Sound Classification for Hearing Aids Using Time-frequency Images

Koji Abe[*1], Hiroyoshi Masaki[2], Haiyan Tian[3]

[*1]School of Science and Engineering, Kinki University, 3-4-1 Kowakae, Higashi-Osaka 577-8502, Japan

[2]Interdisciplinary Grad. Sch. of Sci. and Eng., Kinki University, 3-4-1 Kowakae, Higashi-Osaka 577-8502, Japan

[3]Graduate School of Engineering, Kobe University, 1-1 Rokkoudai-cho, Nada-ku, Kobe 657-8501, Japan

[*1]koji@info.kindai.ac.jp; [2]masyaki.24@gmail.com; [3]tian.haiyan2006@gmail.com

***Abstract*-This paper presents a method for extracting features of real sound data from time-frequency images. The features are used for a sound classification equipped for hearing aids. As an application of hearing aids in mind, four classes of "classical music", "speech", "multi-talker noise", and "speech in the noise" are prepared in order to classify the input signal of a hearing aid into useful classes. Although there are several possible ways to figure out which class the current input signal belongs to, an approach from image processing is utilized to find out appropriate features because 2D image (time-frequency image) can contain multifaceted information compared to 1D information (waveform or frequency response of sound), and can be regarded as comprehensive data. It is found that eight features are required to meet a certain quality of sound classification according to our investigation. Experimental results of the sound classification by some clustering machines using the proposed features have shown that accuracy of the classification was more than 95% with every clustering machine.**

Keywords- Hearing Aids; Sound Classification; Auditory Scene Analysis; Time-Frequency Image

I. INTRODUCTION

Digital hearing aids have been developed with several functions such as noise reduction, microphone directivity, and feedback cancellation systems [1]. Some of the functions are automatically controlled based on the surrounding environment of a hearing aid user. This means that hearing aids force the users to change some parameters of hearing aids manually according to the environment. However, most of the users are the aged and they wish the manual controls would be improved.

To resolve this issue, sound classification could be useful for hearing aids to know in what sort of environment the users are and how the parameters of hearing aids should be optimized according to the environment. It would be allowed to consider that the most important part to establish any sound classification system is "feature extraction". As a trial of the classification, a classification of artificial sounds into four classes (music, speech, noise, and speech in the noise) was reported, where features of amplitude, frequency, and rhythm were extracted by signal processing techniques [2]. However, this method cannot be applied to hearing aids because the sampling rate for the spectral analysis in this method exceeds 16 kHz, which is the maximum rate in the current hearing aids. Besides, in examining performance of sound classification systems, it is convenient to use artificial sound data because we can clearly define boundaries between classes of speech, noise, and speech in noise by SNR (signal to noise ratio) [2, 3]. However, since SNR for speech cannot be measured to real sound signals and we cannot decide the boundaries strictly, we cannot examine performance of the systems in the case when input signals are sound in real environment. On the other hand, there are reports on classifications of real environmental sound signals [4-6]. The sound signals used in them are noises which arise from traffic, wind, and machines in addition to human voice, so that the classifications are effective for users which often send a time under these environments. However, most of hearing aid users are aged people, and they send much time in shops, stations, and common rooms of welfare homes (e.g., a nursing home, protective institution, etc.) in their daily life. Therefore, if we consider applying a sound classification to hearing aids, we should treat noises which hinder the aged users in listening to human voices and important announcements in a space.

