1994 International Symposium on Speech, Image Processing and Neural Networks, 13-16 April 1994, Hong Kong

Voice Message Summary for Voice Services

Yasuhisa Kato

NTT Network Information Systems Laboratories

Abstract

This paper describes a voice message summarizing method for retrieving specific voice messages from a large number of voice messages on voice services, such as voice mails and voice bulletin boards. Voice browsing facilities are tools intended to allow users to handle voice messages as easily and conveniently as browsing books. After surveying methods for voice browsing, I propose a new voice message summarizing method that is based on the important part being spoken slowly and having a higher proportion of unvoiced parts. The effectiveness of this method was demonstrated using actual voices from radio programs.

1 Introduction

The number of voice message services is increasing year by year. In Japan two types of voice storage services are offered by NTT. One is a public voice mail box, called "Dengon-Dial" and the other is a subscription group voice mail box, called "Message-In." Since the market for voice message services is growing, we are developing new techniques for voice messaging service facilities.

The voice bulletin board service (VBBS), one such service, is the counterpart of text-based bulletin boards. Both bulletin boards will be integrated to the multimedia bulletin boards together with the still picture and the animation in the future. In multimedia environment, voice processing techniques for manipulating voice messages are still important.

A VBBS consists of several structured bulletin boards. VBBS users can listen to and record spoken messages on any bulletin board using ordinary telephone sets. To implement VBBS, there are some obstacles for users and information providers that need to be solved. It takes much time for users to listen to all the messages on a particular bulletin board if it has many messages. Some solutions to this include:

(a) Structured voice messages: Voice messages must conform to a predetermined structure, such as subject, sender name, keyword, and body [1, 2].

(b) Message browsing (skimming): This is analogous to the way we read.

One of the most important human interface issues for these services is to establish "message browsing" facilities that allow users to quickly search for and retrieve the messages they want.

We surveyed several current methods of "message browsing," proposed a new voice message summarizing method, and demonstrated its performance.

2 Survey of voice browsing methods

There are many ways to provide voice browsing facilities [3,4]. In this paper I present some simple compression and extracting methods for the voice browsing.

2.1 Compression

A *compression method* divides a voice message into scores of millisecond frames. Then it links these frames, with overlapping according to the playback speed. The overlapped part will increase as the speed increases (Figure 1). It is called the synchronized overlap and add method (SOLA) [5]. Other modified SOLA methods were presented [6, 7].

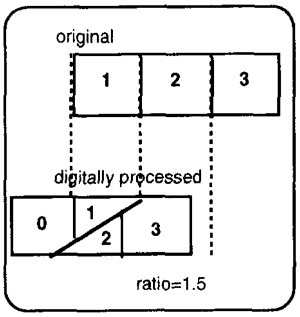

Figure 1: SOLA (Synchronized OverLap Add)

0-7803-1865-X/94/$3.00 © 1994 IEEE

The advantage of this method is that not only speech but also any sound source can be compressed and the computation time is relatively short, compared with pitch depending methods. However, there are limitations on the speed. Our experimental results showed that users cannot understand the whole content when the speed is over three times faster than the original [8, 9].

2.2 Extraction

An extraction method does not compress the whole message. Some parts of the messages are extracted and played back. I present three extraction methods. First, a message is divided into phrases by detecting silences. Then a certain length of the initial part of each phrase is played back at the selected playback speed in Figure 2 [9]. Second, the mean power in each frame is calculated, and the region around frames whose power exceeds a given threshold is played back in Figure 3 [9]. Third, the emphasized part is extracted by using a Hidden Markov Model [10].

These methods are not suitable for understanding the entire content, but give a good grasp of the outline of the message. Our experiments for voice services showed that these methods decreased listening time about a factor of five [8, 9].

3 A new method

I propose a new extraction method that has two steps.

First, a voice message is divided into parts by detecting silent periods and counting zero-crossings [11]. The parts are then adjusted so that they correspond to text-based

(1) words,
(2) phrases,
(3) sentences,

as much as possible (Figure 4). They are classified according to the following algorithm. This algorithm is very simple. Two thresholds of the length of the silence are decided. Dividing into words are based on detecting silent periods and counting zero-crossings. Dividing into phrases binds two voiced parts based on words when the length of the unvoiced part between them is less than the threshold A. Dividing into sentences does as the same as dividing into phrases except using the threshold B. Each unit (word or phrase or sentence) has voiced parts and unvoiced.

