KEER2010, PARIS | MARCH 2-4 2010
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

COMPONENTS RELATED TO SENSORIAL DECISION USING ANN PREPROCESSED WITH AURAL CHARACTERISTICS

Yutaka Suzuki [*a], Takaya Kato [a] , Asobu Hattori [b] and Osamu Sakata [a]

[a] *University of Yamanashi, Japan*
[b] *Tokyo Metropolitan Industrial Technology Research Institute, Japan*

ABSTRACT

It is necessary to develop new signal processing systems to confer on robots "the ability to behave in a *kansei* manner" in response to sounds. In this study, the components of sensorial decision in response to sound are investigated using an acoustic diagnostics system that outputs a sensorial decision modelled after the mechanism of processing in humans. The loudness function is applied to the loudness of the sound, and mel is applied to the pitch. The averaged result of the frequency spectrum is input into an artificial neural network (ANN). Samples with only slight differences are used as diagnostic subjects to examine the ability to differentiate between the sounds. The ratio of correct answers is increased by combining mel and the loudness function, and there is an optimal loudness of sound to obtain the correct answer. Only parts of the aural characteristics of humans are utilized in this study, but the results indicate that it is effective to consider such nonlinear characteristics.

Keywords: *acoustic diagnostics, neural network, aural characteristic*

1. INTRODUCTION

There is increasing research interest in the development of robots capable of sensorial interaction, a field known as *Kansei* Engineering, as evidenced by the 2009 *Kansei* Engineering International Conference. which featured "*KANSEI* Robotics" [1]. In the 20th century, various types of industrial robot were developed to reduce labour. However, only the performance of such robots in performing specialised physical tasks was evaluated.

[*] **Corresponding author**: shimokato 1110 Chuo, Yamanashi, 409-3898, Japan. yutakas@yamanashi.ac.jp

However, robots now interact closely with humans, and studies to develop robots capable of acting as partners are underway. For example, with the aging of society there is a demand for robots capable of assisting in home care. In such applications, communication ability is desirable in addition to the ability to perform physical tasks [2]. In addition, commercial products developed to ease stress on people through communication, such as paro [3] and Ifbot [1, 4] have attracted attention. To further expand the communication abilities of such robots, further studies on "the ability to read sensorial behaviours of humans" as well as the "ability to behave in a sensorial manner" are required, and therefore there is increasing research interest regarding different aspects of sensorial behaviour [1, 2].

In this study, we have focused on acoustic sense, which is an important information receptor, to develop a signal processing system to confer on robots "the ability to behave in a sensorial manner" in response to sounds. Previously, we examined sensorial evaluation" and physical characteristics using acoustic and aural signals [5 – 7] to determine the effectiveness of analysis methods, such as loudness, sharpness, and roughness [8]. However, these analysis methods must be improved for situations in which the subject sound is very small, such as differences in the tone quality of musical instruments manufactured in the same lot and tuned by the same person, or differences in sound with use of different audio amplifiers. Therefore, we examined an acoustic diagnostics system modelled after the acoustic process of humans, and searched for components of this sensorial decision system.

In this study, an artificial neural network (ANN) was used to output the sensorial evaluation of humans [7, 9]. The advantage of this method is that the decision of an experienced expert can be learned as an instruction signal to attain the ability to recognize what is learned. In analysis of acoustic signals, frequency analysis is often performed first, and then an ANN is used [10]. However, incorporating loudness analysis that is effective in sensorial evaluation has not been taken into consideration. Loudness analysis takes loudness characteristics (effect that takes place in the outer hair cells of the human auditory organ, where the amplification would be larger with smaller stimuli) into consideration, and is very useful when differentiating between soft sounds. In a previous study, we processed the frequency analysis results according to the loudness characteristics before inputting into an ANN and found improvements in the ability to differentiate the slightest differences in sound [11]. In the present study, we also considered nonlinear perception characteristics related to the pitch of a sound, and investigated its effects.

Our ultimate goal is to develop a signal processing system that can indicate sensorial evaluation, such as slight differences in tone quality. However, we predicted that it would be difficult to evaluate the nonlinear effects of aural characteristics using a subject sound with many potential evaluation axes, such as played back music, at such an early stage in the research. Here, the sound source was chosen based on whether the instruction signal necessary for learning of ANN can be obtained accurately, and the differences between tone quality are small and are difficult to differentiate; we decided to use the tapping sound of a hammer. The loudness and pitch of the sound were preprocessed according to aural characteristics, and were input into the ANN to be differentiated. A high ratio of correct answers was obtained. The results are present in a later section.

