emotional factors, which play important roles in real conversations. Further, the speech-recognition-based approaches depend strongly on the language.

For skimming content, we focus attention on prosodic information. We propose a new multimedia content skimming technique based on the automatic extraction of emphasized speech (Hidaka, Mizuno, & Nakajima, 2002). The proposed method locates emphasized portions of speech by using prosodic parameters such as pitch, power, and speaking rate. In Hidaka et al. (2002), a simple new skimming method was proposed and no evaluation experiments were done. Because prosodic emphasizing seems to be less dependent on language, the proposed approach can be applied to other languages with minor customization. A few prosody-based skimming approaches are known to exist. Chen and Withgott (1992) considered the use of prosody as autonomous information in identifying emphatic speech for creating summaries of spoken discourse. Although their approach is close to ours, it differs essentially from ours in that they take emphasis as a more static and local phenomena. Their approach "recognizes" emphasized syllables and determines whether an acoustic phrase is emphasized or not by calculating the sum of the syllable-based emphasis probabilities. Because the phrases used for summary generation are limited to those recognized as emphasized, this approach can generate only one fixed summary for each content. As described in a later section, our approach treats emphasis as a relative, dynamic, and global phenomenon. In fact, our method generates "the degree of emphasis" for each acoustic phrase instead of recognizing "emphasized or not." By setting a variable threshold for the degree of emphasis curve, it can generate skimmed contents dynamically to match the user's requested skimming rate. Another minor difference is that they excluded speaking rate from their consideration despite the role it plays in prosodic information. There also appears to be differences with respect to varieties of speech and environments. In our approach, we define emphasized speech as those portions whose prosody deviates from that of typical speech. The method estimates the likelihood of emphasized speech from a global point of view and creates a skimmed content by paying attention to the global deviation of intonation.

3. THE PROPOSED SKIMMING METHOD

3.1. Overview of The Proposed Method

Figure 1 shows the flow of the proposed multimedia content skimming method. Speech information is separated from multimedia content and analyzed to estimate the degree of emphasis by using prosodic parameters. Because the duration of emphasized speech is quite short, the extracted portions do not exhibit semantic continuity. We can define portions of speech that can be understood semantically as *speech blocks*. Speech blocks with a high degree of emphasis are assigned priority. When a user requests to skim the multimedia content, the method refers to the user-specified skimming rate to collect the emphasized speech blocks.

The next subsection describes the estimation of the degree of emphasis. The subsection after that explains the extraction of speech block, and the final subsection describes multimedia content skimming in detail.

3.2. Estimation of the Degree of Emphasis

We first defined a set (eight) of emphasized speech labels for speech database marking. Two persons set 2,208 labels on natural speech material (duration of about 10 hr) created by two or more speakers. The prosodic characteristics of the labels are shown in Table 1. We can define speech having the prosodic deviation shown in the table as emphasized speech. It is clear that the prosodic deviations are concerned with pitch, power, and speaking rate. We used the speech parameters shown in Table 2 to extract prosodic characteristics. The pitch parameters of speech can be obtained by the autocorrelation function of linear prediction based

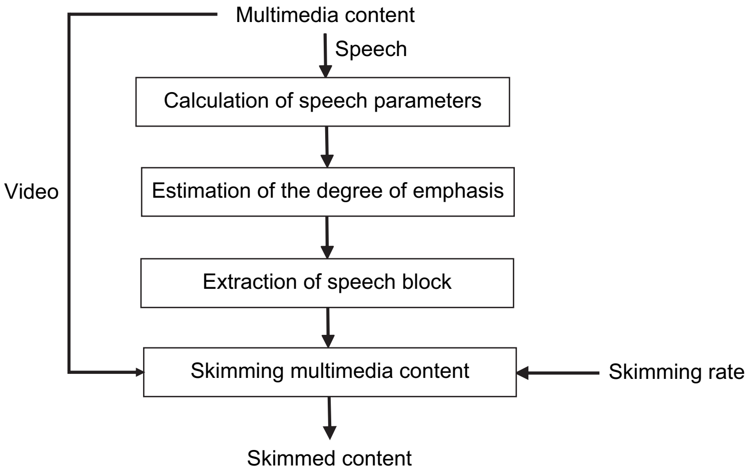

Multimedia content
Speech
Calculation of speech parameters
Video
Estimation of the degree of emphasis
Extraction of speech block
Skimming multimedia content
Skimming rate
Skimmed content

FIGURE 1 Flow of the proposed multimedia content skimming method.

