MSF FORMAT FOR THE REPRESENTATION OF SPEECH SYNCHRONIZED MOVING IMAGE

Cheol-Woo Jo

Department of Control and Instrumentation Eng., Changwon National University

Changwon, Kyeongnam 641-773 Korea

Phone: +82-551-279-7552, Fax: +82-551-262-5064, Email: cwjo@sarim.changwon.ac.kr

ABSTRACT

This paper describes the structure of a new multimedia file format. Also the procedures for implementing its encoder and the player are described. Multimedia Sound File(MSF) format reduced the size of the file. The display software is improved in the points that it requires only small sized image database compared to that of current similar programs require hugh amount of image database. This software tool can effectively display animated facial images and speech sounds together in synchronized form even at PC level. Implemented tool can be used as a plugin or an independent form. Encoder software is implemented to facilitate the production of the msf file. Files from the segmentation of speech signal into phonemic units are used as an input to the encoder.

1. INTRODUCTION

Recently many different kinds of multimedia file formats are being proposed. Such multimedia files can deliver informations more effectively than sound only or image only files. Some of the multimedia files include complex encoding mechanisms.

To output speech signal with image informations, various researches are being reported. 2D and 3D images are used to express speech signal in multimedia form. Recently 3D talking head models are proposed by many research groups. But most 3D implementations used mesh frames of the face and computed positions of the every nodes to decide proper lip shapes and facial expressions corresponding to the contents of the speech signal. Accordingly much computing time is required to generate moving image sequences and big computing power is essential for implementation. It requires at least workstation level computers or high performance PCs. It also requires large image database to store real images when real image was used to be shown.

In this paper we proposes a simple multimedia file format, which can be used for various purposes. And we also describes it's encoder and player softwares. We named that format as Multimedia Sound(Speech) File(MSF).

2. MSF FORMAT

There are various kinds of currently available moving image formats. Some of the formats can include sound and some does not. For example formats such as avi and mpeg can include sound in itself, but gif format cannot include sound in it. When considering expression of speech synchronized moving image, we can use avi or mpeg. They are now widely used as standard. But there are some problems in practical point of view when considering specific applications such as TTS output or some web-based applications. Even though those are being used effectively in current applications, the size of the files in those format is still too big and takes much time to transmit and requires much space on harddisk. They are mainly caused by the size of the included images. Furthermore their coding process to reduce the volume of the image data is too complex to perform. When considering some special application areas such as web-based applications or TTS output etc, current formats are definitely not so economical. In the web-based applications, we would better use only limited number of images shots to make movie according to the contents of sound or speech. In TTS , if we want to express moving facial images, mpeg coding requires too much operations from the processor. The images required to express the movement of the face are limited to facial components of a character. In those cases we can reduce the size of moving image files by removing the image shots from the files and keeping it in the client's memory.

The main motivation of this paper came out of this idea. By storing some sets of images on client's memory before using, we can reduce the time required for downloading multimedia files and the size of multimedia file on server. The idea is very simple. By replacing the image files of current multimedia files into index numbers of image data set, we can reduce the size of the file enormously. Users can choose his or her own image sets according to their applications. Asynchronous output mechanism is used to display sound and image together in synchronized form.

Figure 1: Overall structure of MSF format

Figure1 shows the overall structure of the MSF. In header part of MSF, the following information is recorded.

- identification number of MSF
- total number of Images Index included in the file
- image set ID

Image index consists of each component image numbers from image database for single time slot. Any number of images can be included by attaching image index components after the file

header. But currently the consecutive numbers of images in a file are limited to 65536 because of the size of the variable(word). This is thought to be enough when considering the characteristics of TTS applications, which generate informations discretely. Sound file comes to the tail part of the file. Normal WAV format file is attached to the format directly without modification.

Currently the image index consists of general index number and index numbers for component images such as face, eye, nose, mouth, etc. At this stage our main areas of application is TTS, accordingly facial components are included in the header part. In general case face specific components can be ignored. I.e. only general index number can be used.

Figure1 and Figure2 shows the difference in an example of usage of conventional multimedia formats and msf between client and server environment.

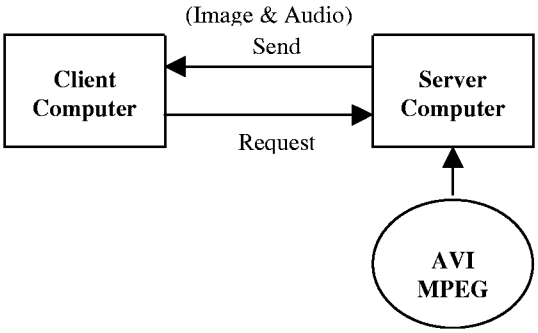

Figure 2: Usage of current moving image files in client-server environment

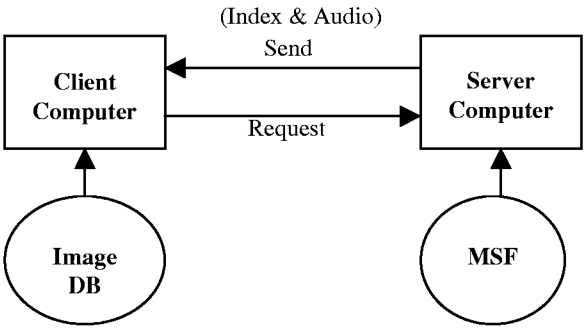

Figure 3: Usage of msf in client-server environment

3. CONSTRUCTION OF IMAGE DB

Unlike conventional multimedia files MSF does not contain image data in itself. Image database is prestored in the computer which contains display program. For facial image database shapes of face, lips, eyes etc are measured and drawn based on real pictures. The size of one image is 141 by 141. Total size of the DB is 1.2MB. So it is very small compared to other implementations.