Second, the units that have a higher proportion of unvoiced parts are extracted and played back. The number of extracted units is based on the extraction ratio, which is usually less than 1/2. Therefore, the speed is more than twice.

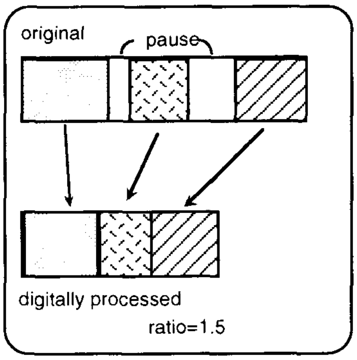

Figure 2: EIPP (Extracting the Initial Part of each Phrase)

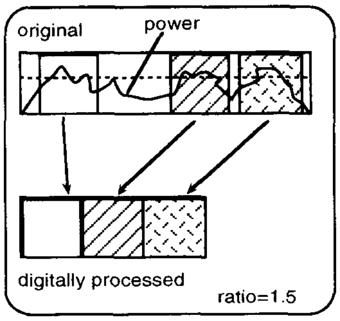

Figure 3: EHPF (Extracting High-Power Frames)

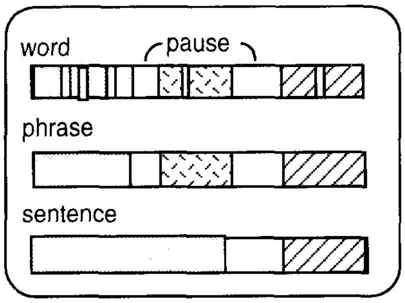

Figure 4: Message dividing

This method depends on the following two hypotheses.

(1) Important parts are spoken more slowly than usual.
(2) When a voice message is divided into some units, the proportion of unvoiced parts are higher if a message is spoken more slowly.

So, the important part has a higher rate of the unvoiced. The voice summary can be implemented automatically to select the units whose proportion of the unvoiced parts is high.

4 Testing the hypotheses

I tested two hypotheses using actual voices spoken by radio announcers on the radio programs and questionnaires completed by test subjects.

4.1 Speech characteristics

Speech samples were divided into units and the speed of each unit (word, phrase, sentence) was measured. The speed was expressed in characters per minute (CPM). I used ten sample messages spoken by radio announcers. The conditions of the messages are showed in Table 1.

Correlations between the speech speed and the ratios of the unvoiced part are shown in Figure 5.

From Figure 5 there is a minus-correlation between speech speed and the proportion of the unvoiced part. The minus correlation means that the speech speed is increasing as the proportion of the unvoiced part is decreasing. Therefore when the part whose unvoiced proportion is high is picked out, there is a high probability of having a slow speech speed. Though there is a correlation between them, it's not so strong. There is a dispersion of messages and how to divide. Sentences have stronger correlation than words.

4.2 Subjective evaluation

The parts that speakers thought them important were determined by questionnaires. Ten sample voice messages were written on the paper. Test subjects picked up the important part from reading the messages on the paper freely. Figure 6 shows the result. The real line shows the mean CPM of the important parts which test subjects selected, and the broken line shows the CPM of the unimportant.

Seven of ten messages have the slower CPM of the important part and three have the faster.

5 Discussion

Though there would be a correlation between speech speed and the ratio of the unvoiced parts, it is not so clear and steady. To apply voice browsing for voice services, the function of exact word search is not needed for voice summary now. And it is very difficult to create an exact summary even if the real voice recognition is available.

Table 1: Conditions of voice messages

Sampling frequency	8 kHz
Quantizing	1 bit sign + 13 bit linearPCM
Message	10 news
Mean length	about 60 seconds
Mean speech speed	453 CPM