2. SIGNAL PROCESSING MODELLED AFTER HUMAN AURAL CHARACTERISTICS

The flow of the acoustic diagnostics system modelled after the sound processing process of humans is shown in Fig. 1. The processing of sound at the basal membrane of the inner ear is equivalent to frequency analysis of acoustic signals [12]. Therefore, we performed FFT analysis and matched the results to the loudness. Next, we considered the aural characteristics of the pitch of the sound, and the frequency band was segmented nonlinearly to match the number of input layer neurons of ANN. The data were averaged, input into the ANN, learned and differentiated.

2.1. Loudness function

The relationship between the loudness L and the intensity of the sound I is proximate in the following loudness function [13 – 16].

$$L = c\left(I^n - I_c^n\right) \text{..........................(1)}$$

Here, c, n and I_c are constants, and their values differ depending on the frequency. Threshold of hearing I_c is determined as the sum of effective physiological murmur and environmental noise; when environmental noise becomes large, I_c also becomes large, and I_c for neurosensory deafness is also large [15]. As the purpose of this study was to mimic the function of human perception recognition, the values provided in the loudness level contour (ISO226) [13] were used as constants c, n and I_c. However, as the masking effect and others are not considered in the present system, these values are treated only as approximations.

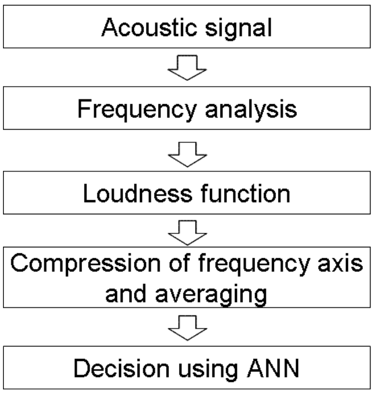

Figure 1: Flow of acoustic diagnostics system developed based on sound processing of humans

2.2. Compression of frequency axis

The resonance and perception of the pitch of the sound in the basal membrane shows nonlinear characteristics. Fletcher proposed scales based on the concept of acoustic sense filters, such as the Bark scale and ERB scale [17] but in this study, the mel scale was obtained from perception characteristics to match the number of input layer neurons of ANN. Mel is a characteristic that indicates the relationship between the intervals of sensorial pitch of the sound and the frequency, which was found by experiment [18] and is approximated in equation (2) , where sound with frequency 1 kHz is 1000 mel [19].

$$Mel = 2595\log(1 + f/700) \text{ [mel] ....(2)}$$

Here *f* is frequency, and it is given in units of Hz. A sound with 2000 mel has a frequency of approximately 3 kHz, and a sound with 500 mel is 400 Hz, which does not correspond to double and half of the frequency 1 kHz equivalent to 1000 mel. This mismatch becomes larger as the frequency becomes higher, *e.g.*, a sound with f = 10 kHz is only equivalent to approximately 3000 mel.

2.3. Artificial neural network (ANN)

ANN is a network in computer simulation modelled after the mechanisms of the human brain. By letting ANN learn the instruction signal ("normal" or "broken"), which act as a sound identifying standard, the ANN can identify between similar data. In this study, a feed-forward type ANN comprised of 3 layers—the input, middle and output layers—was used, and back-propagation was used as a learning rule. A sigmoid function was used for calculation for each neuron. In this study, effects such as roughness and intensity fluctuation were not considered, as averages are used in a feed-forward type ANN, without any consideration of time variation in the tap tone.

3. DIAGNOSTICS

The subjects of diagnostics were chosen based on the following criteria: accurate instruction signal can be obtained, the difference between tone qualities is small, and difficult to differentiate, the differences are seen within the audible band, and the samples are readily available. By trial and error, we decided to use 250-mL bottles of carbonated drinks. Forty bottles were used as samples. Of these, the top part of the 20 bottles was cracked to create broken bottles. Normal bottles and broken bottles were differentiated based on the tap tone and echo (hereafter referred to simply as tap tone) [20]. To maintain the strength of hammering on the diagnostic subjects, an inspection table was produced using a wooden stick and metal spring. The samples were collected in a soundproof room with a noise level of ≤ 22 dBA. Samples were obtained using condenser microphones (SONY ECM-DM5P), and were recorded onto a PC using a microphone amplifier and A/D converter (Roland UA-25) with 24 bit – 48 kHz sampling.