Table 1: Classification of Emphasized Speech by Prosody

	Prosodic Characteristics
1	Strong and slow
2	Strong and high
3	Slow at start of speech
4	Remarkable differential between strong and weak, high and low
5	High suddenly
6	Strong by degrees
7	Strong, high, and fast speech
8	Suddenly weak and low

Table 2: Speech Parameters

Average of fundamental frequency	$\overline{F0}$
Differential component	$\pm\Delta\overline{F0}$
Average of power	$\bar{P}$
Differential component	$\pm\Delta\bar{P}$
Number of peaks in dynamic measure	dp

residual signals. The autocorrelation function tends to peak in synchronization with vocal cord vibration, that is, the pitch of voice. Therefore, pitch, technically fundamental frequency (*F0*) is extracted by peak-picking of the autocorrelation function. Power (*P*) is calculated as the logarithmic root squared sum of amplitude of the sound wave. It is reasonable to assume that averaging the *F0* obtained by using several data items effectively decreases the influence of pitch detection errors. The proposed method averages and defines both analysis window and window shift as 50 msec. Among these prosodic parameters, speaking rate cannot be computed directly. Instead, we use dynamic measure as an estimate of the speaking rate (Sagayama & Itakura, 1979). The dynamic measure is defined as the sum of squared slope coefficients of the cepstrum's regression line.[1] It generally reflects the dynamics of spectral movements in speech and normally peaks at phoneme boundaries. Given this, the number of the dynamic measure peaks per interval is a good estimate of the speaking rate. The number of peaks of dynamic measure (*dp*) is counted at 1-sec intervals; window shift is 50 msec. The deviation component of each parameter (i.e., $\pm\Delta\overline{F0}$ and $\pm\Delta\bar{P}$) is taken as the difference in previous/succeeding frame values. (Because speaking rate does not change drastically in such a short duration, we don't use its differential component.)

To extract emphasized speech, it is essential to distinguish speech portions that are emphasized from those that are not. We define the speech portions to which the prosodic characteristics shown in Table 1 are not applicable as "normal speech."

The likelihood of emphasized speech is estimated by using speech parameters. We examined the duration of emphasized speech, assuming it was always longer

[1] Cepstrum is a standard spectral parameter in the speech processing domain and is obtained by applying inverse-FFT to log area arithmic domain power spectrum (Bogert, Healy, & Tukey, 1963).

than the analysis window. It was found that the average duration of the emphasized speech portions is 700 msec and that 80% of them are less than 1 sec. For this reason, it is important to increase the temporal deviation of the speech parameters; longer portions are defined as "estimating units." Referring to the aforementioned duration analysis, we set the estimating unit to 1 sec and unit shift interval to 0.5 sec. Furthermore, for ascertaining the deviation of the speech parameters, it is reasonable to use vector-quantized prosodic parameters rather than the individual values. We employed the following approach in our analysis:

- Compute probabilities of emphasized and normal per analysis window.
- Estimate likelihoods of emphasized and normal per estimating unit.
- Estimate the degree of emphasis per estimating unit.

The procedure for extracting the degree of emphasis for estimating units can be described as follows. Speech parameters are calculated for each analysis window. The vector quantization technique called LBG method (Linde, Buzo, & Gray, 1980) is performed on the feature vectors. In the method, a set of feature vectors is split into two clusters via the minimum distortion criterion. The cluster-splitting procedure is applied repeatedly on each cluster until proper conditions are satisfied. After N-stage splitting, 2^N clusters are generated, and the centroid vector of each cluster is saved in a codebook of the feature vectors. Given that there are L frames in an arbitrary estimating unit, vectors are given by Cl ($l = 1, 2, \ldots, L$). To calculate the probability of emphasized P_E (f) and normal P_N (f) speech at arbitrary frame f, we apply the trigram model used in a study of natural language (Rabiner & Juang, 1993). The point to which special attention should be paid is capturing temporal pattern of prosodic characteristics with a few brief moments. This is possible with the trigram model based on vector transition. For the reason just stated, Conditional probabilities P_{emp} $(C_f | C_{f\text{-}1}C_{f\text{-}2})$, P_{emp} $(C_f | C_{f\text{-}1})$, P_{nrm} $(C_f | C_{f\text{-}1}C_{f\text{-}2})$, and P_{nrm} $(C_f | C_{f\text{-}1})$, and probabilities P_{emp} (C_f) and P_{nrm} (C_f) were computed in advance via statistical learning, described later, that is,

