Recognition of Hands-free Speech and Hand Pointing Action for Conversational TV

Yasuo Ariki
Dept. of Computer and
System Engineering
Kobe University
1-1 Rokkodai, Kobe,
657-8501, Japan
ariki@kobe-u.ac.jp

Tetsuya Takiguchi
Dept. of Computer and
System Engineering
Kobe University
1-1 Rokkodai, Kobe,
657-8501, Japan
takigu@kobe-u.ac.jp

Atsushi Sako
Dept. of Computer and
System Engineering
Kobe University
1-1 Rokkodai, Kobe,
657-8501, Japan
sakoats@me.cs.scitec.kobe-u.ac.jp

ABSTRACT

In this paper, we propose a structure and components of a conversational television set(TV) to which we can ask anything on the broadcasted contents and receive the interesting information from the TV. The conversational TV is composed of two types of processing; back end processing and front end processing. In the back end processing, broadcasted contents are analyzed using speech and video recognition techniques and both of the meta data and the structure are extracted. In the front end processing, human speech and hand action are recognized to understand the user intention. We show some applications, being developed in this conversational TV with multi-modal interactions, such as word explanation, human information retrieval, event retrieval in soccer and baseball video games with contextual awareness.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*user-centered design, voice I/O*

General Terms

Design, Experimentation

Keywords

Hands-free speech recognition, hand pointing action recognition, conversational TV

1. INTRODUCTION

Digital contents are widely broadcasted to home television sets all over the world. However, they have no function to answer the questions such as "who is he?", "what is this?" or "what is the meaning of the word?" about the objects appearing in the contents or words spoken by the anchor person in the news program. Consequently, we can not get the information exactly we want to know from the present television interactively. In order to solve this problem, we have to provide the television with an intelligent facility of multi-modal interaction as well as broadcasted content analysis.

As the first step to realize this facility, we propose in this paper a structure and components of a conversational television set, especially focusing on the recognition ability of hands-free speech and hand pointing action. When we ask some questions to the television, a keyboard or mouse is not suitable because we want to focus our attention on the content when watching the television. Then speech and hand action are employed as excellent modalities compared to the keyboard and mouse in this study. However, grasping a microphone or wearing a handset, putting LEDs or color markers on the hands and face are extremely cumbersome and unnatural because we want to feel easy when watching the television. From this viewpoint, we employed the hands-free speech recognition using a microphone array and hand pointing recognition with no color markers using two cameras.

The applications being developed in the conversational television with these multi-modal interactions are Word explanation, Human information retrieval, Event retrieval in soccer game videos and Event retrieval with contextual awareness in baseball game videos which are described in section 5.

In this paper, we describe the structure of the conversational television in section 2. Then in section 3 and 4, hands-free speech recognition and hand pointing recognition are described. Applications are described in section 5.

2. STRUCTURE OF CONVERSATIONAL TV

Fig.1 shows a structure of the conversational TV. In the back end processing, meta data are extracted in advance from news, drama, soccer and baseball videos through contents analysis such as speech and video processing and their integration. The meta data are stored in the multimedia database together with their original contents.

In the front end processing, when a user asks questions, hands-free speech recognition, estimation of speaker direc-

MM'05, November 6–11, 2005, Singapore.

tion and recognition of hand pointing action are carried out. Then, the word explanation, human information retrieval, event retrieval in soccer or baseball game videos are carried out using the analyzed contents in the multimedia database and internet. The retrieved data are transformed and presented to the user.

