AN INVESTIGATION TOWARDS VERBALLY CONTROLLABLE GRAPHIC EQUALIZER FOR SINGING VOICES

Seiya Masuda, Eriko Aiba**, and Tetsuro Kitahara**

*College of Humanities and Sciences, Nihon University
Setagaya-ku, Tokyo, Japan, {masuda,kitahara}@kthrlab.jp
**Graduate School of Informatics and Engineering, University of Electro-Communications
Chofu City, Tokyo, Japan, aiba.eriko@uec.ac.jp

ABSTRACT

This paper presents an investigation of the relationship between the equalization of a singing voice and its verbal evaluation. Participants were asked to listen to sound stimuli generated with different equalization settings and evaluate their timbres with respect to 10 words (*warmness, presence, showiness, muddiness, mellowness, softness, brightness, lightness, thickness,* and *clearness*). The mapping between the equalization settings and verbal evaluations were obtained with multivariate linear regression. The obtained results support findings described in know-how books for hobby musicians: for example, *warmness, muddiness, mellowness,* and *softness* are enhanced when a low pitch range is boosted while *showiness, brightness, clearness,* and *presence* are enhanced when a high pitch range is boosted. We also implemented a prototype of a system that estimates an equalization setting from a verbal evaluation vector.

1. INTRODUCTION

The recent development of multimedia and network technologies has enabled non-professional musicians to easily publish musical works. In fact, many non-professional singers record their singing performances, either of their original songs or as covers of songs written by professional musicians, and publish them on web-based video sharing services such as YouTube[1] and Nico Nico Douga[2]. To publish singing recordings, however, non-professional musicians have to go through various processes. To make recorded voices closer to what they desire, for example, they often have to change the timbre of the recorded voices using a graphic equalizer (GEQ). A GEQ has many (typically 10 to 30) controllers, each of which boosts or cuts the tone at a specific frequency range. With a GEQ, they can control the timbre finely, but accordingly it is difficult for people without the know-how to express the timbre that they desire as controller settings.

One approach for solving this problem is to enable them to control a GEQ using verbal expressions. We often express a timbre verbally, such as “bright sounds” and “warm sounds.” It is, therefore, considered that the relationship of timbres and their verbal expressions can be easily understood, even by non-professional musicians.

The relationships between timbres and their verbal expressions have been previously investigated. Bloothooft et al. [1] investigated the timbres of voices sung by various singers using adjective pairs describing the timbres such as light—dark. Yamauchi et al. [2] investigated the effects of materials of flutes and cellos on their timbres, also using adjective pairs describing the timbres. Jan Stepanek [3] investigated verbal attributes describing timbral dissimilarities of sounds of the violin and pipe organ through listening tests (called a bottom-up approach), as well as collecting musicians’ opinions (called a top-down approach). However, no attempts to investigate the timbres of singing voices controlled with a GEQ and their verbal evaluations have been made. Also, a technique has been proposed for automatic equalization [4], but it did not focus on equalization based on verbal expressions.

In this paper, we investigate the mapping between a GEQ setting space and a verbal evaluation space for singing voices. The GEQ setting space is a multi-dimensional space in which each dimension represents a boost/cut level at each frequency band, while the verbal evaluation space is a multi-dimensional space in which each dimension represents sound evaluations with respect to each sound expression word (such as *brightness* and *warmness*). To construct a map between them, we conducted a listening test in which we asked participants to evaluate how strongly they feel characteristics expressed by each word in sound stimuli equalized with different GEQ settings.

2. METHOD

2.1. Sound Stimuli

We recorded a singing voice sung by the first author (male, age: 21) a cappella. The sung melody was taken from “*Yoake to Hotaru*” composed by Nabuna. The singing voice input from a microphone (AT4040, Audio Technica) was transmitted to a PC (Sepctre 13-v007TU, Hewlett-Packard) through a USB-connected audio interface (UM2, Behringer) with a sampling rate of 48 kHz, and then saved in

[1] https://www.youtube.com/
[2] https://www.nicovideo.jp/

the WAV format with 16-bit linear encoding. For equalization, we used a 10-band software GEQ built in digital audio workstation software (Cubase 7.5, Steinberg). The center frequency of each band is 31.5 Hz, 63 Hz, 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, 8 kHz, and 16 kHz. Because the lowest fundamental frequency of our sound stimuli was 119 Hz, the bands with a center frequency of 31.5 Hz and 63 Hz were excluded from equalization. To reduce the participants' burden, we bound each pair of adjacent bands (125 Hz and 250 Hz, 500 Hz and 1 kHz, 2 kHz and 4 kHz, 8 kHz and 16 kHz). Thus, we treated this GEQ as a 4-band GEQ. Below, a band means a bound band (for example, Band 1 is a pair of 125 Hz and 250 Hz). The following 27 stimuli were prepared:

- Sound without any equalization (1 stimulus)
- Sounds in which only one band was boosted by 12 dB (4 stimuli)
- Sounds in which only one band was cut by 12 dB (4 stimuli)
- Sounds in which two bands were boosted by 12 dB (6 stimuli)
- Sounds in which one band was boosted by 12 dB and another band was cut by 12 dB (12 stimuli)

To reduce the effects caused by the sound pressure, we normalized all sound stimuli so that their root mean square of the amplitude is equal. Above, the sounds in which two bands were cut by 12 dB were excluded because their relative boost/cur levels were the same as those for the sounds in which two bands were boosted.

