Voice Activity Detection by Lip Shape Tracking Using EBGM

Masaki Aoki[†]
masamax777@me.cs.scitec.kobe-u.ac.jp

Ken Masuda[†]
masudaken@me.cs.scitec.kobe-u.ac.jp

Hiroyoshi Matsuda[†]
matsuda@me.cs.scitec.kobe-u.ac.jp

Tetsuya Takiguchi[††]
takigu@kobe-u.ac.jp

Yasuo Ariki[††]
ariki@kobe-u.ac.jp

[†]Graduate School of Engineering, Kobe University
[††]Organization of Advanced Science and Technology, Kobe University
1-1 Rokkodai, Nada, Kobe, Hyogo, 657-8501 Japan

ABSTRACT

We propose a voice activity detection of a target speaker (driver) in a car by integrating lip movement and acoustic processing. To prevent the wrong detection caused by non-target speakers using only acoustic processing, the proposed system extracts the lip movement of the target speaker by measuring the lip aspect ratio. An infrared camera is used to cope with the change of lighting environment. In order to extract the lip from gray scale images, Elastic Bunch Graph Matching is employed. Experimental results showed the proposed system improved the precision rate in the voice activity detection by approximately 40% compared to the method using only acoustic processing in a car.

Categories and Subject Descriptors

I.5.m [**Pattern Recognition**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Voice activity detection, lip shape extraction, Elastic Bunch Graph Matching

1. INTRODUCTION

Voice operation is a useful tool for drivers whose hands are occupied with driving a car. But, when a driver speaks under the narrow and noisy environment in a car, the false detection and mis-recognition will be caused, due to the car noise, music and voices other than the driver. To prevent the wrong voice detection, discrimination between the voice and noise is performed mainly using acoustic signals. However, it is difficult to judge whether the detected voice is driver's or not when only the acoustic signal is processed. From this viewpoint, we propose a new method that can track the driver's lip movement and calculate the dynamics of the lip aspect ratio. Using this method, only the driver's utterance can be detected by discriminating noise, music and voices other than the driver.

There are a lot of studies to detect the voice section of the target speaker by using the acoustic signals and face images. In the related works of the lip extraction, various features like RGB, HSV, Illumination[4], edge and moving vector have been used. In addition, various methods like Boosting[2], active search with histogram comparison[3] and template matching[5] have been employed to extract the lip. Since color information is influenced by illumination, we employ an infrared image to improve the robustness of the system for illumination changes. The conventional methods mentioned above are able to extract the lip. But it is difficult to detect voice section by using the extracted lip information due to the accuracy. From this viewpoint, we employ Elastic Bunch Graph Matching (EBGM)[1][6] to extract the lip more accurately.

Since the change of the lip shape gets large when the driver is speaking, the lip aspect ratio is computed to normalize the individual lip size. Finally, the system combines the visual and acoustic processing and detects only the driver's voice section. In the acoustic processing, voice and noise are discriminated by likelihood ratio test (LRT). Then in the visual processing, driver's voice and other voice (or noise) are discriminated by lip movement.

The rest of this paper is organized as follows. In section 2, voice activity detection (VAD) using GMM is described. In section 3, voice activity detection by lip shape extraction using EBGM is described. In section 4, how to combine the visual and acoustic VAD is described. The effectiveness of the proposed system is described in section 5.

2. VAD FROM SPEECH SIGNAL BY GMM

GMMs (Gaussian mixture model) are widely used for voice activity detection from speech signal because the model is easy to be trained and usually powerful. GMMs are expressed for the acoustic signal x_t (MFCC: Mel-Frequency Cepstral Coefficients) at time t as follows, using the mul-

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

tiple normal distribution functions $N(x_t; \mu_m, \Sigma_m)$ with the m-th mixture mean vector μ_m and covariance matrix Σ_m.

$$Pr(x_t) = \sum_m P(m) N(x_t; \mu_m, \Sigma_m) \quad (1)$$

In order to detect the voice section, two GMMs with 64 mixture numbers are trained using clean voice data and non-voice data. Using these two GMMs, the log likelihood ratio is calculated by

$$L(x_t) = \log \frac{\Pr(x_t|voice\_model)}{\Pr(x_t|non-voice\_model)} \quad (2)$$

where $\Pr(x_t|voice\_model)$ and $\Pr(x_t|non\text{-}voice\_model)$ are the voice likelihood and the non-voice likelihood respectively. To avoid the voice section being separated by the short pause, the following smoothing is performed to the log likelihood ratio $L(x_j)$.

$$L'(x_t) = \frac{1}{n} \sum_{j=t-\frac{n}{2}}^{t+\frac{n}{2}} L(x_j) \quad (3)$$

For the time section with $L'(x_t)$ over the threshold θ is regarded as the voice section. The time section with $L'(x_t)$ under the threshold θ is regarded as the non-voice section. Finally, the voice section is extracted by deleting the short gap between the voice sections.

