Proceedings of the 2016 IEEE
International Conference on Robotics and Biomimetics
Qingdao, China, December 3-7, 2016

Laughing Voice Recognition Using Periodic Waveforms and Voice-likeness Features

-- Toward Advanced Human-machine --

Taisuke Sakano, Takahiro Kigawa, Masanori Sugimoto, Fusako Kusunoki, Shigenori Inagaki, and Hiroshi Mizoguchi

Abstract—**With a goal of advancing human-machine collaboration, we propose a method to recognize a laughing voice. A previous study that proposed a laughing voice recognition method was based on the periodic waveform feature of the laughing voice. However, identifying laughing voices from waveforms is problematic, because this can result in false positive results. To overcome this problem, we propose a laughing voice recognition method that incorporates a voice-likeness feature. In addition, we make a laughing voice recognition system using our proposed method. To confirm the efficacy of the laughing voice recognition method, we evaluated this system. In our evaluation, we were able to correctly recognize laughing voice at a high success rate of 97%. In comparison with the previous study, the recognition rate improved by 18%; therefore, our proposed method can be considered an effective in recognizing laughing voice.**

I. Introduction

In recent times, it is becoming increasingly common to have autonomous robots that operate on the basis of their interaction with humans, such as service and social robots, not only in public environments, but also in private homes [1]. These robots are required to carry out functions that naturally blend into the everyday human life; this necessitates a sense of compatibility between these robots and humans [2]. To fulfill their functions, the social robots should be able to estimate a human's feelings. To this end, it is necessary for a social robot to recognize human emotional signals such as a smile, angry face, laughing voice, and crying voice; robots can do this using their built-in camera to recognize a human's face, and their microphone to recognize voices. In the past, many researchers have tried to estimate a human's feelings using face images only; however, considerably few researchers have attempted to use the voice to do so. Therefore, in our study, we attempt to estimate a human's feelings on the basis of their voice and the way it was affected by their feelings. Because attaining a sense of compatibility between social robots and humans is a milestone towards achieving improved human-machine collaboration, and for the purposes of this study, considering that laughter might be associated with friendliness, we considered recognition of a laughing voice as the first step toward estimating a human's feelings.

Some methods of laughing voice recognition have been proposed previously. For example, the "Laughometer" is a bone-conductive microphone attached to the skin of the neck, combined with a hands-free and wireless system that monitors a laughing voice in everyday conversation [3]; another algorithm is based on the idea that the intensity of a laughing voice has characteristics of periodic peaks [4]. However, in these conventional approaches, the researchers relied on the idea that the waveform of a laughing voice exhibits a periodic change. This can lead to the problem of incorrectly recognizing a periodic change in a waveform of a non-laughing voice as a laughing voice, i.e., a false positive result.

To avoid this problem, it is necessary to not solely rely on identifying laughing using periodic waveforms, but also to devise a means of recognizing a voice. Therefore, in this study, we propose a laughing voice recognition method that incorporates the features of voice-likeness. Further, to confirm the efficacy of the proposed method, we developed a laughing voice recognition system based on our method and evaluated this system.

II. Laughing voice recognition method incorporating features of voice-likeness

As previously discussed, to overcome the problem of false positive results in laughing voice recognition systems that use the periodic waveform characteristic in laughing voices, we propose a laughing voice recognition method that incorporates a voice-likeness feature.

A. Voice-likeness

To the best of our knowledge, two previous studies have addressed the feature of voice-likeness. The first study found that the shape-features of a voice can be understood using its frequency [5]. The second study concluded that the features of voice-likeness, such as the phonological features, can be

T. Sakano is with Tokyo University of Science, Chiba, Japan (corresponding author to provide phone: +81-4-7124-1501(ex.3925); fax: +81-4-7123-9814; e-mail: 7516623@ed.tus.ac.jp).

T. Kigawa, is with Tokyo University of Science, Chiba, Japan (e-mail: 7515617@ed.tus.ac.jp).

M. Sugimoto is with Hokkaido University, Hokkaido, Japan (e-mail: sugi@ist.hokudai.ac.jp).

F. Kusunoki is with Tama Art University, Tokyo, Japan (e-mail: kusunoki@tamabi.ac.jp).

S. Inagaki is with Kobe University, Kobe, Japan (e-mail: inagakis@kobe-u.ac.jp).

H. Mizoguchi is with Tokyo University of Science, Chiba, Japan (e-mail: hm@rs.noda.tus.ac.jp).

978-1-5090-4364-4/16/$31.00 ©2016 IEEE

deduced in the low-frequency region [6]. On the basis of these studies, we believe that the shape-feature of voice-likeness becomes apparent in the low-frequency region. We therefore assume that a laughing voice recognition method could be realized using the shape-features of the voice-likeness that appear in the low-frequency region.