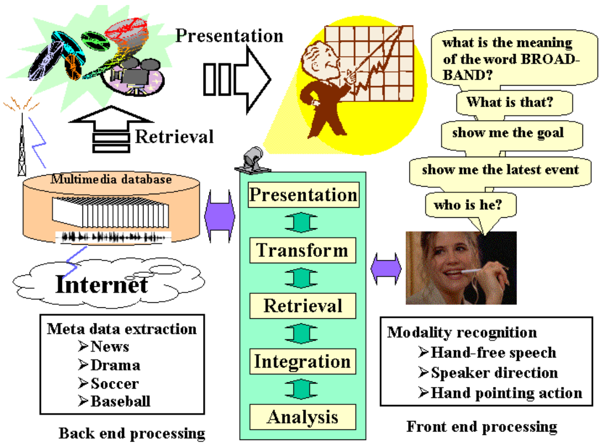
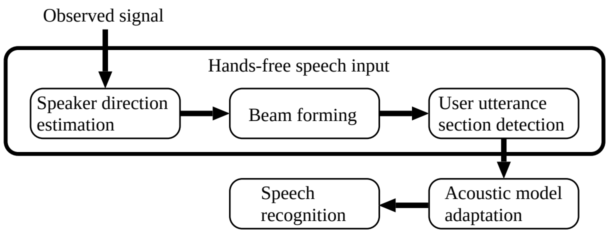


Figure 1: Structure of a conversational TV

3. HANDS-FREE SPEECH RECOGNITION

Fig.2 shows an overview of our proposed hands-free speech recognition in a front end processing of the conversational TV. In the figure, the DOA (Direction of Arrival) of target speech signal is estimated by using a microphone array and the target speech signal is enhanced by beam forming. Then, the time section of user utterance is detected automatically from the continuously observed signal. Furthermore, by applying a 2-levels MLLR based noise adaptation, the enhanced speech signal is recognized accurately.



Observed signal
Hands-free speech input
Speaker direction estimation
Beam forming
User utterance section detection
Speech recognition
Acoustic model adaptation

Figure 2: Diagram of hands-Free speech recognition

3.1 Direction Estimation of Arriving Signal

To capture the target speech signal with high quality, it is required for a microphone array to form the directivity to the target speech signal and suppress the other signals. In this paper, a delay and sum beam former[1] is employed to capture the target speech signal with high quality. Then the DOA is estimated by using a CSP(Cross power Spectrum Phase analysis) method[2].

3.2 Detection of User Utterance

In the conversational TV, it is supposed that the user asks to the TV even in announcer's speech from TV speakers. Therefore, to recognize the user utterance in a hands-free mode, a speech input interface is required to barge into announcer's speech (TV sounds). Generally, this speech input interface is realized by detecting the time section of user utterance from continuously observed signal. We propose here the user utterance detector based on time stability of the DOA. In this paper, it is supposed that the TV sounds arrive from the back of the microphone array. In this case, the TV sounds are captured with various reflections at the microphones. Therefore, the DOA of the TV sounds is not stable in the time sequence. On the other hand, the DOA of the user utterance is stable because it arrives from the front of the microphone array. Under these assumptions, time section with DOA stability is detected as the user utterance section as shown in Fig.3.

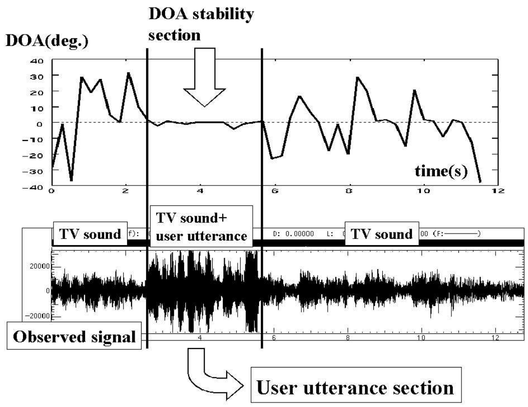


Figure 3: An example of detection of user utterance

3.3 2-Levels MLLR Adaptation

The beam forming can reduce sounds other than the target speech. However, the background noise is still superimposed on the waveform after beam forming. To cope with the residual background noise, we propose here a 2-levels MLLR adaptation. The first level of the 2-levels MLLR is the adaptation of HMMs to the residual noise after the beam forming using 15 sentences spoken by 5 males (noise adaptation). The second level of the 2-levels MLLR is the adaptation of HMMs to an individual speaker using 3 sentences spoken by each user after the first level noise adaptation (speaker adaptation).