Assuming a sound classification for hearing aids, this paper proposes features for classifying sound signals recorded in real environment, and presents classification of the real sound data using the proposed features. In this investigation, as well as the existing systems (for only artificial sounds) [2, 3], the four classes shown above are prepared before extracting features because the sound classification is expected to be applied to the current configuration of commercialized hearing aids. This means it seems to be insignificant for the sound classification to have so many classes to control the current configuration or only a few functions. If the classification is conducted by using the existing methods [7-9], the classification needs a microphone-array and filter-banks which are not equipped in the current hearing aids, hence they would be not practical ways for enhancing hearing aids. On the other hand, the proposed method does not need a new device to classify environmental sound signals in a hearing aid. The proposed method classifies the signals using time-frequency images. The images are generated by the frequency analysis which is applied in hearing aids, and represent the signals in 3 dimensions of time, frequency, and sound pressure. The proposed features for the classification are extracted from time-frequency images and the classification is conducted by a discriminant machine using the features and classifies the signals into the four classes shown above. 2D images can be found as more comprehensive data, and 2D structure of time-frequency components should be

processed in a cooperative way on the basis of computational auditory scene analysis [10]. To deal with sound data in a sense of auditory scene analysis [11], which is based on *Gestalt psychology*, it seems to be convenient to have time-frequency information that can be seen at once rather than have the waveform or frequency response. This is understandable intuitively because human hearing system listens to any sound in several senses.

II. FOUR CLASSES

The proposed method classifies the sound signal obtained from a hearing aid into four classes of "speech", "noise", "speech in noise", and "music" [12]. At the end of the simulation towards the evaluation of the proposed method, it is necessary to define each of four classes. In this investigation, since all the sound materials are prepared by recording in real environment, the four classes are defined as follows;

- "music": classical music recorded in a quiet room (not an anechoic room),
- "speech": speech by adults recorded in the quiet room,
- "noise": multi-talker noise recorded in a clamorous space (e.g., a classroom in a lecture break),
- "speech in noise": speech by adults recorded in the space,

where "quiet" would be roughly assumed as about 30 – 45 dB on a sound level meter. The classification for the four classes enables optimization of parameters which control the functions of howling canceller and noise reduction in hearing aids. For example, the howling canceller is a function of reducing ingredients which cause howling such as a dull hollow sound. Frequency characteristics of classical music are often similar to a dull hollow sound, i.e., it is difficult for hearing aids to distinguish the two sound signals; hence hearing aids could reduce music sound as well as howling. If hearing aids equip the classification, hearing aids can automatically switch off the howling canceller when the input sound is classified into the music class. Besides, when the hearing aids classify the input signal into "speech", the hearing aids can switch off the noise reduction because the signal has no noise. And then, the hearing aids can apply the noise reduction appropriately by distinguishing "noise" with "speech in noise". It means that the hearing aids can save energy consumption if the hearing aids can classify the four classes.

Ideally, the classification should treat various noises. However, in order to cope with them, hearing aids have to equip a filter for each of them. Under the current situation, where the number of filters is limited due to the small capacity, it is difficult for current hearing aids to treat numbers of filters. Hence, as the first step, only the multi-talker noise (MTN), which is much demanded in the noise-cancelling, is considered in the proposed method.

III. TIME-FREQUENCY IMAGE

The time-frequency image utilized in this research is a spectrogram based on short-time Fourier transform (STFT). Fig. 1 shows an example of time-frequency images consisted of Japanese female speech for 5 seconds. The size of the image is 32×500 pixels (H/W) with 256 gray levels. The vertical axis y corresponds to frequency up to 8 kHz (32 bins, linear scale), and the horizontal axis x is time for 5 seconds (i.e., 1 pixel = 10 ms). The bottom left corner of the image is assigned as the origin such as $(x, y) = (1, 1)$ in this paper.