$$P_E(f) = \lambda_{e1}P_{emp}(C_f \mid C_{f\text{-}1}C_{f\text{-}2}) + \lambda_{e2}P_{emp}(C_f \mid C_{f\text{-}1}) + \lambda_{e3}P_{emp}(C_f) \tag{1}$$

$$P_N(f) = \lambda_{n1}P_{nrm}(C_f \mid C_{f\text{-}1}C_{f\text{-}2}) + \lambda_{n2}P_{nrm}(C_f \mid C_{f\text{-}1}) + \lambda_{n3}P_{nrm}(C_f) \tag{2}$$

λ_{ei}, λ_{ni} (i = 1, 2, 3) are weights. The likelihoods of emphasized P_{Xemp} and normal P_{Xnrm} are computed as

$$P_{Xemp} = \prod_{l} P_E(f) \tag{3}$$

$$P_{Xnrm} = \prod_{l} P_N(f) \tag{4}$$

Emphasized speech should be extracted to consider the relative prosodic deviation from normal speech. We define the degree of emphasis K_X as

$$K_X = \frac{\log P_{Xemp} - \log P_{Xnrm}}{L} \tag{5}$$

As previously mentioned, the method captures the relatively local deviation of intonation by estimating the probability of emphasis and normal per analysis window (50 msec) and estimates the degree of emphasis from a global point of view per estimating unit (1 sec). The portion, whose degree of emphasis is greater than the threshold, is assumed to be emphasized speech. The aforementioned probabilities must be computed statistically. The learning procedure is described as follows.

Step 1. Emphasized and normal speech portions are extracted manually from sample speech data. The sample data comprises speech from conversations, lectures, meetings, TV programs and movies to enable a wide variety of speech to be handled.

Step 2. Prosodic parameters are calculated and are vector-quantized by the LBG method; the number of clusters is 64.

Step 3. $P_{emp}\,(C_f \mid C_{f-1}C_{f-2})$, $P_{emp}\,(C_f \mid C_{f-1})$, $P_{emp}\,(C_f)$, $P_{nrm}\,(C_f \mid C_{f-1}C_{f-2})$, $P_{nrm}\,(C_f \mid C_{f-1})$ and $P_{nrm}\,(C_f)$ are calculated statistically using measured the vector transitions. λ_{ei} and λ_{ni} (i = 1, 2, 3) are given by the deleted interpolation method (Jelinek & Mercer, 1980).

3.3. Extraction of Speech Blocks

Speech blocks can generally be extracted by using silent portions. However, we should point out that it is difficult to identify silent portions if the contents have background noise. (To identify silent portions, threshold logic of some kind must be applied to the speech waveform. It is, however, quite difficult to set a proper threshold level, if the input speech includes background noise.) Hence, we need to incorporate a method that extracts continuous speech portions robustly. The proposed method utilizes unvoiced information indicated by the pitch detection process instead of identifying silent portions. Consequently, a speech block is defined as a portion bracketed by unvoiced parts whose durations exceed d.

Figure 2 shows an unvoiced portion histogram when the block between the two unvoiced portions corresponds to a semantic sentence. In this study, we define d, the average duration of manually extracted unvoiced portion, as 1 sec. This value of d ensures that speech blocks that exhibit semantic continuity are seldom divided into smaller isolated parts.

3.4. Multimedia Content Skimming

To view video contents effectively, we skim the multimedia content using the degree of emphasis. Figure 3 shows a schematic diagram of the proposed skimming

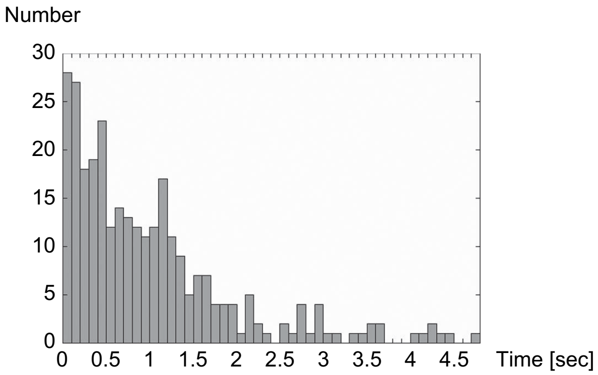


FIGURE 2 Histogram of unvoiced duration.