B. Proposed method for laughing voice recognition

To use the feature of voice-likeness for laughing voice recognition, we extracted the features of voice-likeness by applying Mel-frequency cepstrum analysis [7] and higher-order local auto-correlation (HLAC) [8]. Mel-frequency cepstrum analysis is a voice-processing method used in voice recognition. It involves converting waveform data to data showing the shape-feature in the low-frequency region. HLAC is a feature value extraction method used for applications such as texture classification and character recognition, and can extract the feature value while clearly reflecting the shape of the feature. Therefore, it is possible to extract the feature of voice-likeness whereby waveform data is converted to shape-feature data using Mel-frequency cepstrum analysis, after which the shape-feature data is converted to a feature value using HLAC. Thus, it is possible to extract the 35-dimensional voice-likeness feature that appears in the shape-feature in low-frequency range of a voice signal.

An overview of our method for laughing voice recognition is described below. First, it is necessary to obtain many different types and a sufficiently large amount of laughing and other voice data as leaning data. For this learning data, the features of voice-likeness are extracted using the process described above. Then, the laughing voice data is discriminated from other voice data using the feature of voice-likeness extracted from the learning data in the first step. At this time, to simplify this discrimination, we compress the amount of information of the 35-dimensional voice-likeness using principal component analysis (PCA) [10], which is a multivariate analysis method. Next, the laughing voice or other voice data is discriminated by using by compressing the amount of information for the voice-likeness (Learning phase). Finally, using the criteria used for the discrimination, laughing voice recognition is performed for the voice data which is not recognized as being either a laughing voice or another voice (Recognition phase). The discrimination and recognition are performed using quadratic discriminant analysis (QDA) [11]. QDA is a method of discriminating between data in different groups so that unknown data can be classified using either an elliptical or quadratic function. Fig. 1 shows the overview of the proposed method of laughing voice recognition.

Mel-frequency cepstrum analysis: Mel-frequency cepstrum analysis is a voice processing method used in voice recognition. In particular, this method is an analytical technique that refracts the nature of human hearing whereby sensitively is perceived as the sound of low frequencies based on cepstral analysis. An overview of Mel-frequency cepstrum analysis is given below. First, a voice signal is transformed using a discrete Fourier transform. As a result, the timing of the voice signal relationship of an amplitude is converted to an amplitude relationship for each frequency. Next, because the features of voice-likeness such as the phonology are apparent in the low-frequency region, imparting a weighting to the data for the low-frequency region to the converted data establishes an amplitude relationship for each frequency. Then, considering the human auditory characteristics, a logarithmic function is applied; after which, a discrete cosine transform is performed. The discrete cosine transform is a method that focuses on the low-frequency region based on the discrete Fourier transform. Fig. 2 shows the result obtained for the appearance feature of voice-likeness performed by the Mel-frequency cepstrum analysis on a laughing voice signal.

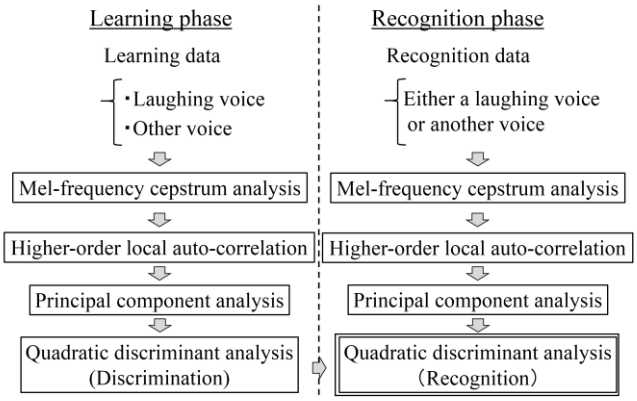
Learning phase
Learning data
・Laughing voice
・Other voice
Mel-frequency cepstrum analysis
Higher-order local auto-correlation
Principal component analysis
Quadratic discriminant analysis (Discrimination)
Recognition phase
Recognition data
Either a laughing voice or another voice
Mel-frequency cepstrum analysis
Higher-order local auto-correlation
Principal component analysis
Quadratic discriminant analysis (Recognition)

