4B-4

音源定位及び唇の動き検出による 複数ユーザ環境における発話者認識

金鉉燉† 駒谷和範† 尾形哲也† 奥乃博†

†京都大学大学院情報学研究科知能情報学専攻

1. Introduction

The objective of this research is to develop the techniques which enable talker identification among multiple people for effective human-robot interaction. In conventional systems to find out a talker, audition systems such as VAD (Voice Activity Detection) and sound localization mainly have been applied into a robot. For this reason, it is difficult to find out a talker in noisy environments because the feature information of speech signals can be hardly extracted. Our system is able to find out a talker by using additional vision information such as lip movement detection even in noisy environments. Moreover, since face detection can remove the misdetection of sound localization and compensate the angle error of that, it is able to distinguish a desired talker among multiple people even if they are in close position. Also, pitch extraction as well as sound localization can identify a desired talker according to the difference of pitch. To verify our system's feasibility, the proposed system is installed in the robot, called SIG2, which has been developed in our laboratory.

2 System Structure

Our system consists of two computers as you see Figure 1. One is in charge of audition system which can perform VAD and sound localization. The other is in charge of vision system which can perform face detection and lip movement detection. Firstly, a robot can find the location of people through sound localization and face detection. Then, a robot can finally find out a person who is talking among multiple people through VAD and lip movement detection.

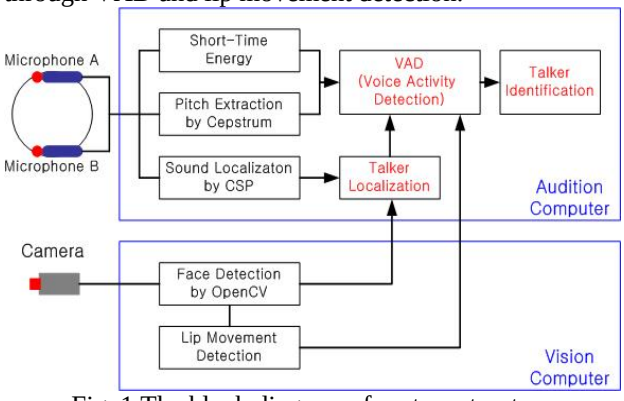

Fig. 1 The block diagram of system structure

Talker Identification among multiple people based on Sound Localization and Lip Movement Detection. Hyun-Don Kim, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno (Kyoto Univ.)

3. VAD (Voice Activity Detection)

To find out a speaker who is talking with robot among people, it is firstly necessary to classify whether input signals are voice or not by using information of short-time energy and voice pitch.

3.1 Short-Time Energy

The short-time energy is used to know whether there are signals or not according to the magnitude. The short-time energy of a frame is denoted as (1).

$$E_{frame} = \frac{1}{k}\sum_{i=0}^{k} x^2(i) \tag{1}$$

where $x(i)$ means the sampling data of i-th step and k is the number of steps.

3.2 Pitch extraction

Cepstrum means the signals made by inverse Fourier transform of the logarithm of Fourier transform of sampled signals [1]. One of the most important features of cepstrum is that if the signal is periodic signal, the signal made by that will also present peaks signal at intervals of each period. Therefore, cepstrum can reliably extract the pitch of a speech signal. Given a signal $x(w)$, the equation of the cepstrum is denoted as (2).

$$c_c(\tau) = IFFT\left\{\log\left|x(\omega)\right|\right\} \tag{2}$$

Then, with the number of samples between two peak signals found, the pitch can be detected by (3).

$$Pitch = \frac{\text{Sampling Frequency}}{\text{Number of samples between the two peaks}} \tag{3}$$

The frequency of a vocal cord concerning human beings exists between 50 and 250Hz in case of a male and between 120 and 500Hz in case of a female. Therefore, if input signals are voice, pitches will show constant frequency within the range of the fundamental frequency of human voice.

4. Sound Localization by CSP

The direction of the sound source can be obtained by estimating the TDOA (Time Delay Of Arrival) between two microphone outputs. When there is a single sound source, the TDOA can be estimated by finding the maximum value of the CSP (Cross-power Spectrum Phase) coefficients [2], as derived from (4)(5).

$$csp_{ij}(k) = IDFT\left[\frac{DFT\left[s_i(n)\right]DFT\left[s_j(n)\right]^*}{\left|DFT\left[s_i(n)\right]\right|\left|DFT\left[s_j(n)\right]\right|}\right] \tag{4}$$

$$\tau = \arg\max\left(CSP_{ij}(k)\right) \tag{5}$$

where k and n are time delay, DFT (or IDFT) is the discrete Fourier transform (or Inverse) and * is the complex conjugate, τ is an estimated TDOA.

Then the sound source direction is derived from the following equation (6)

$$\theta = \cos^{-1}\left(\frac{v \cdot \tau}{d_{\max} \cdot F_s}\right) \quad (6)$$

where v is the sound propagation speed, F_s is the sampling frequency, θ is the sound direction and d_{max} is a distance which has the maximum time delay between two microphones in order to consider the head shape which is assumed to be a sphere.