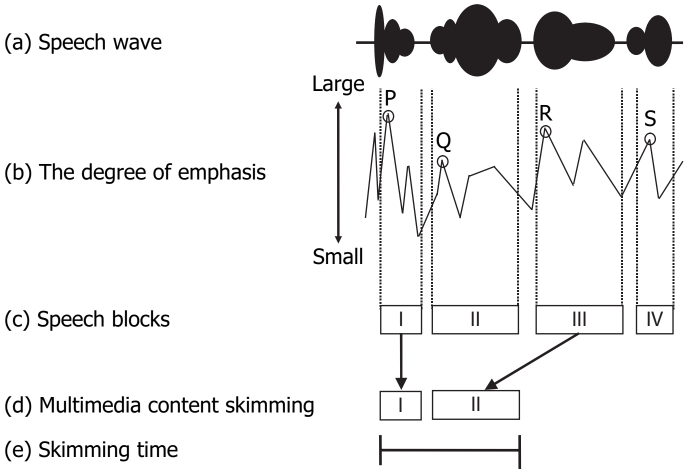


FIGURE 3 The schematic diagram of the proposed skimming method.

method. In accordance with what we stated earlier, (a) speech parameters are calculated by analyzing speech waves, (b) the degree of emphasis is estimated, and (c) speech blocks are extracted. All speech blocks are sorted in descending order of their degrees of emphasis. The degree of emphasis of a speech block is defined by the highest value among the estimating units within it. As stated in the previous section, emphasized portions are relatively short and do not keep semantic continuity. In other words, it is reasonable to assume that the “scope” of emphasis should cover the phrase that keeps semantic continuity, that is, a speech block. Considering this assumption, we define the highest degree of emphasis as being representative for a speech block. As shown in the figure, the degrees of emphasis of speech blocks I, II, III and IV are P, Q, R and S, respectively. Speech blocks are taken from the sorted order just mentioned (Figure 3c) until their total time is just shorter than the user-specified skimming time and then presented to the user in their original temporal order. Figure 3d shows the extraction of two speech blocks,I and III, and Figure 3e shows the resulting skimming time.

4. EVALUATION EXPERIMENTS

4.1. Emphasized Speech Extraction Experiments

Emphasized speech extraction experiments were conducted using both emphasized and normal speech portions. Learning data consisted of 1,110 emphasized and 1,109 normal speech portions, whereas the evaluation data consisted of 599 and 557 portions, respectively.

Speech parameters of learning data were calculated and quantized to make a codebook. From the vectors of each portion, we calculated likelihood values for emphasized and normal by using equations (3) and (4). To compare the two, we estimated the recall ratio and precision ratio for all learning portions. Recall and precision ratio are given by

$$recall = \frac{\{relevant\ \ labels\} \cap \{retrieved\ \ labels\}}{\{relevant\ \ labels\}} \tag{6}$$

$$precision = \frac{\{relevant\ \ labels\} \cap \{retrieved\ \ labels\}}{\{retrieved\ \ labels\}} \tag{7}$$

We define this experiment as the "closed" experiment. Another experiment was carried out using the evaluation data. Vectors of speech parameters were obtained by quantization using the aforementioned codebook. Estimations were made in the same way as in the closed experiment. We define this experiment as the "open" experiment. The results obtained show that recall and precision ratios were about 80% in both experiments (Table 3).

From the standpoint of skimming multimedia content such as video mail, blogs that include audio-visual content, and CGM, it is important to extract emphasized speech robustly. As a measure of extraction accuracy, our proposed method achieved recall and precision ratios of about 80% (Table 3). We note that even the manual extraction of emphasized and normal speech is rather variable, so these results confirm that the proposed method can extract emphasized speech with quite high accuracy.