Figure 1. Overview of method of laughing voice recognition

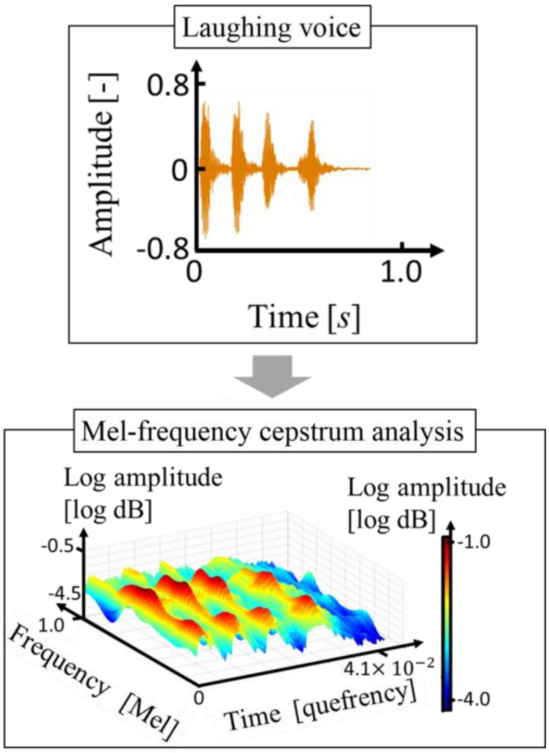

Figure 2. Results of Mel-frequency cepstrum analysis

Higher-order local auto-correlation (HLAC): The HLAC is a statistical feature value extraction technique used in applications such as texture classification, character recognition, and face recognition. This technique is based on a high-order correlation to satisfy positional invariance and additivity. This approach enables the extraction of a feature value that characterizes the structure that clearly appears.

N in the following HLAC feature H_N is represented by the following equation.

$$H_N\,(\mathrm{a}_1, \mathrm{a}_2, \ldots \mathrm{a}_N) = \int f\,(r)\,f\,(r+\mathrm{a}_1)\ \ldots f\,(r+\mathrm{a}_N) \qquad (1)$$

where r is the position vector (x, y), $f\,(r)$ is the data value at the focus point, and a_1, a_2, $\ldots\mathrm{a}_N$ are the displacement vectors. In this study, we assume that N can take a value of up to 2, and therefore the range of the displacement vector around the target point r is up to 3×3. At this time, the HLAC feature value of the number is 1 for $N = 0$, and 5 for $N = 1$, and 29 for $N = 2$, thus giving a 35-dimensional vector. Fig. 3 shows the HLAC feature value in this case. A specific calculation method is applied, which is as follows. Step 1: Calculate the product of the data values of the reference point and the point of interest. Step 2: Shift focus point and perform Step 1 for all the data. Step 3: Perform calculation for every possible combination of the 35 species. Fig. 4 shows the HLAC calculation process for discrete data (applying the Mel-frequency cepstrum analysis to a laughing voice signal) as an example of using a combination.

Principal component analysis (PCA): PCA is an analytical technique, which, in the case of many variables, considers the structure of correlation and can be analyzed using one or a small number of the overall index (the principal part) while losing as little information as possible.

The principal component is obtained by solving the following equation.

$$R\boldsymbol{u} = \lambda\boldsymbol{u} \qquad (2)$$

where R is the correlation coefficient matrix of HLAC feature vector of 35-dimensions. λ is an eigenvalue of R, and $\boldsymbol{u}$ is an eigenvector corresponding to λ.

Principal component 1 (PC1) is calculated from the first eigenvector $\boldsymbol{u}_1$ corresponding to the first eigenvalue λ_1.

Similarly, Principal component 2 (PC2) is calculated from the second eigenvector $\boldsymbol{u}_2$ corresponding to the second eigenvalue λ_2. In a similar manner, Principal components 3 to 35 are obtained too. Fig. 5 shows the example of PCA that compressed the information of the 35-dimensional HLAC feature vector into a 2-dimensional first principal component and second principal component.

