*MVA '98* IAPR Workshop on Machine Vision Applications, Nov. 17-19, 1998, Makuhari, Chiba, Japan

3—14

# A Novel Application of Real-time Face Tracking and Microphones Array to Pick up Human Voice Remotely and Clearly

Hiroshi Mizoguchi Takaomi Shigehara Muneharu Yokoyama* Taketoshi Mishima
Faculty of Engineering, Saitama University
Address: 255 Shimo Okubo, Urawa 338-8570, Japan
E-mail: hm@ics.saitama-u.ac.jp

## Abstract

This paper proposes a novel application of computer vision and signal processing, called Virtual Wireless Microphone (VWM). It integrates real-time face tracking and sound signal processing. VWM is intended to be used as a speech signal input method for human computer interaction(HCI), especially for autonomous intelligent agent that interacts with humans like as "digital secretary". Utilizing VWM, the agent can clearly listen human master's voice remotely as if a wireless microphone were put just in front of the master.

## 1 Introduction

Both semiconductor technology and computing power have rapidly grown in these years. These remarkable trends have recently raise an expectation for more friendly computer-human interface[1]-[8]. If computer could listen human voice command and behave properly, it would be very convenient and much desirable for us. Emerging agent technology draws a dream towards such a user-friendly computer. Typical example is an autonomous intelligent agent that interacts with humans like as a *digital secretary*.

However even state-of-the-art speech recognition technology requires high quality noiseless input. Users of current speech recognition system are forced to put headset microphone to prevent background noise. Because of this difficulty and inconvenience of usage, the speech recognition still cannot be applied to the human computer interaction (HCI).

To make the speech recognition practically applicable to the HCI, some novel technique to pick up speech sound clearly and remotely is keenly required. In other words, a technique to form *acoustic focus* is needed. To realize such technique there are two problems to be solved. One is how to track and obtain location of human's face in real time. The other is how to form the acoustic focus at the measured face location in three dimensional space. The location of the face is not fixed according to motion of the human's head and body. Especially, in case of the digital secretary application, the user may walk around in his/her office room.

To solve these problems this paper proposes a novel computer human interface, named Virtual Wireless Microphone (VWM), which integrates computer vision and sound signal processing.

As for the first problem, real time face tracking and binocular stereo vision are utilized. Face tracking is realized with skin color extraction. Binocular stereo vision calculates three dimensional location of the tracked face.

As for the second problem, the proposed VWM utilizes microphones array. Setting gain and delay of each microphone properly enables to form acoustic focus at desired location. Based upon the location measured by the above mentioned stereo vision, gain and delay of each microphone are determined. By repeating this calculation at video rate enables the acoustic focus to track the mouth in real time.

Dai et al.[9] discussed microphones array as a mean to realize telescopic microphone system with variable focus. Our work reported in this paper is an extension of their work by integrating with real-time computer vision.

In the following, section 2 describes principle of the proposed virtual wireless microphone. In section 3, real time face tracking is discussed. Section 4 describes mechanism of acoustic focusing by microphones array. Results of simulation are also described. Section 5 is conclusion.

## 2 Virtual Wireless Microphone

Fig. 1 illustrates principle of the proposed system. The system consists of face tracking part and microphones array part. The face tracking part detects and tracks

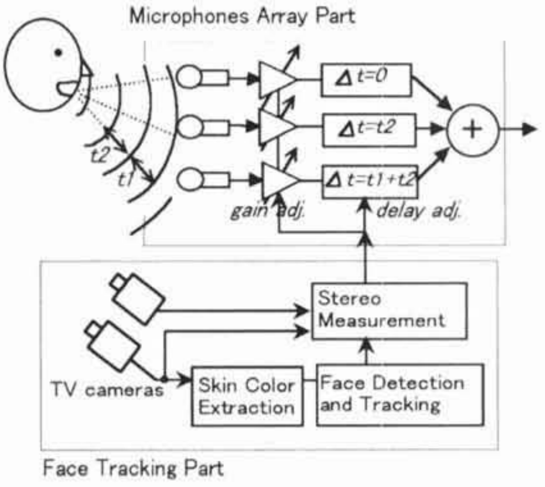

Fig. 1 Principle of proposed system

* Currently with Chukyo University

human face. And it continuously calculates three dimensional coordinates of the face utilizing binocular stereo. To detect and track face area, skin color extraction method is utilized. Stereo matching is performed with dedicated hardware to calculate correlation. The calculated three dimensional coordinates is sent to microphones array part as output.