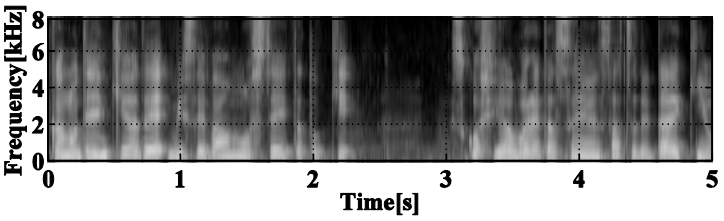


Fig. 1 A time-frequency image consisted of a female speech

The time-frequency image is created as follows. The input signal is digitalized at a sampling rate of 16 kHz. After Hanning window is utilized to 64 points, FFT is applied to the 64 points. The FFT is applied at 75% overlap to the 64 points, i.e., the STFT is conducted to 16 points (1 ms) each. The result of the FFT for 10 ms is added at each frequency in the power spectrum. The sound pressure levels for the 10 ms are calculated by transforming the added value with the base of logarithm 10. The levels which are no more than 35 dBSPL are assumed as gray level 0 on the image, and which are more than 120 dBSPL are as gray level 255. The frequency is linearly divided into 32 matching with the frequency resolution of hearing aids. Hence, from the input signal for 10 ms, the STFT produces 32 pixels which have a gray value out of 8 bits at 1 column in the time-frequency image. After the STFT is applied to the signal for 5 seconds, a time-frequency image is created as a bunch of power spectra for the input signal over 5 seconds such as Fig. 1.

IV. PROPOSED METHOD

First, the features $f_1 \sim f_5$ among the eight proposed features are utilized to classify the data into "music" and "the others". Then, for the images which were classified into "the others", the features $f_4 \sim f_8$ are utilized to classify the data into "speech", "noise" and "speech in noise". The details of the eight features are described below. In order to save the capacity of hearing aids, the calculations for extracting each of the features is sequentially conducted.

A. Pre-processing

In the time-frequency image $O(x, y)$, the mean value μ_x over 256 gray levels except for zeros is sequentially obtained by Eq. (1), and the image is binarized into the black and white image $I(x, y)$ by the threshold μ_x as shown in Eq.(2). Note x and y correspond to the axes of time and frequency, respectively, and the origin point of the image array is defined as the corner at the lower left of the image.

$$\mu_x = \sum_{x=1}^{500} \frac{1}{x}\left(\mu_{x-1}(x-1) + \frac{1}{n_x}\sum_{y=1}^{32} O(x,y) \right) \tag{1}$$

μ_x represents the threshold at a bunch of columns between the first and the x-th columns, and n_x is the number of pixels which are nonzero at the x-th column. Here, $\mu_0 = 0$ and $n_0 = 1$ for convenience of the calculation.

$$I(x,y) = \begin{cases} 1, & O(x,y) > \mu_x \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Fig. 2 shows the binarized image for the time-frequency image shown in Fig. 1.

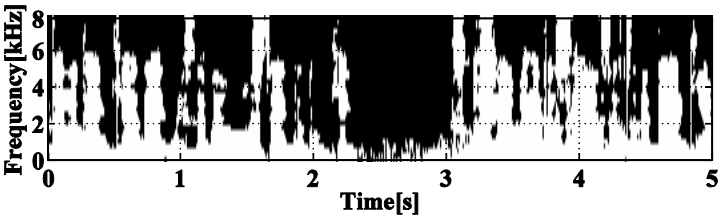


Fig. 2 Binarized image made from Fig. 1

Besides, the histograms of $H_1(y)$ and $H_2(x)$ are calculated by Eqs. (3) and (4), respectively. Fig. 3 and Fig. 4 show the histograms obtained from Fig. 2, respectively.

$$H_1(y) = \sum_{x=1}^{500} I(x,y) \tag{3}$$

$$H_2(x) = \sum_{y=1}^{32} I(x,y) \tag{4}$$

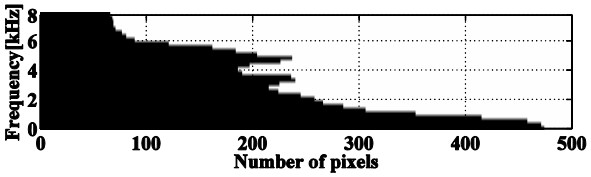


Fig. 3 $H_1(y)$ obtained from Fig.2.