3. VAD FROM LIP MOVEMENT BY EBGM

3.1 Gabor Wavelets

Gabor wavelets can extract global and local features by changing spatial frequency, and can extract features related to wavelet's orientation.

Eq.(4) shows a Gabor Kernel used in Gabor wavelets. This function contains Gaussian function for smoothing as well as wave vector $\vec{k}_j$ which indicates simple wave frequencies and orientations.

$$\psi_j(\vec{x}) = \frac{k_j^2}{\sigma^2} exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right)\left[exp\left(i\vec{k}_j\vec{x}\right) - exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (4)$$

$$\vec{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\nu \cos\varphi_\mu \\ k_\nu \sin\varphi_\mu \end{pmatrix} \quad (5)$$

Here, $k_\nu = 2^{-\frac{\nu+2}{2}}\pi$, $\varphi_\mu = \mu\frac{\pi}{8}$. We employ a discrete set of 5 different frequencies, index $\nu = 0, ..., 4$, and 8 orientations, index $\mu = 0, ..., 7$.

3.2 Jet

A jet is a set of convolution coefficients obtained by applying Gabor kernels with different frequencies and orientations to a point in an image. To estimate the positions of facial feature points in an input image, jets in an input image are compared with jets in a facial model.

A jet $\mathcal{J}$ is composed of 40 complex coefficients (5 frequencies $\times$ 8 orientations) and expressed as follows:

$$\mathcal{J}_j = a_j exp(i\phi_j) \, (j = 0, ..., 39) \quad (6)$$

where $\vec{x} = (x, y)$, $a_j(\vec{x})$ and $\phi_j(\vec{x})$ are the facial feature point coordinate, magnitude of complex coefficient and phase of complex coefficient, which rotates the wavelet at its center respectively.

3.3 Jet Similarity

For the comparison of facial feature points between the facial model and the input image, the similarity is computed between jet set $\{\mathcal{J}\}$ and $\{\mathcal{J}'\}$. Locations of two jets are represented as $\vec{x}$ and $\vec{x}'$. Difference between vector $\vec{x}$ and vector $\vec{x}'$ is given in Eq.(7).

$$\vec{d} = \vec{x} - \vec{x}' = \begin{pmatrix} dx \\ dy \end{pmatrix} \quad (7)$$

Here, let's consider the similarity of two jets in terms of the magnitude and phase of the jets as follows:

$$S_D(\mathcal{J}, \mathcal{J}') = \frac{\sum_{j=0}^{N-1} a_j a'_j \cos(\phi_j - (\phi'_j + \vec{k}_j \vec{d}))}{\sqrt{\sum_{j=0}^{N-1} a_j^2 \sum_{j=0}^{N-1} a'^2_j}} \quad (8)$$

The similarity $S_D(\mathcal{J}, \mathcal{J}')$ is optimized when the best $\vec{d}$ is estimated which maximizes the similarity including the both magnitude and phase.

3.4 EBGM

3.4.1 *Graph*

A set of jets extracted at all facial feature points is called a graph. In this study, a lip graph composed of facial feature points around a lip is constructed as shown in Fig.2 to extract the lip shape.

3.4.2 *Bunch Graph*

A set of jets extracted from many people at one facial feature point is called a bunch. A graph constructed using bunches at all the facial feature points is called a bunch graph. In searching the location of facial feature points, the similarity described in Eq.(8) is computed between the jets in the bunch graph and a jet at the point in an input image. The jet with the highest similarity, achieved by moving $\vec{d}$, is chosen as the target facial feature point in the input image. In this way, using a bunch graph, the location of facial feature points can be searched for under several conditions.

3.4.3 *Elastic Bunch Graph Matching*

Fig.1 shows a flow of an elastic bunch graph matching. First, given an image, the bunch graph is pasted to the image, and then local search starts using the greedy-like method. Finally the graph is extracted after all the locations of the feature points are matched.

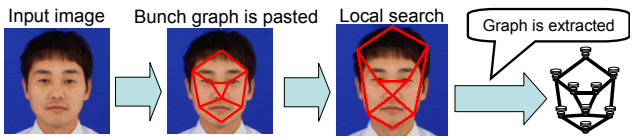

Figure 1: Elastic Bunch Graph Matching procedure

3.5 VAD by Lip Shape Tracking

A lip model is composed of facial feature points locating at an outline of a lip, nasal cavities, ears and jaw. Fig.2 shows an example of a lip model of facial feature points and the extracted result. After the lip extraction, top and bottom ends as well as right and left ends are determined on the extracted lip shape. Then the height of the lip, Lip_{Height}

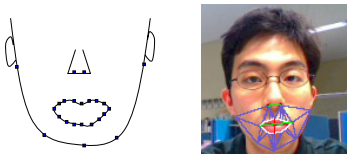

Figure 2: An example of feature points used for EBGM and the extracted lip shape

and the width of the lip, Lip_{Width} are computed. Finally, aspect ratio is computed by Eq.(9).

$$Aspect\ ratio = \frac{Lip_{Height}}{Lip_{Width}} \tag{9}$$

Fig.3 shows how the robustness is achieved for the variation of lip sizes by using the aspect ratio. In the figure, it is depicted that the aspect ratios, R_{close1} and R_{close2} are almost same when lips are closing irrespective of their sizes, but the aspect ratio R_{open} changes when they open.