The microphones array part makes *acoustic focus* at the face location that is detected, tracked and measured by the face tracking part. Since distances from a sound source to each microphones are different, a sound wave of the same phase reaches each microphones at different time. In other words, phases of the microphones output signal at a time are so different. The acoustic focus can be realized by equalizing the phases by adjusting delay and gain of each microphones output based upon the measured location of the face.

## 3 Face Tracking

The face tracking part consists of skin color extraction and binocular stereo. A method used in the skin color extraction part is based upon YIQ color representation instead of commonly used RGB one, because I and Q components of the YIQ are independent of brightness and free from brightness change. Original idea of the method is found in Mori et al[10]. Skin color is possible to be defined as some continuous region in IQ plane. Thus skin color area of the input image can be extracted by filtering each pixel of the input image based upon whether the corresponding point in the IQ plane is within the region or not.

The extracted face area in one eye is matched with corresponding area in other eye image for binocular stereo calculation. Stereo pair matching is based upon cross correlation utilizing dedicated hardware. The hardware is originally developed by Inoue et al. of Univ. of Tokyo [11][12][13] and commercialized by Fujitsu [14][15][16]. We utilize Fujitsu's product. It can calculate cross correlation of 8 by 8 template within search space at about 500 times during one video frame(1/30 sec). Since the calculation of cross correlation is performed at far beyond video rate by the hardware, the stereo calculation can be done in real time.

## 4 Microphones Array

This section describes principle of the acoustic focus. Fig. 2 shows coordinate system of microphones array. As shown in the figure the microphones are laddered on *x*-axis. Each distance between microphones is denoted as *d*. In this discussion, frequency of sound is assumed as 1KHz, i.e. wavelength of the sound is about 34cm at ordinary room temperature. Distance between microphones, *d*, is set to half of wavelength, about 17cm. Microphone is indexed as *Mi*,, where *i* ranges from *-N* to *N*. Thus number of microphones is 2*N*+1.

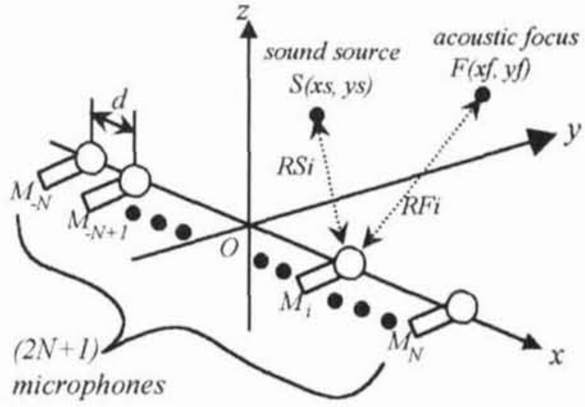

Fig. 2 Coordinate system of microphones array

In the figure, *F*(*xf, yf*) denotes the acoustic focus. *S*(*xs, ys*) denotes the sound source. Distance from the focus to *i*-th microphone is represented as *RFi*.
Similarly *RSi* denotes distance from the sound source to *i*-th microphone. When the sound source is sine wave with amplitude 1 and frequency *f*, it is
written as

$$s(t) = \sin 2\pi ft \qquad (1)$$

Output of *i*-th microphone is as follows.

$$xi(t) = \frac{1}{RSi} \sin 2\pi f(t - \tau i) \qquad (2)$$

Where $\tau i$ is the time delay from the source to the *i*-th microphone. The delay is proportional to the distance from the source to the microphone.