3.4 Experiments

We evaluated the hands-free speech recognition in the conversational TV environment. The experimental materials are 100 sentences spoken by 5 Japanese male subjects and include 20 keywords appeared in the TV news. Each utterance is spoken by the subject in front of a microphone array. The distance from the subject to the microphone array(linear type, 16 microphones and 2cm intervals) is 2m and sampling rate is 16kHz. By using these materials, we evaluated the hands-free speech recognition by sub-word model based keyword spotting. The noise sources are the TV sounds and the fan noise generated from 4 digital projectors and 9 PCs

(Noise level is about 55dB.). The TV news is NHK TV news broadcasted at 12:00 on November 30 in 2001.

The acoustic models used in the speech recognition are the speaker independent monophone HMMs(3 states and 12 mixtures). They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data were taken from the JNAS(Japanese Newspaper Article Sentences) database. The feature parameters are composed of 39 MFCCs with 12 MFCCs, Log-energy and their first and second order derivatives(frame length: 20ms and frame shift: 10ms).

3.5 Experimental Results

In the experiments of time section detection of the user utterance and the DOA estimation, 89 utterance sections were correctly detected among 100 sentences and the DOA estimation rate to the correctly detected sections was about 95.5%. Here, the DOA estimation rate allows the estimation error of ±5 degrees. Table 1 show the keyword extraction rate of the correctly detected sections. Here, the baseline keyword extraction rate(without adaptation) was 48.3%. In Table 1, by using the proposed supervised noise and speaker adaptation, the best keyword extraction rate 83.2% was obtained by improving 34.9 points compared to the conventional method. These hands-free speech recognition methods (DOA, beam forming, user utterance section detection, speech recognition) are carried out in a real time.

Table 1: Keyword extraction rate by 2-levels MLLR adaptation (%)

	Unsupervised speaker adaptation (2nd level)	Supervised speaker adaptation (2nd level)
Supervised noise adaptation (1st level)	71.9(64/89)	**83.2(74/89)**

4. HAND POINTING RECOGNITION

Fig.4 shows an estimation flow of three dimensional coordinate on the screen pointed by a user. Two cameras are used; one locates in right side of the user and the other locates on the bottom side. A finger point and head are extracted on the camera images. Then the three dimensional coordinates of the finger point and head are estimated by camera calibration. Finally three dimensional point is extracted on the screen as the crossing point between the screen plane and three dimensional line connecting the finger point and head.

4.1 Skin Color Region Extraction

Input camera images are converted from RGB to luv, HSV, RGBY, Wr color spaces[3]. These four images are thresholded into binary images and integrated into one binary image. Skin color regions corresponding to hand and face are extracted on the binary image. The one near screen is regarded as finger and the other is regarded as face (head).

4.2 3D Coordinate Estimation

Fig.5 shows how to obtain three dimensional coordinates of the finger point and head using two camera calibration.

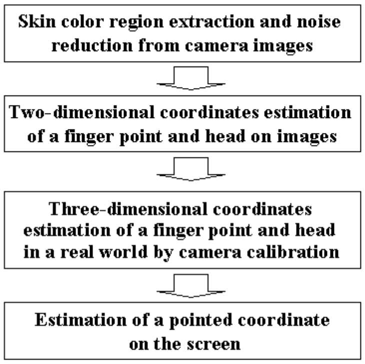

Skin color region extraction and noise reduction from camera images
Two-dimensional coordinates estimation of a finger point and head on images
Three-dimensional coordinates estimation of a finger point and head in a real world by camera calibration
Estimation of a pointed coordinate on the screen