4. RESULTS AND DISCUSSION

4.1. Characteristics of the subject tap tone

Tap tone is a damped oscillation, and changes depending on the differences between individual bottles as well as depending on the location where the bottle is hit. The duration of tap tone from 95% of maximum amplitude to 5% is approximately 0.1s. To include the durations of all tap tones examined, FFT analysis was performed on approximately 0.17 s of each sample. Figure 2 shows a sample of a result displayed in the amplitude spectrum. Large spectra were observed around 4 kHz and 7 kHz, and components in other areas were small.

Figure 3 shows the results displayed with the sound pressure level of the largest spectrum in all examined tap tones assumed to be 80 dB SPL, applied in the loudness function of equation (1), and with frequency axes converted using linear, mel, and logarithmic scales. In this study, this sound pressure level will be referred to as the input level of the tap tone, with units in dB. With the amplitude spectrum in Fig. 2 and the results shown in Fig. 3 (a) with loudness display, we can see how a small amplitude spectrum in Fig. 3(1) has become relatively large, and the spectrum is observed even around 10 kHz. This indicates that when ANN is used to differentiate the sounds, even a tiny signal could contribute to the decision. In addition, it can be seen that the parameters are different depending on the frequency band, and that there is a frequency band with a large amplification factor. For example, in amplitude display, the peak value at 7 kHz is larger than that at 4 kHz, but in loudness display the peak at 4 kHz is larger.

Comparison of the differences in spectrum display due to different frequency axes indicates how the low frequency region is amplified in mel compared to the linear scale, and the high frequency region is reduced. The ratio of expansion in mel is small compared to the logarithmic scale, and therefore we can see that the spectral distribution of mel is between that of linear and logarithmic scales. Large spectra observed at 4 kHz and 7 kHz are located on the right side, and can be heard as high-pitched sounds, which matches the perception sensitivity of humans.

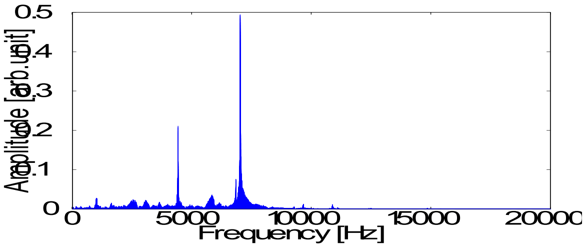

Figure 2: Sample of amplitude spectrum of tapping sound

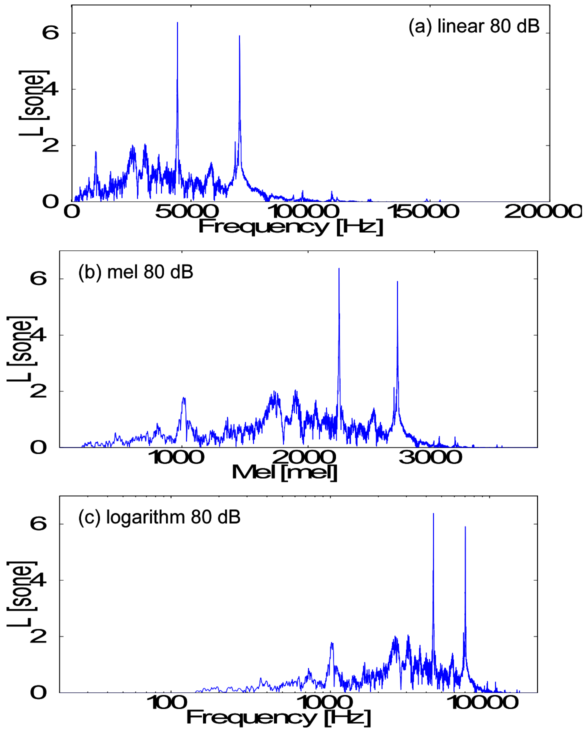

Figure 3: Spectral distribution (input level 80 dB) with different frequency axes. The results with (a) linear, (b) mel scale and (c) logarithmic frequency axes.