5. Vision System

5.1 Face Detection by OpenCV

For the purpose of the detection of human faces, we used OpenCV (Open Computer Vision), the open source vision library made by Intel Company. This vision library supplies the function concerning human face detection. Therefore, we can know the number and the coordination of the detected faces through developing a face detection system using OpenCV.

5.2 Lip Movement Detection

It is difficult to find out a desired speaker at a place with multiple people even if a robot succeed in finding sound direction through classifying voice signals and sound localization. This is because sound localization using two microphones not only is difficult to accurately extract a sound direction but also is easily affected by an environment noise. Therefore, we developed lip movement detection using an OpticalFlow function in OpenCV in order to distinguish a speaker who is talking to a robot accurately among adjacent people. This function has the ability to detect the variation between a former picture and a present picture.

Figure 2 shows the feature masks which are applied to the area of detected faces so as to detect lip movement. If the quantity of the variation detected by lower feature mask is large, it will infer a talker who is talking. However, if the quantity of the variation detected by upper feature mask is large, it will regard a talker whose head is moving as an undesired talker. This is to remove the misdetection as a desired talker by reason of also increasing the quantity of lip movement.

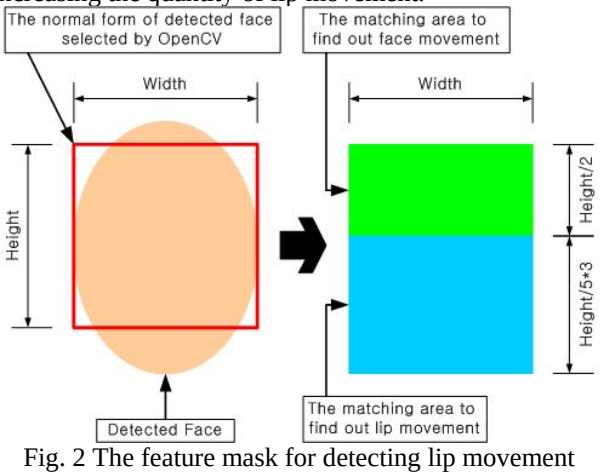

Fig. 2 The feature mask for detecting lip movement

The left part of Figure 3 shows the result of lip movement detection. At this picture, a red box (left) indicates a detected face and we can see a blue box (right) when lip movement is detected among detected faces. The right part of Figure 3 shows applied feature masks.

Fig. 3 The results of lip movement detection

6. Talker Identification

We installed developed systems in a humanoid robot, called SIG2 developed at our laboratory. (see the Fig. 4) [3]. Figure 4 shows results of executing our algorithms when someone told to SIG2 at about 1.5m and 10° from the front of SIG2. In these results, we can know that pitch and sound direction represent constant values in the interval of speech signals. Also, we can realize that lip movement detection can be detected at the same time and angle. Consequently, as long as a robot finds a sound direction roughly, it is possible to find out desired talker among people through a vision system. Also, since this system can classify the period of speech in real environment including a lot of noise, it can be applied to robust voice activity detection.

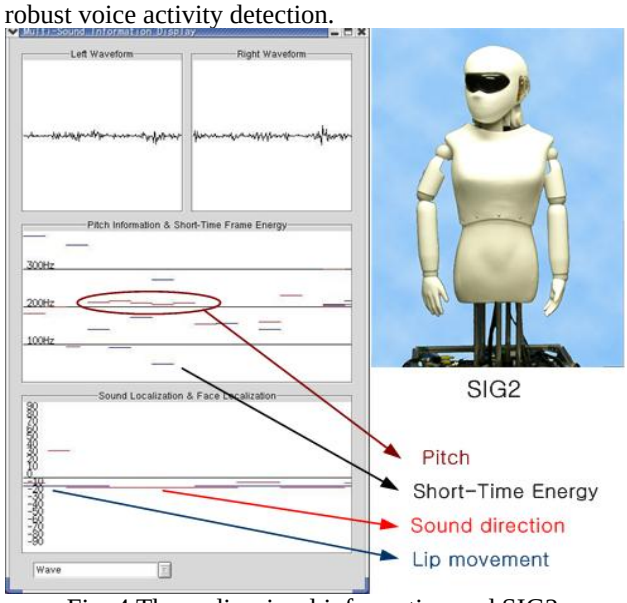

Fig. 4 The audio-visual information and SIG2

Acknowledgement

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research, and COE program of MEXT, Japan

Reference

[1] H. Kobayashi, and T. Shimamura, "A Modified Cepstrum Method for Pitch Extraction," *IEEE/APCCAS* - 1988.

[2] T. Nishiura, T. Yamada, S. Nakamura and K. Shikano, "Localization of Multiple Sound Sources based on a CSP analysis with a microphone array," *IEEE/ICASSP* - 2000.

[3] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-Time Speaker Localization and Speech Separation by Audio-Visual Integration," IROS - 2002.