Figure 4: Estimation flow of pointed 3D coordinate

Three dimensional line is drawn from the finger point coordinate on each camera image into a direction computed by the camera calibration. Intersection of these two lines from two cameras is computed as three dimensional coordinate of the finger point as shown in Fig.5 and the following equations;

$$\mathbf{L_1}(\mathbf{s_1}) = \mathbf{R_1^{-1}}(s_1\mathbf{m_1} - \mathbf{t_1}) \tag{1}$$
$$\mathbf{L_2}(\mathbf{s_2}) = \mathbf{R_2^{-1}}(s_2\mathbf{m_2} - \mathbf{t_2}) \tag{2}$$

Here $\mathbf{m_1}$, $\mathbf{m_2}$ are the finger point coordinates on the $Camera1$ and $Camera2$, and $\mathbf{R}$, $\mathbf{t}$ are the camera rotation and translation parameters computed by the camera calibration. Practically, line $\mathbf{L_1}\mathrm{C}\mathbf{L_2}$ does not always intersect, then the nearest point of the two lines are computed as the intersection point by the following equation.

$$s_1 = \det\{(\mathbf{P_2} - \mathbf{P_1}), \mathbf{V_2}, \mathbf{V_1} \times \mathbf{V_2}\}/|\mathbf{V_1} \times \mathbf{V_2}|^2 \tag{3}$$
$$s_2 = \det\{(\mathbf{P_2} - \mathbf{P_1}), \mathbf{V_1}, \mathbf{V_1} \times \mathbf{V_2}\}/|\mathbf{V_1} \times \mathbf{V_2}|^2 \tag{4}$$

Here, $\mathbf{V_i} = \mathbf{R_i^{-1}}\mathbf{m_i}$, $\mathbf{P_i} = -\mathbf{R_i^{-1}}\mathbf{t_i}$. The three dimensional face point is computed in the same way.

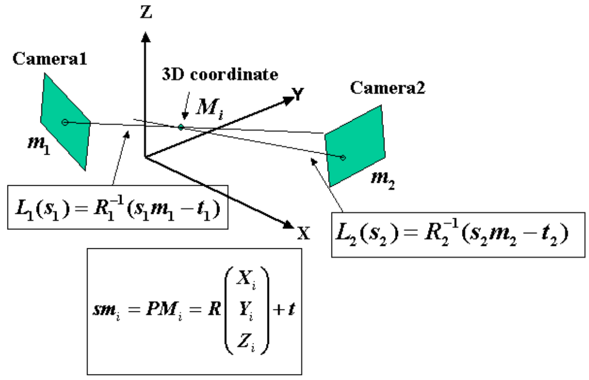

Z
Camera1
3D coordinate
M_i
Y
Camera2
m_1
m_2
X
$L_1(s_1) = R_1^{-1}(s_1 m_1 - t_1)$
$L_2(s_2) = R_2^{-1}(s_2 m_2 - t_2)$
$sm_i = PM_i = R\begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} + t$

Figure 5: 3D estimation of finger point and head

4.3 3D Screen Point Estimation

Three dimensional point on the screen is estimated as the crossing point between the screen plane and three dimensional line connecting the finger point and head three dimensional coordinates as shown in Fig.6. The black line shows the arm line, but it does not point the exact objects. On the other hand, the white line we propose is the line connecting the finger point and head in the three dimensional space and shows the exact objects.