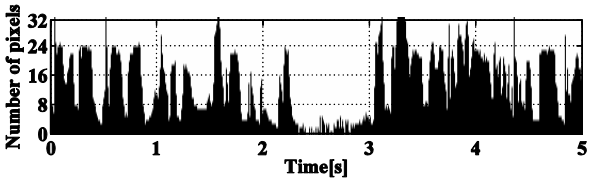


Fig. 4 $H_2(x)$ obtained from Fig. 2

B. Extraction of Features for the Classification

Feature f_1 is defined as

$$f_1 = \sum_{y=1}^{32} g(y) \tag{5}$$

where

$$g(y)=\begin{cases} H_1(y)-H_1(y-1), & \text{condition 1} \\ H_1(y)-H_1(y+1), & \text{condition 2} \\ 0, & \text{otherwise} \end{cases}$$

condition1: $H_1(y-1)<H_1(y), H_1(y)>H_1(y+1)$, and $H_1(y+1)\geq H_1(y-1)$

condition2: $H_1(y-1)<H_1(y), H_1(y)>H_1(y+1)$, and $H_1(y+1)<H_1(y-1)$

It could be said that the feature f_1 calculates how sharp the peaks of the histogram are.

Feature f_2 is defined as

$$f_2=\sum_{x=1}^{500}\sum_{y=1}^{32}K(x,y) \tag{6}$$

where

$$K(x,y)=\begin{cases} 1, & I(x,y)+I(x+1,y)=1 \\ 0, & \text{otherwise} \end{cases}$$

Feature f_3 is defined as

$$f_3=\sum_{x=1}^{500}\sum_{y=SY(x)}^{EY(x)}(1-I(x,y)) \tag{7}$$

where $SY(x)$ and $EY(x)$ are the minimum and the maximum values of y-coordinates at a column x in the case $I(x, y) = 1$, respectively.

Feature f_4 is defined as

$$f_4=\begin{cases} \frac{f_3}{p}, & p\neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where p is the number of segments which are black and vertical lines with the width just 1 in $I(x, y)$.

Feature f_5 is defined as

$$f_5=\frac{1}{m}\sum_{x=1}^{500}\sum_{y=1}^{32}(\mu_x-O_1(x,y))^2 \tag{9}$$

where

$$O_1(x,y)=\begin{cases} \mu_x, & O(x,y)\leq\mu_x \\ O(x,y), & \text{otherwise} \end{cases}$$

and m is the number of pixels which have a grey value higher than μ_x. f_5 represents the variance of grey values in the original image for the pixels which were transformed into the white pixel by the binarization.

f_6 is defined as

$$f_6=\sum_{x=2}^{500}\left|H_2(x)-H_2(x-1)\right| \tag{10}$$

f_7 is defined as

$$f_7=\sum_{x=1}^{500}\sum_{y=1}^{16}I(x,y)-\sum_{x=1}^{500}\sum_{y=17}^{32}I(x,y) \tag{11}$$

f_7 represents difference of the numbers of white pixels between two bunches of lines for 0–4 kHz and 4–8 kHz in the binarized image.

f_8 is defined as

$$f_8 = \frac{1}{n}\sum_{x=1}^{500}\sum_{y=1}^{32}(\mu_x - O_2(x,y))^2 \quad (12)$$

where

$$O_2(x,y) = \begin{cases} \mu_x, & O(x,y) = 0 \\ O(x,y), & \text{otherwise} \end{cases}$$

and n is the number of pixels which are nonzero in the original image.

C. Classification

First, the five features $f_1 \sim f_5$ are utilized as variables in a discriminant analysis for classifying the data into "music" and the other classes (Step 1). Second, the five features $f_4 \sim f_8$ are utilized as variables in a multiple discriminant analysis for classifying the data into "speech", "noise" and "speech in noise" (Step 2). The reason why the classification of the four classes is conducted by the two steps is because sound signals of "music" are composed of various tones by instruments and the other classes are composed of only human voices, i.e., characteristics of "music" are much different from the other classes. Therefore, the classifications by the proposed method could get better performance than a single classification.

Each of the features is normalized as the mean value is zero and variance is one.