The delay is written as

$$\tau i = RSi / v \qquad (3)$$

where *v* is velocity of sound at current temperature. The amplitude of the output from the *i*-th microphone decays in inverse relation to the distance from the sound source.

In case the sound source is located at the focus point, there must be the furthest microphone from the focus and distance between the microphone and the focus is represented as *RFmax*. Putting additional delay to the output signal of each microphone by *(RFmax-RFi)/v* equalize phase of each microphone's output. And setting gain of *i*-th microphone as *RFi* compensates the amplitude decaying. In other words, putting the additional delay by *(RFmax-RFi)/v* and setting gain as *RFi* realizes the acoustic focus at the location *F(xf, yf)*.
Location of the face is obtained with real-time binocular stereo vision as described in the previous section.

Summation of each microphone's output *y(t)* can be written as;

$$y(t) = \sum_{i=-N}^{N} RFixi(t) \qquad (4)$$

Thus this equation becomes

$$y(t) = (2N+1)\sin 2\pi f(t - \frac{RF\max}{v}) \qquad (5)$$

because of $RFi = RSi$.

In case the sound source is not located at the focus point, i.e. $S(xs, ys) \neq F(xf, yf)$, the summation, $y(t)$ is written as follows.

$$y(t) = \sum \frac{RFi}{RSi} \sin 2\pi f(t - \frac{RSi + RF\max - RFi}{v})$$

$$= \left( \sum_i \alpha i(xs, ys; xf, yf) \right) \sin 2\pi ft$$

$$+ \left( \sum_i \beta i(xs, ys; xf, yf) \right) \cos 2\pi f$$

$$= A(xs, ys; xf, yf) \sin(2\pi ft + B(xs, ys; xf, yf)) \qquad (6)$$

$A(xs, ys; xf, yf)$ represents the sensitivity of the microphones array for given focus $F(xf, yf)$.

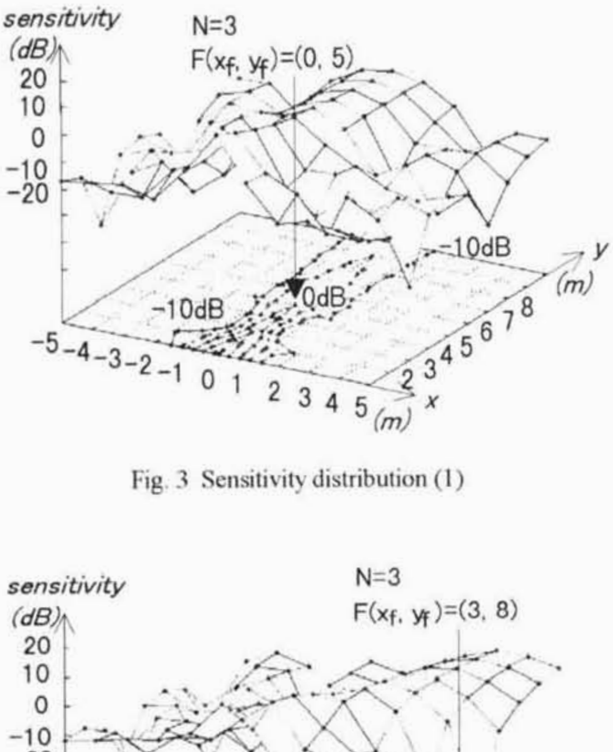

Fig. 3 Sensitivity distribution (1)