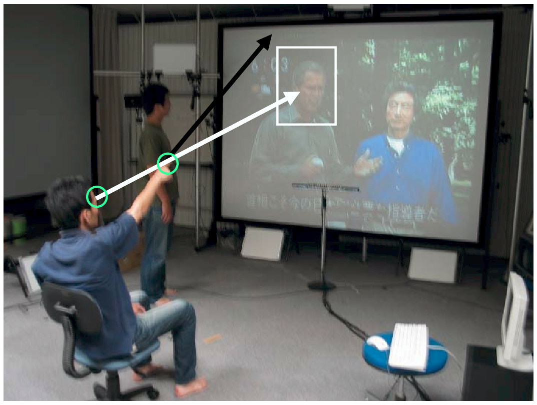


Figure 6: Pointed 3D coordinate on the screen

4.4 Experimental Results

We carried out an experiment to measure the averaged time required to point the target on the screen for more than one second. The time was measured by one person repeating 100 times of the target pointing. The target moves after the success of the pointing. The result is shown in Table2. The proposed method was carried out in real time and compared with the pointing methods of conventional motion tracker and mouse. The result indicates that the mouse device is fastest and the proposed method is almost same as the method of motion tracker.

Table 2: Time required for pointing

Method	Averaged time(ms)
Proposed	2420
Motion tracker	2132
Mouse	1520

5. APPLICATIONS

As the applications of the conversational TV with hands-free speech recognition and hand pointing action recognition, we mention here what we are developing at present.

5.1 Word Explanation

Keywords are automatically recognized in real time on the anchor person speech in the news program using originally developed speech recognition software. Keywords are extracted using TF-IDF and listed at the bottom of the display with the first frame of the corresponding shot. When a user asks a question ”what is the meaning of the word BROAD-BAND?” or ”what is that?” by his voice just after the anchor person said the sentence including the word ”broad-band” in the news program, the conversational TV recognizes the word, retrieves it in the internet and explains it by the synthesized voice.

5.2 Human Information Retrieval

Human faces are automatically extracted and recognized in real time using boosting techniques on the broadcasted videos. When a user asks a question ”who is he?” by his voice and pointing the person on the television by his hand, the conversational TV recognizes the pointed person's name and retrieves his information from the internet and shows his related URL.

5.3 Event Retrieval in Soccer Game Videos

Soccer events such as goal, corner kick, free kick are automatically recognized by integrating positions of a ball, players, goal and corner as well as speed of the ball, after tracking the players and a ball by normalized cross correlation method. When a user asks a question ”show me the goal” or ”show me the latest event”, then the television retrieves the event and replays the corresponding video clips.

5.4 Retrieval with Contextual Awareness

Batter name is automatically recognized from announcer speech by using speech recognition software. His face is also recognized used in the same software mentioned in 5.2. When a user asks a question ”show me the grand slam” when the batter is Mr. Matsui, then the recent Mr. Matsui's grand slam is replayed. However, when the batter is Mr. Ichiro, then the recent Mr. Ichiro's grand slam is replayed. In this way, the same question ”show me the grand slam” retrieves the different grand slam video clips depending on the context of who the batter is. We call this application as the contextual awareness because both the television and user understand who the batter is. The video clips such as grand slam are automatically extracted by pitcher-catcher scene extraction on the video sequence and keyword extraction on announcer speech.

6. CONCLUSIONS

We proposed in this paper the conversational TV with multi-modal interaction of hands-free speech recognition and hand pointing recognition. In hands-free speech recognition, 89 % user speech were correctly detected and their direction was 95.5% correctly estimated. The keywords asked by the user were 83.2% correctly recognized. In the hand pointing action, the pointing accuracy was almost same as that by a motion tracker. Using these multi-modal interaction, we mentioned the developing application software such as word explanation, human information retrieval, event retrieval in soccer and baseball game videos with contextual awareness. In future, we plan to develop the LSI and install it into the conversational TV.

7. REFERENCES

[1] J.L. Flanagan, J.D. Jhonston, R. Zhan and G.W. Elko, “Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms,” *J.Acoust. Soc. Am.*, **78**, 1508-1518 (1985).

[2] M. Omologo and P. Svaizer, “Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique,” *Proc. ICASSP'94*, **1**, 273-276 (1994).

[3] G. Gomez. On selecting colour components for skin detection . Proc. of the ICPR, vol. 2, pp. 961-964, Aug. 2002.