Cartoon image is created for describing the details of face. Real picture images or mesh structure images doesn't look so natural than expected when imaplemented and requires more computing. To be displayed easily on normal PCs(not a high performance PCs), we used 2D cartoon images and limited its size. Another reason that cartoon images are used is sometimes over emphasized images are more efficient to give the meaning to users. Figure4 shows the real image sample collect. Figure5 shows cartoon character images implemented.

Table1 is matching table between phonemes and corresponding images.

Phonemes	Image(Small)	Image(Large)
/a/	M11	M12
/eo/	M21	M22
/o/	M31	M32
/u/	M41	M42
/ae/	M51	M52
/e/	M61	M62
/i/	M71	M72
/eu/	M81	M82
/w/	M31,M41	
/j/	M71,M81	
/m/,/b/,/p/,/pp/	J11	J12
/ng/,/h/	M21	M22
/g/,/gg/,/n/, /d/,/dd/,/r/, /s/,/ss/,/z/,/zz/, /ch/,/k/,/t/,/tt/	J31	J32

Table 1: Phonemes vs corresponding images

Other image files can be added to the database. Any kind of images can be designed to express visual part of the sound.

4. ENCODER AND PLAYER

To facilitate generation of MSF and display, sample encoder and player software is written.

4.1. Encoder

Encoder consists of two parts, i.e. segmentation part and msf generator. Segmentation part can be implemented by automatic segmenter or manual segmenter. Text or phonemic information is provided as an input. Segmentation part is just a conventional labelling program. Figure4 shows the screen from

the segmentater. This is an example of manual segmenter. Automatic segmenter is undevelopment.

Each phoneme bounday is decided by referencing spectrogram, evergy and zero crossing rate. The resulting timing information is stored to the form of timing file. Time and each phoneme number is written in the file sequencially. The current extension of timing file is 'tdd'. 'tdd' stands for time delay data.

Using 'tdd' file msf file is generated by generator. The procedures are as follows.

- Read 'wav' file
- Read 'tdd' file
- Generate 'msf' file and store it with specified name

Figure4 shows the structure of encoder.

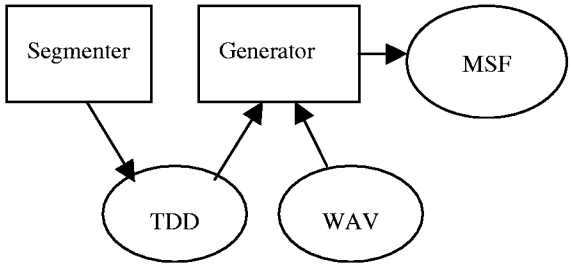

Figure 5: Structure of encoder

4.1. Player

Player program is written to display 'msf' file in synchronous manner. By using style sheets we can make the Proceedings more attractive and useful to everyone. Using styles can insulate you from the tedious repetition of setting font and spacing parameters for each paragraph individually. Taking a few minutes to set up a style sheet can save hours of work later. For instance, by setting "before" and "after" spacings for paragraph and heading styles you can do without extra empty paragraphs to get the white space between paragraphs. But be careful with "before" and "after" spacing in adjacent paragraphs—some word processors add the two spacing quantities together, leaving too much white space between paragraphs.

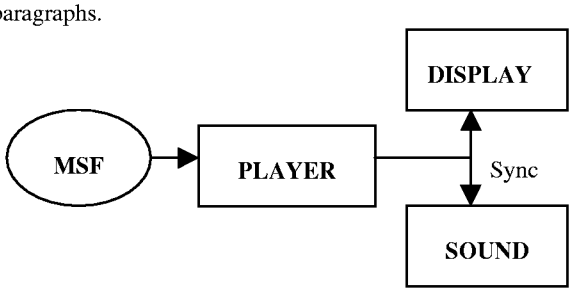

Figure 6: Function of player

5. Applications of MSF

There are many possible areas of applications of msf format.

- Adding sequential image files to the wave file to convert it in multimedia form.

  Predefined various image sets can be used to add moving images or sequential images. Automatic indention of proper images according to the emotional characteristics of audio signal.

- Output module of TTS

  TTS has its timing information inherently. Using those informations it is easy to generate 'msf' file.

- Plugin output module for webbrowser

  Msf format can be used to store multimedia informations with reduced size.