Table 3: Emphasized Speech Extraction Results

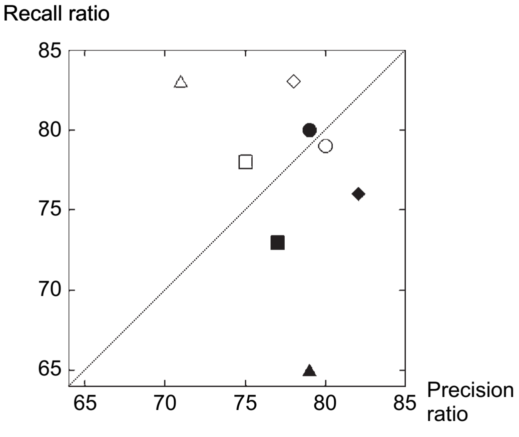
	Emphasized Speech		*Normal Speech*	
	Recall	*Precision*	*Recall*	*Precision*
Closed	81%	81%	81%	81%
Open	80%	79%	79%	80%

Turning now to an account of the speech parameters in this study, we proposed using pitch, power, and speaking rate to extract emphasized speech. After changing the speech parameter set, we performed experiments identical to the one just described. The results obtained are shown in Figure 4. In this figure, the broken line plots identical recall and precision ratios. Emphasized speech extraction is more robust the closer the ratios are to the line. As can be seen in the figure, the extraction performance is closest to the broken line when all parameters are used. It follows that using pitch, power, and speaking rate yield better extraction performance than using only pitch and power, the combination adopted in past studies.

4.2. Assessment of the Skimmed Content Quality

To assess the quality of skimmed content produced by our method, we conducted preference experiments using seven different content items. The content was obtained from 3 participants (2 male, 1 female), and comprised natural, unedited conversation including shouts, cheers, and the like, captured in a noisy environment (e.g., S/N ratio was under 10dB). It is reasonable to assume that such contents might be relevant to video mail, blogs that include video, and CGM (i.e., when users want to send images of a noisy party in progress).

Reference data were created by extracting a set of 10-sec portions at a fixed interval. Ten participants (all female) were asked to select either the skimmed contents created by our method or those created using the fixed interval



Emphasized Speech	Normal Speech	Parameters
●	○	$\overline{F0}, \pm \Delta\overline{F0}, \overline{P}, \pm \Delta\overline{P}, dp$
■	□	$\overline{F0}, \pm \Delta\overline{F0}$
▲	△	$\overline{P}, \pm \Delta\overline{P}$
◆	◇	$\overline{F0}, \pm \Delta\overline{F0}, \overline{P}, \pm \Delta\overline{P}$

FIGURE 4 Result of extracting emphasized speech with various sets of speech parameters.

approach. The skimming rates of 1/10, 1/15, and 1/30 were examined. Both the skimmed contents and skimming rates were selected at random for the evaluations. Table 4 shows the results obtained; overall, our method was selected by 80% of the participants. However, a few exceptional results were found; the results from No. = A-5, skimming rate = 1/30 and No. = A-7, skimming rate = 1/15 were relatively low. In both cases, the proposed skimmed data and reference data happen to be very similar. As a result, the participants could not find any remarkable difference between them. Figure 5 shows the average preference ratio of our method. The results suggest that the proposed method can generate much better skimmed contents than the fixed interval reference data and its effectiveness strengthened as the skimming rate increased.

4.3. Assessment of Skimmed Content Adequacy

We then evaluated the adequacy of the generated skimmed contents. That is, we evaluated how well the skimmed contents reflected the major characteristics of the original contents. We also have an application in mind that provides users with an attractive preview automatically generated by our skimming method.

Table 4: Result of Preference Test

Content No.	*Original Duration*	*Skimming Rate 1/30*	*Skimming Rate 1/15*	*Skimming Rate 1/10*
A-1	600 sec	100%	80%	70%
A-2	600 sec	80%	100%	80%
A-3	900 sec	80%	80%	70%
A-4	900 sec	90%	90%	70%
A-5	500 sec	60%	90%	90%
A-6	1200 sec	100%	90%	70%
A-7	1800 sec	90%	50%	70%

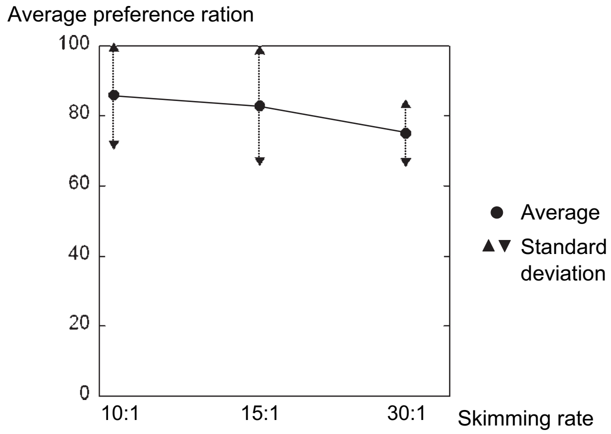


FIGURE 5 Average preference ratio of our method selected by participants.