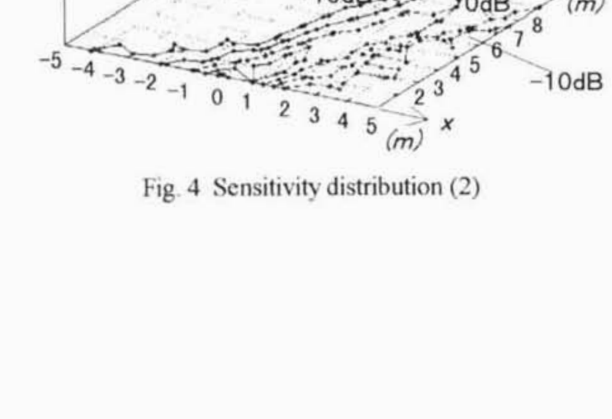

Fig. 4 Sensitivity distribution (2)

To confirm the feasibility of the proposed idea, we have conducted computer simulation. Result of the simulation is as follows. Fig. 3 and 4 plots the spatial distribution of the sensitivity for different focus points. In the figures, $z$-axis denotes decibel representation of the sensitivity, $A(xs, ys; xf, yf)$.

For better visibility, contour line is plotted on $x$-$y$ plane. We can see higher sensitivity spans from microphones ($x$-axis) to the focus point and the sensitivity steeply decreases away from the focus. Fig. 5 shows a result of different number of microphones. As the numbers increase, the sensitivity distribution becomes steeper.

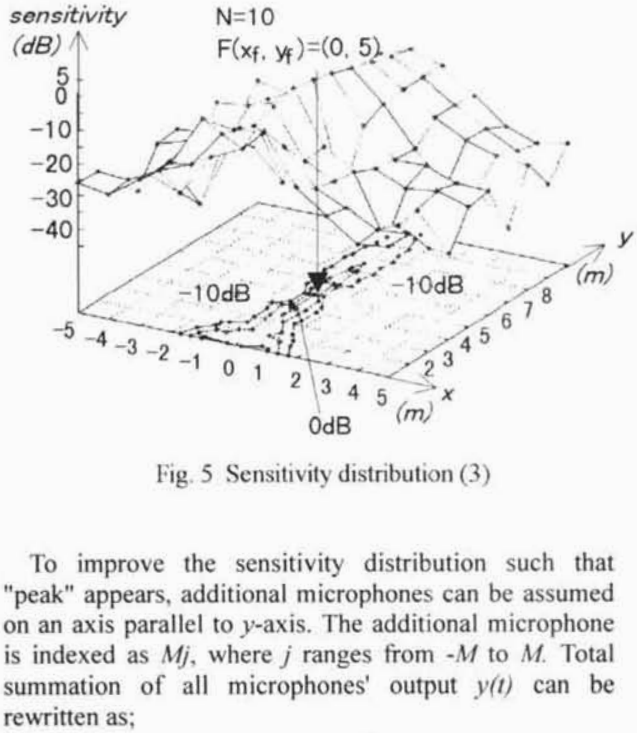

Fig. 5 Sensitivity distribution (3)

To improve the sensitivity distribution such that "peak" appears, additional microphones can be assumed on an axis parallel to $y$-axis. The additional microphone is indexed as $Mj$, where $j$ ranges from $-M$ to $M$. Total summation of all microphones' output $y(t)$ can be rewritten as;

$$y(t) = \sum_{i=-N}^{N} RFixi(t) + \sum_{j=-M}^{M} RFjxj(t) \qquad (7)$$

This equation is also rewritten as follows.

$$y(t) = \left( \sum_i \alpha i + \sum_j \chi j \right) \sin 2\pi ft$$

$$+ \left( \sum_i \beta i + \sum_j \delta j \right) \cos 2\pi ft \qquad (8)$$

where

$\alpha i = \alpha i(xs, ys; xf, yf), \chi j = \chi j(xs, ys; xf, yf),$

$\beta i = \beta i(xs, ys; xf, yf), and\ \delta j = \delta j(xs, ys; xf, yf).$

The equation (8) can be rearranged to a form similar to (6). Fig. 6 shows a result of the extended system with additional microphones. Steeple peak of sensitivity is appeared at the focus point. The *acoustic focus* or *spot* can be realized at the desired location.

