Automatic Video Caption Generation for Sound Containing Voice and Music

Takafusa Kitazume, Masahide Sugiyama

The University of Aizu

Abstract. This paper describes our progress of automatic caption generation project. VCML project has developed video caption markup language and its player (VCML and VCML Player) to reduce labor and cost of making captions. Voice-Pause method[1], which was originally developed to align between voice intervals and their corresponding written text, is improved to align sound data containing both voice and music intervals. The results of the alignment experiment show that the improved method, Voice-Music-Pause method, can align both voice, music and pause intervals effectively.

1 Introduction

Today, 350,000 Japanese hearing-impaired people are recognized as physically disabled persons by the Ministry of Health and Welfare. Considering hearing-impaired seniors the number increases to 6 million. Only 17.1% hearing-impaired people can use sign language[2] and most of them have difficulty to get information by sign language. Therefore, broadcasting with captions is very important for providing information to them.

Our VCML project has developed video caption markup language (VCML) and its player (VCML Player) to solve the above problem, and Voice-Pause method has been proposed and the effectiveness was reported[1]. This paper proposes a new alignment method for sound data containing both voice and music intervals[3].

2 VCML system

VCML (Video Caption Markup Language) is the original script language for description video caption based on the XML (eXtensible Markup Language). VCML document is created by the flow of the process shown in Fig.1. The time alignment module needs a scenario and the sounds in the video. The VCML Player can display the video and captions on a PC screen in real-time. VCML has several TAGs to describe information about displaying time and objects. VCML Player generates caption images from a VCML document, compounds new images from video and caption images and displays them.

In alignment module, Voice-Pause method supposes that the speaker alternately transit two states: “voice state” and “pause state”. Using DP matching,

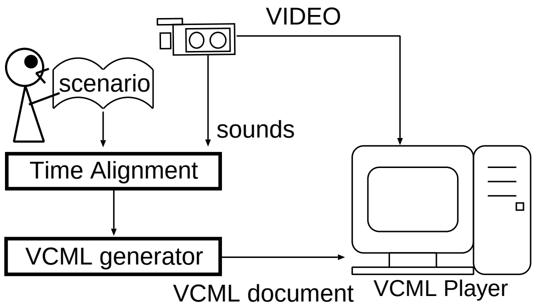

Fig. 1. VCML project

the correspondence between the input speech data and the states is determined. The local distance $d_{i,j}$ between the input speech feature vector of a_i and the phoneme feature vector in the state of j at the lattice point (i,j) is defined as Eq.(1).

$$d_{i,j} = \begin{cases} \min_{1\le p\le N}\ d^2_{\mathrm{CEP}}(a_i, v^1_p)\ (j:\mathrm{Voice}) \\ \qquad\qquad d^2_{\mathrm{CEP}}(a_i, v^2)\ (j:\mathrm{Pause}) \end{cases} \tag{1}$$

where $v^1_p (1 \le n \le N)$ is the voice feature vectors and v^2 is the pause feature vector automatically created using the Otsu method for given voice database. The number of the voice feature vector, N, is 38. d_{CEP} is power weighted LPC cepstrum distance between two vectors. $g_{i,j}$ is the minimum value of distance accumulations from the starting point $(0,0)$ to lattice point (i,j). It is recursively calculated as follow:

$$g_{i,j} = \begin{cases} \min_{0\le k\le \lceil w_l l_j \rceil}(D_{i,j,k} + w_L\ L_{j,k} + g_{i-k-1,j-1}) \\ +\infty \qquad\qquad\qquad\qquad\qquad\qquad (i,j<0) \end{cases} \tag{2}$$

Here, $D_{i,j,k}$ is the accumulation of the local distance from a lattice point $(i-k,j)$ to (i,j). $L_{j,k}$ is the difference in the frame length from estimated duration of state j , when the state j continues k frames. l_j is an estimated frame length of state j. w_l is a weight parameter for controlling the duration of all states. $D_{i,j,k}$ and $L_{j,k}$ are defined as follows:

$$D_{i,j,k} = \sum_{n=0}^{k} d_{i-n,j}, \quad L_{j,k} = (l_j - (k+1))^2 \tag{3}$$

where i is the number of input voice frames. After the calculation of $g_{i,j}$, back tracing from the end point $(I-1, J-1)$ to $(0,0)$ provides the best cost search path.

l_j can be given from the number of characters using $y = ax + b$. The reading units (voice and pause) are decided considering the meaning of the contents and listeners. In order to divide the text into reading unit, the 3 rules are applied.

3 Evaluation using Sound Containing Music

3.1 Database Containing Voice and Music

Using recitation database (database A) without music intervals alignment module was evaluated[1]. The experiment results showed that the average time gap is 0.108 sec and the Voice-Pause method works well.