- Independent MSF player

  MSF file is small in its size and can be used to store various sound files with attached image. This format can be applicable not only to speech but also to song, music etc.

6. CONCLUSION

We proposed a new MSF file format for expressing speech with synchronized moving image. And procedures to implement encoder and player is described. Msf format is useful for applications which require only some limited image sets. The performance of displaying and storing multimedia file will be greatly improved using this format.

Further research topics relating this research are designing a more efficient encoding tools so that this format can be widely used by public. Applying automatic segmentation technology for encoding speech file is also under consideration.

ACKNOWLEDGEMENT

This paper is a partial result from the research funded by Korea Science and Engineering Foundation. Grant no.971-0917-104-2 The author wish to acknowledge the financial support of the Korea Science and Engineering Foundation made in the program year of 1997.

7. REFERENCES

1. Shigeo Morishima, Hiroshi Harashima, "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface", IEEE Journal on Selected Areas in Communications, pp.594-600, Vol.9, No.4, May 1991

2. W.Goldenthal, K.Waters, J-M Van Thong, O.Glickman, "Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe!", Proc. Eurospeech'97, Rhodes, Vol.4, pp.1995-1998, 1997

3. F.Lavagetto, "A New Algorithm for Visual Synthesis of Speech", Proc. Eurospeech'95, Madrid, pp.303-306, 1995

4. Jonas Beskow, "Rulebased Visual Speech Synthesis", Proc. Eurospeech'95, Madrid, pp.299-302, 1995

5. Cheol-Woo Jo, In-Hwa Chung, "Implementation of Multimedia Speech Player Using Animated Cartoon Image", Trans. On Institute of Information, Vol.2, pp.147-150, 1998

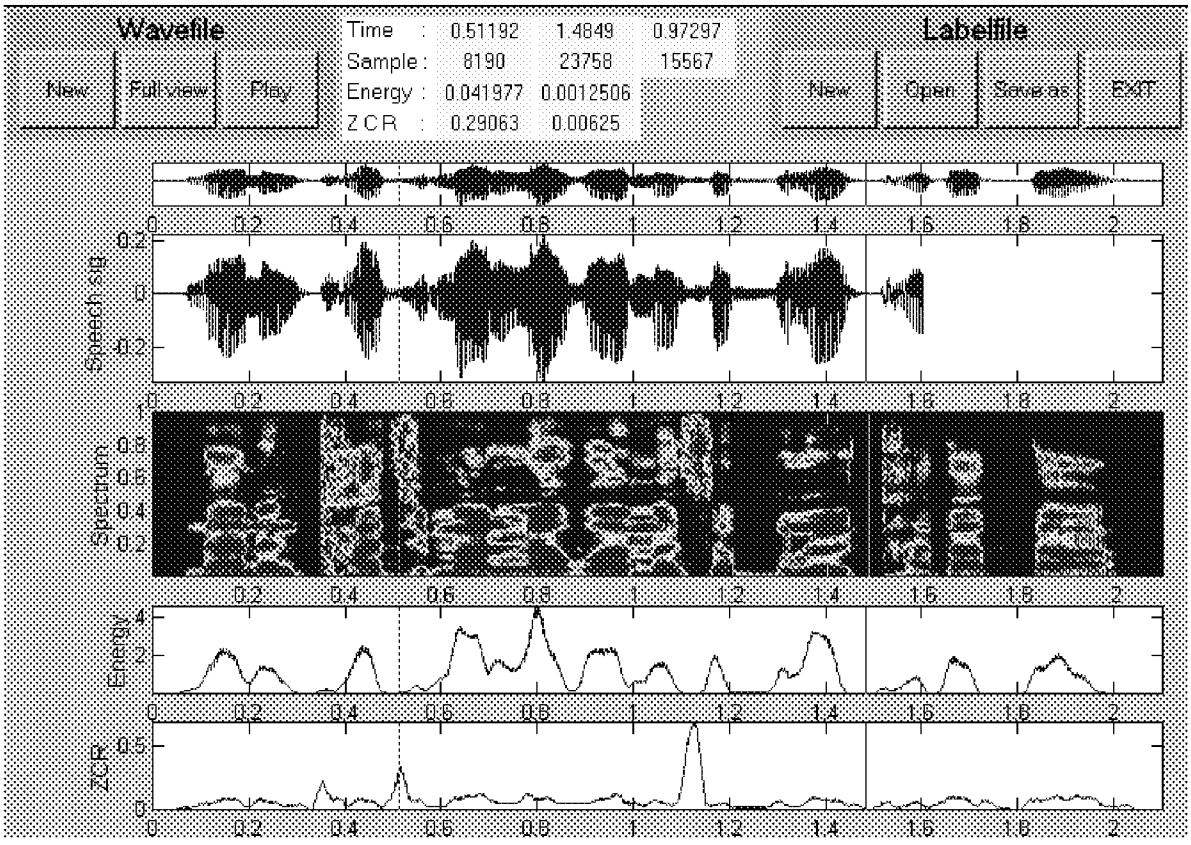

Figure 4: Segmentation tool