Taking account of applications of this sort, we evaluated the attractiveness of previews generated by our method. We conducted a subjective evaluation using six content items to verify how well the users could grasp the original atmosphere from the skimmed content and the attractiveness of the skimmed content. The content was obtained from 4 participants (2 female, 2 male), under the same conditions as the preference test. Each content item was about 10 min in length. Three 1-min skimmed contents were created by our method, and three (different) reference contents were prepared in the same manner as described in section 3.1. Figure 6 shows a schematic diagram of the experiment's procedure. The 11 participants (all female) were separated into two groups (α and β). Both groups evaluated three skimmed contents and the three (different) reference contents. Table 5 shows the items evaluated. Evaluation items indicate the prior (S1, S2, and S3) and the posterior (O1 and O2) evaluation. They answered three questions (S1, S2, and S3) after watching the skimmed contents (created by either our method or the

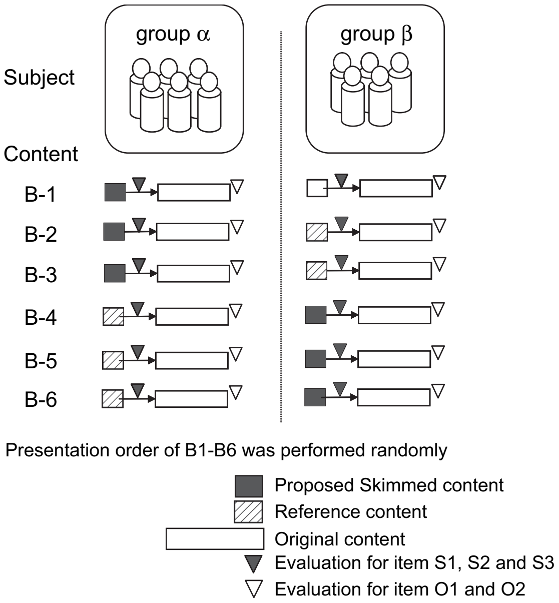

group α
group β
Subject
Content
B-1
B-2
B-3
B-4
B-5
B-6
Presentation order of B1-B6 was performed randomly
Proposed Skimmed content
Reference content
Original content
Evaluation for item S1, S2 and S3
Evaluation for item O1 and O2

FIGURE 6 The schematic diagram of the experiment procedure.

Table 5: Evaluation Items

S1	Did the skimmed content/reference content include a large quantity of emphasized speech?
S2	Would you like to watch the original content?
S3	Could you grasp the characteristics or mood of the original content?
O1	Did the skimmed content/reference content include a large quantity of emphasized speech?
O2	Did the skimmed/reference content accurately reflect the characteristics or mood of the original content?

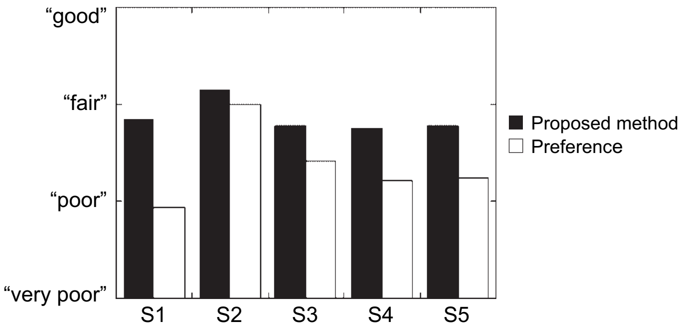


FIGURE 7 The result of assessment of the skimmed content adequacy.

fixed interval method), watched the original contents, and then answered two more questions (O1 and O2). Items S1, S2, and S3, intended for preview applications, evaluated the skimmed content or reference content in terms of their inclusion of emphasized speech and elucidating the original content to the users. O1 and O2, intended for summary applications, evaluated the adequacy of the skimmed content compared to the original. Participants assigned the grades of good, fair, poor, and very poor to the items. Please note that questionnaires S1, S2, and S3 should be presented to the participants before watching the original. This is why we separated the participants into two groups. For both groups, the presentation order of contents was selected randomly.