Using the loudness function, any spectral component below a threshold becomes 0, but depending on the frequency axis, the amount of component that becomes 0 changes. When the audible frequency bands between 20 Hz and 20 kHz are equally divided into 100, and in each divided part, the values are averaged to be input into the ANN, 23% of the inputs on the linear scale become 0 with a sound pressure level of 80 dB SPL. In mel and logarithmic scales, 9% and 30% become 0, respectively. These observations indicate that when mel is used, the number of inputs contributing to the decision increases.

4.2. Examination results and discussion

Ten bottles were taken from each of 20 normal bottles and 20 broken bottles. Twenty tap tones were taken from each of these selected bottles, for a total of 400 tap tones as learning samples. The 400 tap tones from the remaining 20 bottles were used as samples for evaluation, and we determined the ratio of correct answers. Using the segment averaging technique, we input 100 data per tap tone. Thus, the number of input layer neurons was 100, and we set the number of middle layer neurons to 5. The number of output layer neurons was 1, and was either Normal or Broken. The learning of ANN differs depending on the initial weight, and therefore we provided 5 patterns of random numbers for each round of processing, then learning and evaluation took place. The criteria to finish learning was for the sum of squares error of the instruction signal and learning set to reach 0.05; if it did not reach 0.05, the learning ended once the number of times of iterations reached 100,000 times. We were unable to obtain the ratio of correct answers from an experienced expert with this hammering test, because there is no such expert. Even for the person who carried out the experiment, who became very familiar with the tap tone, it was difficult to determine whether there was damage to the bottle or not by listening to the sound when blindfolded.

Figure 4 shows the ratio of correct answers in the evaluation set when the scale used to convert the frequency axis was changed to linear, mel and logarithmic. The errors on the vertical axes in the figures are within the 95% confidence interval of the ratio of correct answers that depend on the initial weight.Although not shown in this paper, when we simply divide the frequency band of the amplitude spectrum of the audible frequency band in FFT analysis into 100 equal parts and input into the ANN using the segment averaging technique, the ratio of correct answers was 73.3%. As shown in Fig. 4, when the amplitude spectrum was treated with the loudness function, the ratio of correct answers was 85% with an input level of 80 dB. In addition, we can also see how as the input level was increased, the ratio of correct answers increased, reaching the maximum value, followed by a decrease. This indicated that there is an optimal loudness of the sound to make a decision. When mel was used as the frequency axis, we were able to obtain an even higher ratio of correct answers, *i.e.*, 91.3% with an input level of 70 dB. As mentioned above, the number of inputs contributing to the decision increases when the segment averaging technique is performed with mel, and we believe this effect has a significant contribution. When the logarithmic scale was used as the frequency axis, the ratio of correct answers decreased in general, and the changes according to the input level became small. We believe this is because when conversion takes place using logarithmic axis, the number of low frequency components input into the ANN increases, and also because many of the frequency components from the subjects used in this study were high. Based on the above, we were able to confirm that the ratio of correct answers becomes high when sensorial decision is mimicked with consideration to mel for the frequency, and by applying loudness function to the loudness of the sound, and that there is an optimal loudness of sound.

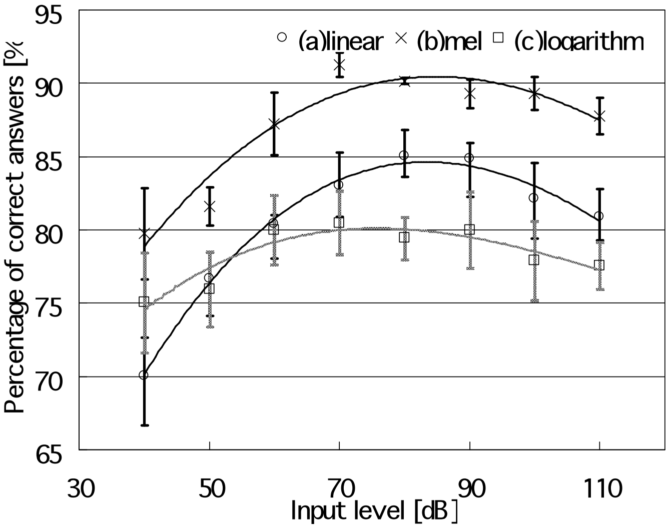

Figure 4: Diagnostic results

5. CONCLUSION

By examining an acoustic diagnostics system that outputs a sensorial decision modelled after the mechanism of processing in humans, the components of sensorial decision on sound was investigated. In this study, the loudness function was applied to the loudness of the sound, and mel was applied to the pitch. The averaged result of the frequency spectrum was input into the ANN. Samples with only slight differences were used as diagnostic subjects to differentiate the sounds. We found that the ratio of correct answers was increased by combining mel and the loudness function, and there was an optimal loudness of sound to obtain the correct answer. Only parts of the aural characteristics of humans were utilized in this study, but the results indicated that it is effective to consider such nonlinear characteristics.