V. COMPUTER SIMULATION

A. Preliminaries

All the sound materials were prepared by recording real environments. The recordings were conducted with sound recorder (SONY, PCM-D50), where the dial for changing the degree of collecting the sound is equipped. The recordings were conducted by adjusting the dial watching the meter which displays how small the sound pressure is from 120 dBSPL. The range of the dial is between 1 ~ 10 and the pressure of the recorded sound changes 8 dBSPL as the dial changes 1. According to this function, in recording a sound, the dial or the distance between the sound source and microphone is adjusted as the meter can catch the range from −12 to −9 dB. This adjustment is because the sound pressure is depended on the distance and sound volume under various environments. The dial is fixed for a recording time. And then, the recorded sounds are stored in the recorder as the sampling rate is 44.1 kHz and the quantization bit rate is 16.

Materials for "music" are prepared by recording the signals of classical music from a speaker (Towa Electronics: Olasonic TW-S5). Materials for "speech" are prepared by recording adults' speeches in a quiet room, where the participants aloud read a book. Materials for "noise" are prepared by recording environmental sounds in a restaurant, a lecture room, and a lobby, where MTN is heard. Materials for "speech in noise" are prepared by recording speeches in the environment where MTN is heard. When the sound materials for "music" and "speech" were recorded in a quiet room, the sound level for 10 min. in the quiet room was 34.9 dB by a sound level meter (RION, NL-62).

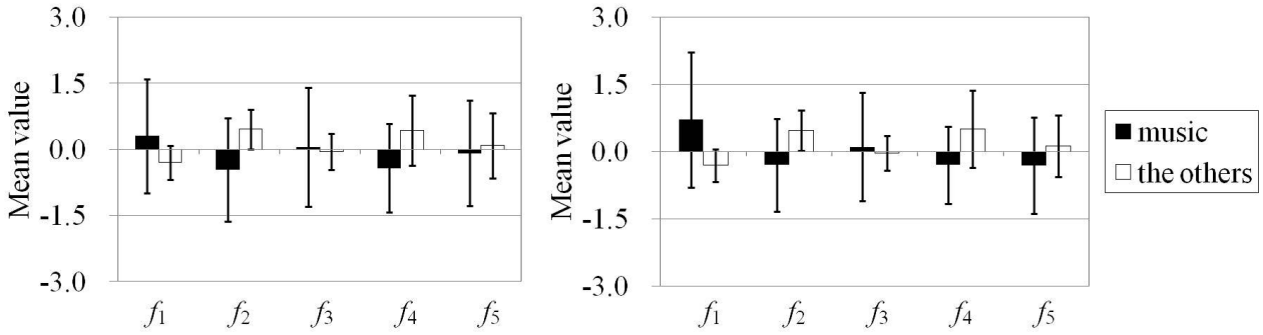
The signals recorded for 5 min. are divided into 60 segments at even intervals hence 60 time-frequency images (for 5 sec.) are created. Table 1 shows the number of data prepared for the experiments. Each of learning and test data for each class was chosen randomly in the class, where the sound sources for all the test data are different from the learning data for the test data.

TABLE 1 DATA SETS OF LEARNING AND TEST DATA IN EACH CLASS

	Learning data	Test data	Sound source
music	1800	600	78
speech	600	600	20
noise	600	600	20
speech in noise	600	600	20