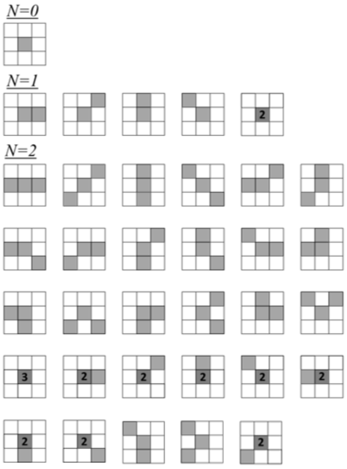

Figure 3. Thirty-five calculation patterns for HLAC [9]

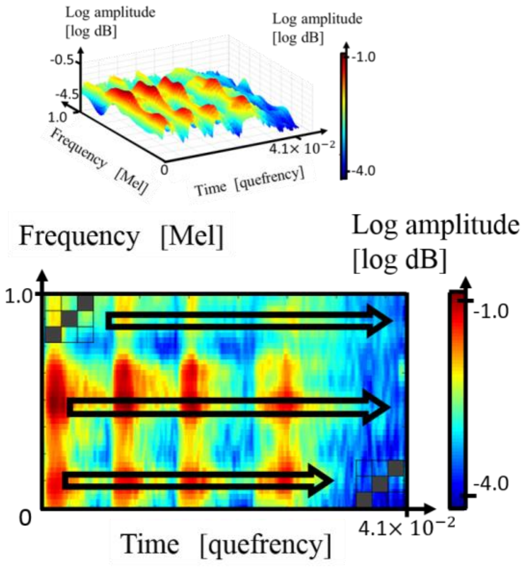

Figure 4. HLAC calculation process

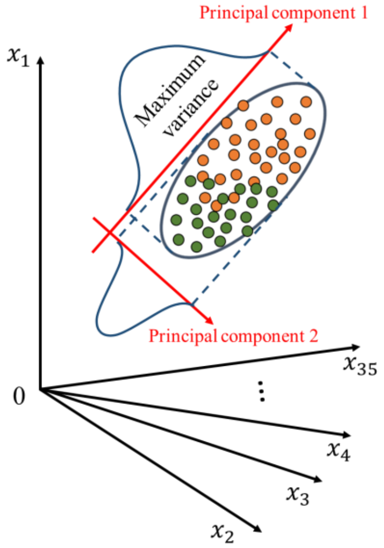

Figure 5. Example of principal component analysis

Quadratic discriminant analysis (QDA): As specified earlier, QDA is a method of discriminating between data in different groups so that unknown data can be classified using either an elliptical or a quadratic function.

As an example, QDA of multivariate data consisting of two populations (Population A, Population B) is shown by performing the following calculations (equation (3)–(5)).

In this case, Population A and Population B is expressed as a Gauss distribution; density function of a multi-dimensional Gauss distribution is represented by the following equation.

$$f(x) = 1/((2\pi)\, |\Sigma|^{1/2})\ exp(1/2\ (x - \mu)'\ \Sigma^{-1}\ (x - \mu)) \qquad (3)$$

where $f(x)$ is density function of multi-dimensional Gauss distribution, π is the abundance ratio, Σ is the variance-covariance matrix, and μ is the population mean.

Next, taking the logarithm of equation (3), we obtain:

$$\begin{aligned} &log\, f(x) = \\ &\quad - log\ (2\pi) - 1/2\ log\ |\Sigma| - 1/2\ (x - \mu)'\ \Sigma^{-1}\ (x - \mu) \end{aligned} \qquad (4)$$

The domain that unknown data X is classified as population A is represented by the following equation.

$$\begin{aligned} &D_{[A]} = \\ &\{x|\ log\ (2\pi_{[A]}) - 1/2\ log\ |\Sigma_{[A]}| - 1/2\ (x - \mu_{[A]})'\ \Sigma^{-1}\ (x - \mu_{[A]}) > log\ (2\pi_{[B]}) - 1/2\ log\ |\Sigma_{[B]}| - 1/2\ (x - \mu_{[B]})'\ \Sigma^{-1}\ (x - \mu_{[B]})\} \end{aligned} \qquad (5)$$

where $\pi_{[A]}$ is the abundance ratio of A in all populations, $\Sigma_{[A]}$ is the variance-covariance matrix for population A, $\mu_{[A]}$ is the population mean of population A, $\pi_{[B]}$ is the abundance ratio of B in all populations, $\Sigma_{[B]}$ is the variance-covariance matrix for population B, and $\mu_{[B]}$ is the population mean of population B.