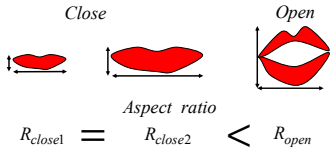

Figure 3: Aspect ratio

Under the assumption that the human lip moves in utterance, it is expected that the lip aspect ratio will change frequently in utterance. From this viewpoint, the difference of the aspect ratio between consecutive frames is computed as the lip movement. Fig.4 shows the difference of the aspect ratio between consecutive frames in both the utterance and non-utterance section. The horizontal axis and the vertical axis indicate the frame number and the difference of aspect ratio respectively. In this study, if the change of the aspect ratio exceeds some threshold, it is regarded as voice section of the target person.

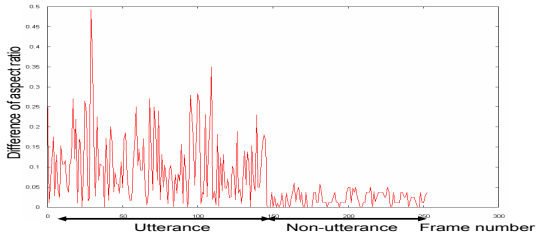

Figure 4: Change of the aspect ratio

4. INTEGRATION OF TWO VADS

This section describes a method of voice activity detection of the target speaker by integrating the results from the acoustic signal processing and the lip shape tracking described in section 3. Fig.5 shows a total system flow. In a car, the image sequence and speech signal are captured with driver's voice, other person's voice, noise and the lip movement of the driver as shown in Fig.5(a). The voice section is detected by the acoustic processing technique using likelihood ratio test (LRT) under the voice or noise hypothesis based on Gaussian mixture model (GMM) as shown in Fig.5(b). At this point, noise is excluded and the voice section is only extracted. Then the lip shape is extracted and the change of the aspect ratio is computed. If the change of the aspect ratio is over the threshold, the lip is regarded as moving. Otherwise, it is regarded as stationary as shown in Fig.5(c).

In this way, the voice section of the target person (driver) is detected as shown in Fig.5(d). The case where the driver moves his lip without voice is excluded at the stage in Fig.5(b) because there are no voice and noise. Fig.6 shows an example of voice activity detection. The horizontal axis indicates the frame number. The vertical axis indicates the change of the aspect ratio. The threshold ζ is determined to maximize the evaluation value of the precision to be described in Section 5.2.

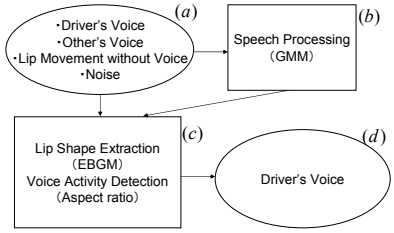

Figure 5: System flow

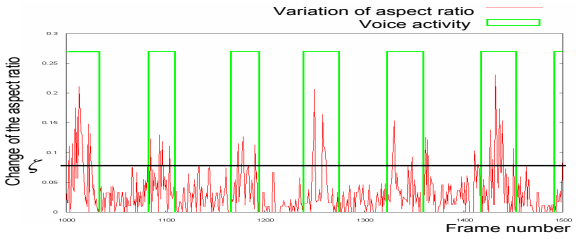

Figure 6: Voice activity detection by difference of aspect ratio

5. EXPERIMENTAL RESULTS

5.1 Experimental Condition

To create the Bunch Graph for EBGM, 142 frontal face photographs were selected from SoftPia Japan database and the 23 facial feature points were manually specified. Two moving images were used as test data. One included a Japanese male driver and the other included a Japanese female driver.

100 words of Japanese city name were uttered in the car under the idling condition in the daytime. The acoustic signal with 10〜20dB SN ratio was filtered by the high-pass filter with the cut off frequency at 200Hz to eliminate the low frequency engine noise. An infrared camera was used to cope with the lighting environment at night. The camera was set at the front of the driver's seat.

In the experiment, since a driver uttered 100 Japanese city names in a car, the acoustic signal included driver's voice and car noise, but not other voices. Therefore before

testing, the voice of 100 city names spoken by other person were manually inserted into the intervals between the driver's voice sections and it was used as the test data.