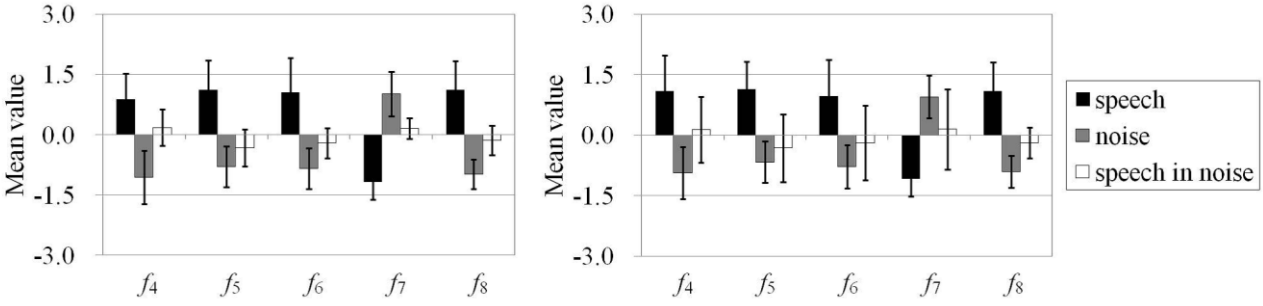
B. Experimental Results

Fig. 5 shows the mean value and the standard deviation of the features in each class of "music" and "the others", where Fig. 5 (A) is results obtained from the learning data and Fig. 5 (B) is results obtained from the test data. And then, Fig. 6 shows the mean value and the standard deviation of the features in each class of "speech", "noise", and "speech in noise", where Fig. 6 (A) is results obtained from the learning data and Fig. 6 (B) is results obtained from the test data. In Fig. 5 and Fig. 6, the horizontal axis represents $f_1 \sim f_5$ and $f_4 \sim f_8$, and the vertical axis represents values of the features. From the figures, we can confirm that there is significant difference between each class.



(A) Learning data (B) Test data

Fig. 5 Mean value and standard deviation of the features in each of "music" and "the others"



(A) Learning data (B) Test data

Fig. 6 Mean value and standard deviation of the features in each of "speech", "noise", and "speech in noise"

Next, according to the classification method shown above, experiments of the classification for the 4 classes were conducted. As a discrimination machine for the classification, linear discriminant analysis (LDA), discriminant analysis by Mahalanobis' generalized distance (MD), and a neural network (NN) were applied. The neural network was a feedforward neural network which consisted of three layers in both steps of the classification. In Step 1 of the classification, the input and output layers of the NN had 2 nodes and the hidden layer had 3 nodes. In Step 2 of the classification, its input, hidden, and output layers had 3, 10, and 3, respectively.

Table 2 (A) shows experimental results for Step 1 of the classification, where performance of the classification is represented by Recall and Precision. Table 2 (B) shows the number of images classified into each class to the number of input images in each class. *Recall* and *Precision* were calculated by

$$Recall = \frac{X_{h \cap c}}{X_h} \times 100 \tag{13}$$

$$Precision = \frac{X_{h \cap c}}{X_c} \times 100 \tag{14}$$

where X_h is the number of images in the image group h for conducting the discrimination, X_c is the number of images in the image group c discriminated by the proposed method, and $X_{h \cap c}$ is the number of images which belong in $h \cap c$.

TABLE 2 RESULTS OF THE DISCRIMINATION BETWEEN "MUSIC" AND "THE OTHERS"

(A) Recall and Precision

Tool	Class name	*Recall* (%)	*Precision* (%)
LDA	music	89.3	91.2
	the others	97.1	96.5
MD	music	99.5	63.0
	the others	80.5	99.8
NN	music	95.5	93.5
	the others	97.8	98.5

(B) The number of images

Tool	Output / Input	music	the others
LDA	music	536	64
	the others	52	1748
MD	music	597	3
	the others	351	1449
NN	music	573	27
	the others	40	1760

Similarly, Table 3 (A) and (B) shows experimental results and the numbers for Step 2 of the classification.