III. Laughing Voice Recognition System Based on the Proposal Method

To confirm the efficacy of our proposal method, we developed a laughing voice recognition system using the proposed method (cf. Section II) and evaluated this system.

A. Laughing voice data and the other voice data (Learning data)

We obtained the laughing voice data from a radio program, and the other voice data from the speech corpus "Spoken Language" and the DSR Projects Speech Corpus (PASL-DSR). In total, we obtained 192 items of laughing voice data, and 192 items of other voice data. For both, the laughing voice data and other voice data, the sampling frequency was 16,000 Hz, and the data was saved in the .wav file format.

B. Compressing the amount of information using PCA

We calculated the contribution ratio of the principal component to represent the original information. We selected the principal component as that for which the sum of the contribution ratio is at least 99.9%. Because the sum of the contribution rate of the first principal component and the second principal component was at least 99.9%, we compressed the information of the 35-dimensional voice-likeness to a 2-dimensional first principal component and second principal component. Fig. 6 shows the contribution ratio of the principal components as a pie chart. Table 1 lists the contribution ratios of the first, second, and third principal component.

TABLE I. CONTRIBUTION OF FIRST TO THIRD PRINCIPAL PARTS

Principal part	Contribution rate [%]
PC1	99.79
PC2	0.16
PC3	0.03

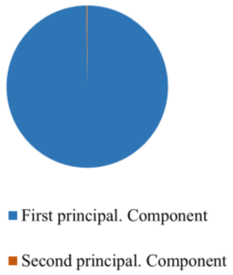

Figure 6. Resultant contribution rates

TABLE II. DISCRIMINATION RESULTS

		Result		Total data amount
		Laughing voice	Other voice	
Learning data	Laughing voice	185 (96%)	7 (4%)	192 (100%)
	Other voice	3 (2%)	189 (98%)	192 (100%)

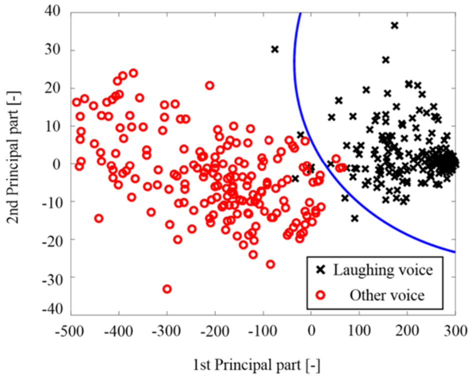

Figure 7. Discrimination of voice- likeness results

C. *Discrimination of laughing voice and other voice using QDA*

We discriminated the laughing voice and other voice data, compressed from a 35-dimensional voice-likeness to a 2-dimentional first principal component and second principal component using QDA. Out of the 192 items of laughing voice, 96% or 185 items of the laughing voice data were correctly discriminated as laughing voice data. Moreover, of the 192 items of other voice data, 98% or 189 items were correctly discriminated as being other voice data. In all, out of the 384 laughing voice and other voice data items, 97% or 374 items of laughing voice and other voice data items were discriminated correctly. Table 2 lists the classification results. Fig. 7 shows the discrimination of the feature values of the voice-likeness results obtained using QDA.

D. *Evaluation of results using 10-fold cross validation*

To test the validity of the amount of learning data and recognition precision, we evaluated the laughing voice recognition system using 10-fold cross validation [12]. Cross validation is a popular method for evaluation of the validity of the amount of learning data and recognition precision in the field of machine learning. The 10-fold cross validation is a type of cross validation. An overview of 10-fold cross validation is described below. First, the entire learning data is divided into 10 groups. Next, among these 10 groups, one group is used for recognition, while the remaining nine groups are used for learning. Then, the recognition error rate is calculated. This process is repeated by exchanging groups for recognition and learning for all the ten possible combinations. Finally, an average of these ten recognition error rates is calculated.

As a result, the 10-fold cross validation value is obtained; this value represents the expected recognition rate. If 10-fold cross validation value and discrimination error rate is the same or nearly the same, then the amount of learning data is sufficient.

We evaluated the laughing voice recognition system using the 10-fold cross validation. The 10-fold cross validation rate is approximately 0.03. This suggests that the expected recognition rate is 97%. The discrimination error of the developed recognition system is approximately 0.03 (cf. Section III-C). Because the 10-fold cross validation rate and the discrimination error rate is equal, the amount of learning data was sufficient.