ACKNOWLEDGMENTS

We would like to express our gratitude to Mr. Itsuo Ishikawa, a former professor at the University of Yamanashi for his advice and cooperation. This research was supported in part by a Grant-in-Aid for Scientific Research (B) 20300076).

REFERENCES

1. Kansei Engineering International, *Journal of Japan Society of Kansei Engineering*, Vol. 8, No. 1, 2009.
2. K. Tomiyama, Virtual KANSEI: KANSEI for robot [in Japanese], *J. Acoust. Soc. Japan*, Vol. 64, No. 8, pp. 481-488, 2008.

3. T. Shibata, Ed., Human Interactive Robots for Psychological Enrichment, *Special Isshue, Proceedings of the IEEE*, Vol. No. 92, 2004.
4. Business Design Laboratory Co., Ltd., *Communication Robot Ifbot*, <http://www.business-design.co.jp/product/ifbot/>, [Accessed 2010 January 10].
5. J. Marozeau, A. Cheveigne, S. McAdams, S. Winsberg, The dependency of timbre on fundamental frequency, *J. Acoust. Soc. Am*, Vol. 114, No. 5, pp. 2946-2957, 2003.
6. W. Klippel, Multidimensional relationship between subjective listening impression and objective loudspeaker parameters, *Acoustica*, Vol. 70, pp. 45-54, 1990.
7. G. Onishi, Hidemi, I. Kimura, T. Miho, H. Yamada, A Kansei model for musical timbre using neural networks [in Japanese], *Transactions of the Japan Society of Mechanical Engineers. C, N*, Vol. 652, No. 66, pp. 3977-3983, 2000.
8. H. Fastl, E. Zwicker, *Psychoacoustics- Facts and Models*, Springer, 2006.
9. K. Suzuki, H. Yamada, S. Hashimoto, A similarity-based neural network for facial expression analysis, *Pattern Recognition Letters*, Vol. 28, No. 9, pp. 1104-1111, 2007.
10. M. Sansalone, Impact-echo signal interpretation using artificial intelligence, *ACI Material*, Vol. 89, No. 2, pp. 178-187, 1992.
11. Y. Suzuki, A. Hattori, T. Kato, I. Ishikawa, O. Sakata, Signal Processing for Sensuous Judgment by Neural Network, *The International Conference on Kansei Engineering and Emotion Research KEER2007*, B-7, 2007.
12. B. C. J. Moore, *An Introduction to the Psychology of Hearing*, ACADEMIC PRESS, pp. 19-28, pp. 72-74, 2003.
13. ISO226, Acoustics-Normal equal-loudness-level contours, 2003.
14. J. J. Zwislocki, R. P. Hellman, On the psychophysical law, *J. Acoust. Soc. Am*, Vol. 32, p.924, 1960.
15. J. P. A. Lochner, J. F. Burger, Form of the loudness function in the presence of masking noise, *J. Acoust. Soc. Am*, Vol. 33, pp. 1705-1707, 1961.
16. L. E. Humes, J. F. Jesteadt, Models of the effects of threshold on loudness growth and summation, *J. Acoust. Soc. Am*, Vol. 90, pp. 1933-1943, 1991.
17. M. Akagi, Auditory filter and its modeling [in Japanese], *J. Institute of Electronics, Information, and Communication Engineers*, Vol. 77, No. 9, pp. 948-956, 1994.
18. S. S. Stevens, J. Volkmann, and E. B. N. Newman. A scale for the measurement of the psychological magnitude pitch, *J. Acoust. Soc. Am.*, Vol. 8, No. 1, pp. 185–190, 1937.
19. G. Fant, *Speech Sound and Features*, MIT Press, Cambridge, MA., 1973.
20. Y. Suzuki, A. Hattori, T. Kato, I. Ishikawa, Auscultating type health monitoring equipment modeled on the human loudness function, *Japan Society for Welfare Engineering*, Vol. 9, No. 2, pp. 19-24, 2007.