TABLE 3 RESULTS OF THE CLASSIFICATION BETWEEN "SPEECH", "NOISE", AND "SPEECH IN NOISE"

(A) Recall and Precision

Tool	Class name	*Recall* (%)	*Precision* (%)
LDA	speech	98.7	100.0
	noise	97.0	95.4
	speech in noise	95.3	95.7
MD	speech	100.0	97.6
	noise	95.7	98.3
	speech in noise	96.3	96.2
NN	speech	100.0	98.7
	noise	97.2	97.7
	speech in noise	96.3	97.1

(B) The number of images

Tool	Input \ Output	speech	noise	Speech in noise
LDA	speech	592	0	8
	noise	0	582	18
	speech in noise	0	28	572
MD	speech	600	0	0
	noise	3	574	23
	speech in noise	12	10	578
NN	speech	600	0	0
	noise	0	583	17
	speech in noise	8	14	578

VI. DISCUSSIONS

Table 2 (A) shows that the proposed method can correctly discriminate the sound in Step 1 of the classification in more than 90% of probability. From the view that a hearing aid manufacturer currently aims at 75% of the success ratio in the classification and the numbers of sound sources and data are sufficient accordingly, the experimental results in Step 1 have shown performance of the proposed method is high enough.

In Table 2 (A), *Precision* for "music" in MD is significantly lower than the others. To examine this source, a parameter D is prepared. D is difference between the Mahalanobis distance from one of the data to "music" (D_{music}) and the distance from the one to "the others" (D_{others}), i.e., D is defined as $D = D_{others} - D_{music}$. From the experimental results shown in Table 2 (B), D was obtained from every of 1449 images which were correctly classified into "the others" in MD, and the mean value of them is defined as D_1, and the standard deviation is defined as S_1. Similarly, D_2 is defined as the mean value obtained from 281 images which were incorrectly classified into "music" in MD (its standard deviation is defined as S_2), and D_3 is defined as the mean value obtained from 19 images which were incorrectly classified into "music" in all the discrimination machines (its standard deviation is defined as S_3). Table 4 shows these values. In Table 4, comparing D_2 with D_1 and D_3, D_2 is notably close to zero. Therefore, considering the value of S_2, we can deduce the 281 images would cluster around the boundary of the two classes. And then, Fig. 5 shows that the standard deviations of $f_1 \sim f_5$ of the images in "music" are higher than "the others", and the mean value and the standard deviation obtained from absolute values for variance-covariance of $f_1 \sim f_5$ extracted the images in "music" were 1.183 and 0.326, respectively. On the other hand, the two values obtained from the images in "the others" were 0.230 and 0.152, respectively. From these results, we can deduce the images which belong in "the others" and are located around the boundary shifted to the area where the images in "music" are distributed; because MD standardizes the distribution of two classes before the distance calculation. The reason that the standard deviations for the features in "music" are higher than "the others" is because image patterns of signals extracted from music are far more various than human voice, hence there would be significant difference of the distribution area between "music" and "the others". Therefore, MD would be not suitable as a discrimination machine for Step 1 of the classification.

TABLE 4 $D_1 \sim D_3$ AND $S_1 \sim S_3$ IN THE DISCRIMINATION BETWEEN "MUSIC" AND "THE OTHERS"

	$i = 1$	$i = 2$	$i = 3$
D_i	-6.368	2.160	8.139
S_i	5.679	2.513	4.360

Referring to Table 3 (A), we can confirm all the discrimination ratios are more than 95% in Step 2 of the classification. As well as Step 1, the experimental results in Step 2 have shown performance of the proposed method is high enough.

In Table 3 (B), most of errors occur between "noise" and "speech in noise" in every machine. Listening to the sound data of "noise" for 7 images which were incorrectly classified into "speech in noise" in every machine, it was found that sudden noises which are not MTN such as the sound of laughter, clapping hands, or dropping something on the floor are included in the data. Since the proposed features were designed by considering characteristics of MTN, the proposed method cannot classify composite noises which include strong noises other than MTN into "noise" correctly. Therefore, as the next trial, it is necessary to cope with the composite noises and to increase type of noise.

VII. CONCLUSION

This paper has presented features of sound data for a sound classification equipped for hearing aids. The features have been extracted by using image processing techniques to time-frequency images. As an application of hearing aids in mind, the four classes of "music", "speech", "multi-talker noise" and "speech in the noise" are prepared to classify the input signal. Experimental results of the sound classification by some clustering machines using the proposed features have shown that accuracy of the classification was more than 95% with every clustering machine.