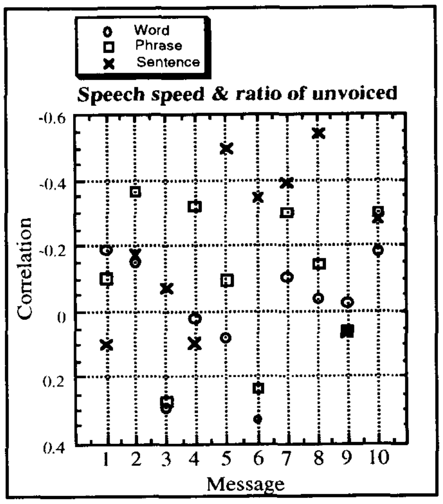

Figure 5: Speech speed and the ration of unvoiced part

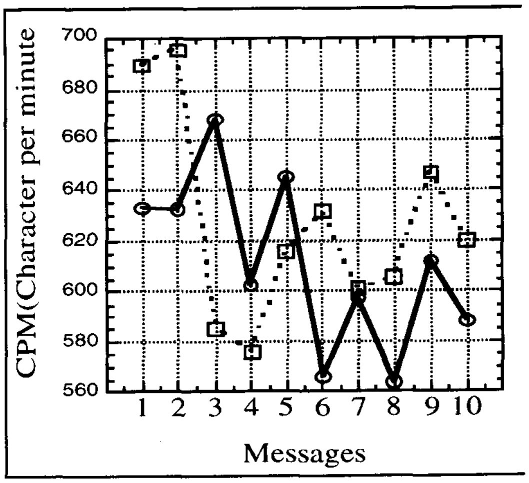

Figure 6: CPM (important & unimportant)

Since in this subjective evaluation test subjects had no restrictions with selecting the important parts, there is a variation to selecting. Someone selected a few words from one message, and another selected many words. I found the difference of what part people think it important.

From these results this algorithm is very useful for the voice browsing to summarize voice message and to save the listening and searching time. Since single algorithm or method is not sufficient enough, it is necessary to properly combine other methods including the skip method.

However, I think more analysis is necessary to confirm the two hypotheses and the real situation tests should be done that real users record messages and they listen to them.

6 Future work

These results involve several problems. In this paper, I used announcers' speech; use of the speech of professional may affect the results. I will examine more voice messages spoken by ordinary users.

The method will be evaluated by experiments, involving a comprehension test as same as the real situation. Subjects listen to messages extracted at various ratios. After listening to each extracted message they write down a summary of the message. The performance will be compared to other extracting methods.

Other problems are posed by the experience of users. People are accustomed to the fast-speed speech and the extraction speech. How are they accustomed to ?

All evaluation experiments were made in Japanese in this paper. I'd like to try the experiments in English and other languages.

The length of pause between extracted units has an important role in the extraction method. I'll determine how long and where the pause is inserted.

I'll improve the method and combine the extracting method and compressing methods and other methods. This integration will show the high performance and the evaluation test should be done under real conditions.

In the future, I will improve our method to make it applicable to actual voice services and build integrated system together with voice recognition technique.

7 Conclusion

We surveyed several methods for message browsing. A new voice message summarizing method was proposed and evaluated. The effectiveness of the method was confirmed for the speech of professional radio announcers. The capability and problems were discussed.

Acknowledgments

I thank Hideyuki Koike for his guidance and encouragement. Thanks for our research colleagues and laboratory members who helped with much of the experimental and construction work.

References

[1] C. Schmandt & B. Arons (1984). A Conversational Telephone Messaging System. *IEEE Transactions on Consumer Electronics*, CE-30(3):xxi-xxiv.

[2] P. Resnick & R.A. Virzi (1992). Skip and Scan: Cleaning Up Telephone Interfaces. *Proceedings of CHI'92*, 419-426.

[3] B. Arons (1992). Techniques, Perception, and Applications of Time-Compressed Speech. *Proceedings of Voice I/O Systems Applications Conference*, 169-177.

[4] B. Arons (1993). SpeechSkimmer: Interactively Skimming Recorded Speech. *ACM Symposium on User Interface Software and Technology*.

[5] S. Roucos, A. M. Wilgus (1985). High quality time-scale modification for speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:493-496.

[6] R. Suzuki & M. Misaki (1990). An implementation of a time-scale modification method on a dsp (in Japanese). *IEICE Technical report Speech*, SP90-34:1-8.

[7] M. Misaki, R. Suzuki & H. Naono (1989). A study on time scale modification (in Japanese). *IEICE Technical report Engineering Acoustics*, EA89-94:17-24.

[8] Y. Kato & K. Hosoya (1992). Fast message searching method for voice mail service and voice bulletin board service. *Proceedings of Voice I/O Systems Applications Conference*, 215-222.

[9] Y. Kato & K. Hosoya (1993). Message Browsing Facility for Voice Bulletin Board Service. 14th International Symposium Human Factors in Telecommunications, 321-330.

[10] F. R. Chen & M. Withgott (1993). The Use of Emphasis to Automatically Summarize a Spoken Discourse. *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 1:229-232.

[11] Y. Yatsuzuka (1980). A High Sensitive Speech Detector Based on Sign Bit Sequence Manipulations (in Japanese). *The Transactions of the Institute of Electrics, Information and Communication Engineers*, J64-A(7):413-420.