Database B which is recitation of a fairy-tale story is used for experiments. The total length is 31 minutes. This contains female voice, instrumental music played using violin, viola, cello, piano and synthesizer. The data are divided to 15 parts separated by music sections which contain 10 seconds or more music.

As voice feature vectors a set of Japanese phoneme feature vectors[1] is generated using ATR phonetically balanced 216 isolated words; four females uttered 54 words without overlap. As the feature analysis the standard LPC cepstrum analysis is applied. Time gap between boundaries of reading units and label data is calculated. Here, the label data is given manually.

3.2 Evaluation of Voice-Pause Method

Using the first half 53 reading units in database B, the duration of voice state l_j (j:voice) in Eq.(2) is estimated by one dimensional polynomial and the optimal value $a = 0.19, b = -0.09$ in $y = ax + b$ are given. Speech rate in database B is slightly slower than that of database A. Using the estimated duration for l_j and $w_p = 50.0$, $w_L = 0.40$ of Eq.(2) for the database B, the average of time gap is 2.658 sec and using the estimation from database A the average time gap is 4.553sec. Fig.2 shows the time gap and the number of reading units in each section. The section 2 and 14 have the large time gap.

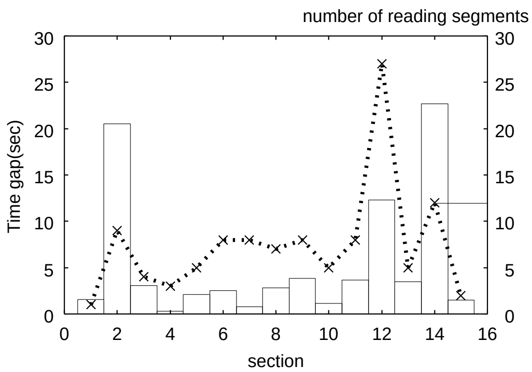

Fig. 2. Time gap of each reading unit

3.3 Evaluation of Voice-Music-Pause Method

The Voice-Pause method supposes that the speaker alternately transits two states. Voice-Music-Pause method, supposes that the data transits three states: voice, music and pause states. Eq.(1) is replaced as follow:

$$d_{i,j} = \begin{cases} \min_{1\leq p\leq N} \ d^2_{\mathrm{CEP}}(a_i, v^1_p) \ (j : \mathrm{Voice}) \\ \qquad\qquad d^2_{\mathrm{CEP}}(a_i, v^2) \ (j : \mathrm{Pause}) \\ \min_{1\leq p\leq M} \ d^2_{\mathrm{CEP}}(a_i, v^3_p) \ (j : \mathrm{Music}) \end{cases} \tag{4}$$

Voice and Music states are specified by symbols of music section in a scenario file. When $j = \mathrm{Music}$, music feature vectors are used. The VQ codebook is generated from input music intervals based on LBG algorithm. It is not easy to estimate duration length of music intervals, therefore, Eq.(2) is defined as follows:

$$g_{i,j} = \begin{cases} \min_{0\leq k\leq\lceil w_l l_j\rceil}(D_{i,j,k} + w_L \ L_{j,k} + g_{i-k-1,j-1}) & (j : \mathrm{Voice}) \\ \min_{0\leq k\leq\lceil w_l l_j\rceil}(D_{i,j,k} + g_{i-k-1,j-1}) & (j : \mathrm{Music}) \\ \qquad +\infty & (i, j < 0) \end{cases} \tag{5}$$

As the previous experiment in section 3.2 suggests that the method does not work for voice with BGM, in this experiments sound data without BGM is only used. This subset data corresponds to 6-9th and 15th sections, and five music sections. A violin, viola and cello are played in the four music sections as the foreground music, and a piano and synthesizer are played in the last section.

A set of music feature vectors with 32 vectors was generated. Using l_j of database B and same w_p, w_L in 3.2, the time gap is 0.769 sec and the proposed Voice-Music-Pause method works well.

4 Conclusion

This paper proposed Voice-Music-Pause Method to align text and sound data containing voice and music intervals. The experimental result showed that the proposed method works well for sound data where voice and music intervals are not overlapped.

References

1. T. Suzuki, et al, Enhancement of VCML Player, Proc. of MMSP2001, pp.365-370 (Oct. 2001).
2. Ministry of Health, "The outline of survey for handicapped person and a physical disability child.", (Jan. 1999) (in Japanese). http://www1.mhlw.go.jp/toukei/h8sinsyou_9/index.html,
3. T. Kitazume, Automatic Generation of VCML Documents, Graduation Thesis of The Univ.of Aizu, UGT01-1061061 (Feb. 2001).