As future works, it is necessary to recognize sudden noise in real environment, and then enhance sound quality of announcements via hearing aids in stations.

REFERENCES

[1] J. M. Kates, *Digital Hearing Aids*, Plural Publishing Inc., 2008.

[2] M. Büchler, S. Allegro, and N. Dillier, and S. Launer, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Applied Signal Processing*, Vol. 2005, pp. 2991-3002, 2005.

[3] K. Abe, H. Sakaue, T. Okuno, and K. Terada, "Sound classification for hearing aids based on time-frequency image processing," In *Proc. of 2011 IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing*, pp. 719-724, Victoria, BC, Canada, 2011.

[4] M. A. Haque, S. Cho, and J. Kim, "Audio classification and scene recognition and for hearing aids." In *Proc. of IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 860-863, Kobe, Japan, 2005.

[5] L. Luc, G. Christian, G. Wail, A. Tyseer, and O. Hisham, "Adaptive environment classification system for hearing aids," *J. Acoust. Soc. Am.*, vol. 127, no. 5, pp. 3124-3135, 2010.

[6] J. Xiang, M. F. McKinney, K. Fitz, and T. Zhang, "Evaluation of sound classification algorithms for hearing aid applications," In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 185-188, Dallas, TX, USA, 2010.

[7] J. Wei, L. Du, Z. Chen, and F. Yin, "A new algorithm for howling detection," In *Proc. IEEE International Symp. on Circuits and Systems*, vol. 4, pp. IV-409-IV-411, Beijing, China, 2003.

[8] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching kalman filter," *IEICE Trans. Info. and Syst.*, vol. E91-D, no. 3, pp. 467-477, 2008.

[9] Tao Yu and John H. L. Hansen, "An efficient microphone array based voice activity detector for driver's speech in noise and music rich in-vehicle environments," In *Proc. 2010 IEEE Int. Conf. on Acoustics Speech and Signal Processing*, vol. 4, pp. 2834-2837, Dallas, TX, USA, 2010.

[10] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982-994, 2007.

[11] A. Bregman, *Auditory Scene Analysis*, The MIT Press, Cambridge, MA, USA, 1994.

[12] A. Schaub, *Digital Hearing Aids*, Thieme Medical, New York, NY, USA, 2008.

Koji Abe received his B.S. and M.S. degrees from Kogakuin University, Japan, in 1996 and 1998, respectively. And, he received his Ph.D. degree in Engineering from Kanazawa University, Japan, in 2001.

He affiliated in the Institute for Image Data Research, University of Northumbria at Newcastle, UK, as an honorary research fellow in 2002. He was an assistant professor in the Department of Information and Computer Science, Kanazawa Institute of Technology, Japan, in 2003. He was a lecturer in the Department of Informatics, School of Science and Engineering, Kinki University, Japan, in 2006. He has been an associate professor in Kinki University since 2010.

His research interests include pattern recognition, medical image processing, CBIR, multimedia database, and artificial intelligence. He is a member of IEEE, IEICE (Japan), IEEJ (Japan), and IPSJ (Japan).

Hiroyoshi Masaki received his B.E degree from Kinki University in 2014. He is currently a master student at Graduate School of Science and Engineering, Kinki University.

His research interests include pattern recognition and image processing. He is a member of the IEICE.

Haiyan Tian received her B.S., M.E., and Ph.D. degrees from Chongqing University in 1992, 1995, 2002, respectively.

She was a lecturer in the University of Electronic Science and Technology of Xi'an in 1997. She was an associate professor in Chongqing University in 2003. In addition, she was affiliated as a research staff with the Institute of Fluid Science, Tohoku University, Japan. Currently, she is affiliated with Graduate School of Engineering, Kobe University.

Her research interests include signal processing, network security, and telecommunication network. She is a member of the IPSJ.