We compared our study with a previous study, i.e., [4]. The study in [4] was selected for comparison because, in [4], only voice data is used without special measuring instruments as they are in our study. On comparing the method in [4] with our proposed system, we found that the recognition rate with our system was 18% better than in [4]; this is depicted in Fig 8. These results indicate the efficacy of the proposed method.

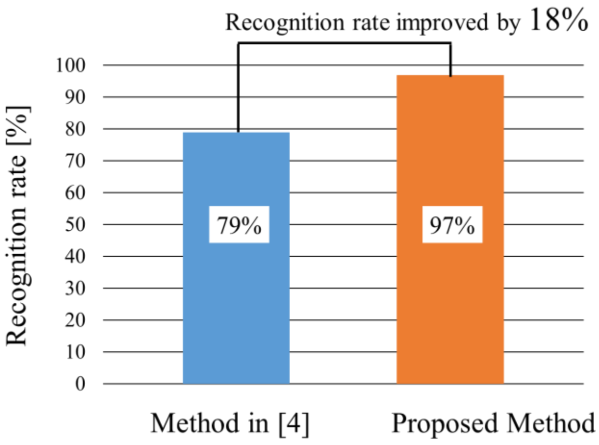

Figure 8. Comparison between recognition rates using method in [4] and the proposed method

IV. Conclusion

Laughing voice recognition based on periodic change recognition has a disadvantage that there is a possibility of incorrectly recognizing a periodic change in the waveform of a voice other than a laughing voice as being that of a laughing voice, i.e., false positive. To overcome this problem, it is necessary to not only use the periodic feature of the waveform in a laughing voice, but also another feature that can be used to discriminate a laughing voice from other voices. Therefore, in this study, we proposed a laughing voice recognition method that incorporates the feature of voice-likeness. In addition, we develop a laughing voice recognition system using the proposed method, and evaluate its performance to confirm the efficacy of our laughing voice recognition method. We were able to recognize correctly a laughing voice with a high success rate of 97%. In comparison with a previous study, the recognition rate improved by 18%. These results indicate the efficacy of the proposed method.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Nos. 15H02936 and 26560129.

References

[1] H.G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, H. Kitano," Human-robot interaction through re-al-time auditory and visual multiple-talker tracking," In 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems, vol.3, pp.1402-1409. IEEE Press, Maui, Hawaii, 2001.

[2] B. A. Christian, H. Ishiguro," Evaluating facial displays of emotion for the android robot Geminoid F," *Affective Computational Intelligence (WACI), 2011 IEEE Workshop on.* IEEE, 2011.

[3] M. Matsumura, Y. Kawabata, R. Suzuki, M. Yoshida," Nonrestrictive and long-term monitoring system of utterance and a laughing voice: Laughometer," *The Journal Acoustical Society of America*, vol.120(5), p.3040, 2006.

[4] S. Tanaka, K. Suzuki," Development of a laughter detection algorithm for robot in response to laughter," In ROBOMECH2015, 1A1-S06. The Japan Society of Mechanical Engineers Press, Kyoto, 2015.

[5] Y. Ariki, S. Kato, T. Takiguchi,“Phoneme recognition based on Fisher weight map to higher-order local auto-correlation,” Transition, 1(1), p.1.

[6] M. Araki,” An illustrated guide to automatic speech recognition,” Kodansha Ltd., 2010.

[7] M. Slaney, “Auditory Toolbox,” Apple Computer, Inc. Technical Report #45, Cupertino CA. (1994).

[8] T. Toyoda, O. Hasegawa, “Extension of higher order local autocorrelation features,” *Pattern Recognition*, vol.40(5), pp.1466-1473 , 2007.

[9] N. Otsu, T. Kurita,” A new Scheme for Flexible and Intelligent Vision Systems,” In Proceedings of IAPR Workshop on Computer Vision --Special Hardware and Industrial Applications--, pp.431-435. IARP Press, Tokyo, 1998.

[10] Y.Ishii, T. Ogitsu, H. Takemura, H. Mizoguchi,” Real-time eyelid open/closed state recognition based on HLAC towards driver drowsiness detection,” In 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp.2449-2454. IEEE Press, Bali, 2014.

[11] J.Kalina,” Classification methods for high-dimensional genetic data,” *Biocybernetics and Biomedical Engineering*, vol. 34(1), pp.10-18, 2014.

[12] Juan D. Rodriguez, Aritz Perez, Jose A. Lozano, “Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32(3), pp.569–575, 2010.