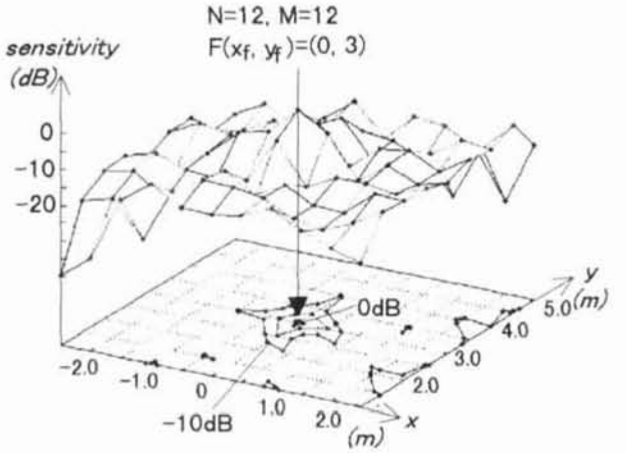

Fig. 6 Improved sensitivity distribution

## 5 Conclusion

This paper proposes a novel application of real-time computer vision and sound signal processing. It is named Virtual Wireless Microphone (VWM). It integrates real time visual tracking of face and sound signal processing. Key technologies are real-time skin color extraction, correlation based stereo matching, and acoustic focusing with microphones array. Preliminary experiments and simulation results prove the feasibility of the proposed idea. Implementing a real system based upon the idea and experimenting by the system are future works.

## References

[1] P. Maes, "Agents that reduce work and information overload", *CACM*, Vol.37, No.7, pp.31-40, 1994.

[2] M. C. Torrance, "Advances in Human-Computer Interaction: The Intelligent Room", *Working Notes of the CHI95 Research Symposium*, 1995.

[3] T. Sato, et al.,"Robotic room: Symbiosis with human through behavior media", *Robotics and Autonomous Systems*, No.18, pp.185-194, 1996.

[4] H. Mizoguchi, et al.,"Robotic office room to support office work by human behavior understanding function with networked machines", *IEEE/ASME Trans. on Mechatronics*, Vol.1, No.3, pp.237-244, 1996.

[5] A. Pentland, "Smart Rooms", *Scientific American*, pp.54-62, 1996.

[6] H. Asada et al., "Total Home Automation and Health Care/Elder Care", *Tech. Rep., Dept. of Mech. Eng.*, MIT, 1996.

[7] R. P. Picard, *Affective Computing*, MIT Press, 1997.

[8] S. Mann, "Wearable computing: A first step toward personal imaging", *Computer*, pp.25-31, 1997.

[9] W. Dai et al., "Basic Study for Realizability of a Telescopic Microphone System", *Transactions of SICE*, Vol.35, No.8, pp.843-845, 1997.

[10]T. Mori et al.,"Action Recognition System based on Human Finder and Human Tracker", Proc. of IROS'97, pp.1334-1341, 1997.

[11]H. Inoue et al., "Robot Vision System with a Correlation Chip for Real-time Tracking, Optical Flow and Depth Map Generation", *Proc. ICRA'92*, pp.1621-1626, 1992.

[12]H. Inoue et al., "Real-Time Robot Vision System based on Correlation Technology", *Proc. of ISIR*, pp.675-680, 1993.

[13]T. Mori, M. Inaba and H. Inoue, "Visual Tracking based on Cooperation of Multiple Attention Regions", *Proceedings of ICRA'96*, pp.2921-2928, 1996.

[14]T. Uchiyama et al.,"Hardware Implementation of the Video-rate Tracking Vision", *Proc. of the 12th Ann. Conf. of RSJ*, pp.345-346, 1994.

[15]N. Sawasaki et al.,"Design and Implementation of High-speed Visual Tracking System for Real-time Motion Analysis", *Proc. of ICPR'96*, pp.478-483, 1996.

[16]T. Morita et al., "Color Tracking Vision", *Proc.of the 14th Ann. Conf. of RSJ*, pp.279-280, 1996.