5.2 Experimental Result

Fig.7 shows the results of the lip extraction by EBGM. Fig.8(a) shows an example of the driver's voice sections. The horizontal axis indicates the frame number. Fig.8(b) shows the voice sections detected by the proposed method. Fig.8(c) shows all the voice sections including the driver and the other person. The number of all the voice sections was

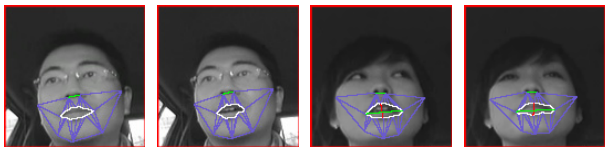

Figure 7: Example of EBGM results for test data

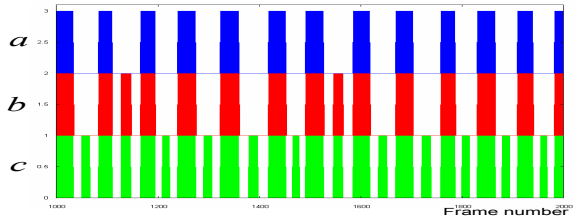

Figure 8: Experimental results

200, while the number of the driver's voice sections was 100. As a result of acoustic processing, the recall rate was 100% and the precision rate was 50%. The recall and the precision rate are defined by the following expressions respectively.

$$Recall = \frac{Detect\ \ True}{100} \times 100(\%)$$
$$Precision = \frac{Detect\ \ True}{DetectAll} \times 100(\%)$$

where, Detect True and Detect All indicate the number of correctly detected driver's voice sections and the number of all the detected voice sections respectively.

Table 1 shows the results of voice activity detection by the proposed method. The threshold was fixed for the recall to be 100%. From the table, it can be said that the proposed system improved the precision rate in the voice activity detection by approximately 40% compared to the method using only acoustic processing in a car.

Table 1: Results by the proposed method

Speaker ID	Detect All	Detect True	Recall	Precision
Male	106	100	100(%)	94.33(%)
Female	118	100	100(%)	84.75(%)

Some problems of the proposed method were found through the experiment. Fig.9 shows an example of the voice activity detection failure when the driver spoke the word "Asahi" but the change of the lip aspect ratio was too small to be detected as the voice section. The other problem is EBGM error as shown in Fig.10. When the driver's face was not frontal, EBGM failed because the images with the frontal faces were only used for constructing the Bunch Graph in the experiment. This problem will be solved by constructing the Bunch Graph using the faces in all directions.

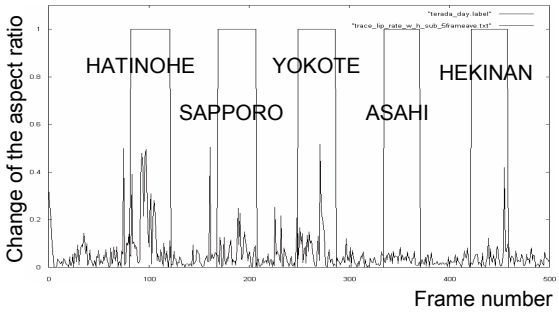

Figure 9: Changes of aspect ratio for test data

Figure 10: EBGM Error

6. SUMMARY

In this paper, we proposed the method to discriminate the driver's voice from the other person's voice or the acoustic noise such as engine or abrupt noise, using infrared image and acoustic signal. The effectiveness of the proposed method was proved by the experimental result. A future issue is to extend the proposed method for the real time processing and to realize dynamic thresholding to the change of the aspect ratio. In addition, the system will be extended to detect driver's voice activity in natural conversation instead of city names (destination) in a car.

7. ACKNOWLEDGMENTS

The facial data in this paper were used by permission of Softpia Japan. It is strictly prohibited to copy, use, or distribute the facial data without permission.

8. REFERENCES

[1] D. S. Bolme. Elastic bunch graph matching. Master's thesis, Colorado, June 2003.

[2] R. E.Schapire and Y. Singer. Improved boosting algorithm using confidence-rated prediction. *Machine Learning*, 37(3):297–336, 1999.

[3] H.Murase and V.V.Vinod. Fast visual search using forcused color matching-active search. *System and Computers in Japan*, 31(9):81–88, 7 2000.

[4] M. J.Lyons, C.-H. Chan, and N. Tetsutani. Mouthtype: Text entry by hand and mouth. *Proceedings, CHI 2004*, 1(1):1383–1386, 4 2004.

[5] O. Vanegas, K. Tokuda, and T. Kitamura. Location normalization of hmm-based lip reading: Experiments for the m2vts database. *Proc. of ICIP*, 2(2):343–347, 10 1999.

[6] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):775–779, July 1997.