Figure 7 shows the evaluation results. The proposed method outscored the reference content in every question. We performed a *t* test on the results and the *p* values of S1, S2, S3, O1, and O2 were .000002, .2449, .0324, .0032, and .0009, respectively. Because the *p* values of S1, S3, O1, and O2 were less than .05, we consider that these differences were significant. Its superiority was significant in S1, O1, and O2. The results obtained for S1 and O1 suggest that the proposed method yields skimmed contents that include a comparatively large quantity of emphasized speech. This implies that the method adequately reflects the characteristics and mood of the original content. In the case of multimedia contents such as video mail, blogs that include audio/visual content, and CGM, it is safe to say that being able to quickly view and grasp the mood of the original content is more important than understanding the content semantically. On the other hand, there was a case for S2 where there was no distinction between the proposed method and the reference. We note that the fixed interval method is not such a bad strategy for making a summary content and it reflects the time structure of the original as much as possible by definition. In terms of creating the desire to watch the original, we have to admit that there is no drastic difference between the previews made by our method and the fixed interval one.

5. CONCLUSIONS

In this article we have proposed a new multimedia content skimming method based on the extraction of emphasized speech. In our method, speech parameters associated with prosody are extracted and the degree of emphasis is estimated according to likelihoods of being emphasized and normal speech; these are calculated from the transition probabilities of speech parameter vectors. To obtain skimming relevant to semantic continuity, speech blocks are extracted, and those having a high degree of emphasis are given skimming priority. The speech blocks with higher priority are combined so as to satisfy the user-specified skimming rate.

Emphasized speech extraction experiments indicated that the proposal could extract emphasized speech with high accuracy, as both the recall and precision ratios were about 80%. However, emphasized speech also involves such factors as pause duration and "clearness of pronunciation." In our future work, we plan to address these topics.

In evaluating the quality of the proposed skimming method, preference experiments showed that about 80% of the test participants selected our method over one that uses the fixed interval approach and that its superiority increases with the skimming rate. This indicates that our method is a more effective approach to skimming. We also evaluated the adequacy of the generated skimmed content via subjective experiments; the results showed that the proposal yielded skimmed contents that well reflected the characteristics and mood of the original content. The proposed method was confirmed to be quite effective for multimedia contents with audio data such as audio/video mail, home video, and any CGM contents. In future work, we plan to perform another preference test involving the use of manually summarized content.

The proposed approach utilizes prosodic information and does not refer to any linguistic information. Because prosodic emphasizing seems to be not so language specific, it is possible that the proposed method can be easily extended to multilingual content skimming. This extension is also left as future work.

REFERENCES

Aoyagi, S., Sato, T., Takada, T., Sugawara, T., & Onai, R. (2005). Evaluation of video-skimming method to educational purpose movies. *IPSJ Journal*, *46*, 1297–1305. Evaluation of Video Skimming.

Bogert, B. P., Healy, M. J. R., & Tukey, J. W. (1963). The Quefrency analysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, CrossCepstrum, and Saphe Cracking. In M. Rosenblatt (Ed.), *Proceedings of the Symposium on Time Series Analysis* (pp. 209—243). New York: Wiley and Sons.

Chen, F. R., & Withgott, M. (1992). The use of emphasis to automatically summarize a spoken discourse. *Proceedings of the IEEE International Conference on Acoustic Speech Signal Process*, 229–232.

He, L., Sanocki, L., Gupta, A., & Grudin, J. (1999). Auto-summarization of audio-video presentations. *Proceedings of the ACM Multimedia 1999*, 489–498.

Hidaka, H., Mizuno, O., & Nakajima, S. (2002). A new speech content summarization technique based on speech emphasis extraction. *Proceedings of the ASJ Fall Conference 2002*, 99–100.

Hori, C., & Furui, S. (2003). A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, *5*, 368–378.

Jelinek, F., & Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In E. S. Gelsema & L. N. Kanal (Eds.), *Pattern recognition in practice* (pp. 381–397). Amsterdam: Elsevier.

Lienhart, R., Pfeiffer, S., & Effelsberg, W. (1997). Video abstracting. *Communication of the ACM*, *40*, 55–62.

Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communication*, *28*, 84–95.

Moriyama, T., & Sakauchi, M. (2001). Video summarization based on the psychological unfolding of a drama. *Proceedings of the IEEE International Conference on Consumer Electronics 2001* (J84-D-II), 1122–1131.

Rabiner, L., & Juang, B. H. (1993). *Fundamental of speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Sagayama, S., & Itakura, F. (1979). On individuality in a dynamic measure of speech. *Proceedings of the ASJ Spring Conference 1979*, 589–590.