2.2. Selection of Sound Expression Words

We extracted words expressing the effects of GEQ from know-how books for hobby musicians [5–7]. Then, we selected the 10 most frequently appearing words as follows: *warmness*, *presence*, *showiness*, *muddiness*, *mellowness*, *softness*, *brightness*, *lightness*, *thickness*, and *clearness*.

2.3. Procedure of Listening Test

After receiving the instructions about the experiment, participants listened to each of the sound stimuli and evaluated how strongly they felt characteristics expressed by each of the above sound expression words in the stimulus. The detailed procedure for each stimulus is as follows:

Step 1 Listen to the pre-equalization stimulus once

Step 2 Rest for 3 seconds

Step 3 Listen to the equalized stimulus three times.

Table 1: Musical experience etc. of the participants

Do you play an instrument?	yes: 3, no: 8
Have you performed on a stage?	yes: 4, no: 7
Have you recorded singing voices?	yes: 0, a little: 3, almost no: 3, no: 5
Have you used DAWs?	yes: 0, a little: 1, almost no: 1, no: 9
Have you tried mixing?	yes: 0, a little: 0, almost no: 1, no: 10
Have you tried equalization?	yes: 1, a little: 0, almost no: 0, no: 10
How often do you go to karaoke?	don't go: 1, a few per year: 4, about once per month: 4, about once per week: 2, more often: 0
Do you adjust the volume at karaoke?	yes: 4, sometimes: 3, almost no: 1, no: 3
Do you have confidence in your answers?	yes: 1, a little: 3, neither: 0, not much: 5, no: 2
Do you know the used song?	yes: 2, I've heard: 1, no: 8

Step 4 At the same time as **Step 3**, evaluate, on a scale of -10 to 10, how stronger (or weaker) the characteristics expressed by each word became in the equalized stimulus than the pre-equalized one.

Step 5 Rest for 10 seconds (before moving to the next stimulus)

We asked participants to listen to the pre-equalization stimulus before every equalized stimulus in order to reduce the effects of the order of presenting the stimuli. We asked the participants to listen to every equalized stimulus three times in order to enable them to evaluate the sound stimulus while listening to it. The order of presenting the stimuli was random, but some participants listened in the same order because up-to-three participants participated in the experiment at the same time. All participants used headphones (MDR-900ST, Sony) to listen to the sound stimuli.

2.4. Participants

The participants were 11 university students (age: 18 to 22; 7 male and 4 female). Their music experiences are listed in Table 1.

2.5. Method of Analysis

For each sound stimulus i, the equalization setting was described as a 5-dimensional vector $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})^\top$ ($\top$: the transposition operator), where each x_{ik} is 0 (neither boosted nor cut), 12

Table 2: Parameters b_{jk} obtained with multivariate linear regression, coefficients of determination R_j, and the stimulus-wise average σ_j of evaluations of the same stimulus among all participants

	b_{j0}	b_{j1}	b_{j2}	b_{j3}	b_{j4}	R_j	σ_j
warmness	0.9692	0.0880	-0.0256	-0.0326	-0.0572	0.8394	2.1615
presence	0.9929	-0.0293	-0.1158	0.0534	0.0869	0.6989	1.9926
showiness	-0.0268	-0.0674	-0.1110	0.0885	0.0847	0.6659	1.8198
muddiness	0.5798	0.0087	0.0201	-0.0254	-0.0266	0.2224	2.0464
mellowness	0.7111	0.1072	0.0359	-0.0569	-0.1068	0.7223	1.8420
softness	0.6424	0.1030	0.0367	-0.0447	-0.1375	0.7623	1.9544
brightness	0.1722	-0.0375	-0.0590	0.0705	0.0660	0.6275	1.9851
lightness	0.0889	-0.1025	0.0155	0.0717	0.0572	0.7867	2.0711
thickness	0.3374	0.0889	-0.0557	-0.0273	-0.0216	0.7585	2.0325
clearness	-0.1394	-0.1045	-0.0395	0.0792	0.0937	0.7169	2.1486

(boosted by 12 dB), or -12 (cut by 12 dB). Then, the average of the evaluations of stimulus i given by all participants with respect to sound expression word j is denoted by y_{ij} ($j = 1, \cdots, 10$). Here, y_{ij} is approximated with multivariate linear regression, that is, $y_{ij} = \boldsymbol{b}_j^\top \boldsymbol{x}_i + u_{ij}$, where $\boldsymbol{b}_j = (b_{j0}, \cdots b_{j4})^\top$ is a parameter vector and u_{ij} is an approximation error to be minimized. If b_{j1} and b_{j2} are positive values for a certain word j, characteristics expressed by this word is expected to be enhanced when a low pitch range is boosted. If b_{j3} and b_{j4} are positive values, it will be enhanced when a high pitch range is boosted.

3. RESULTS AND DISCUSSIONS

Table 2 lists the parameter vector $\boldsymbol{b}_j$ obtained with multivariate linear regression, the coefficients of determination R_j, and the stimulus-wise average σ_j of the standard deviation of the evaluations of each stimulus among the participants with respect to each sound expression word j.

For *warmness*, b_{j1} was a positive value. This means that warmness is enhanced when the range of 125 to 250 Hz is boosted. This result matches the fact that a know-how book for hobby musicians [5] says that the range between 90 Hz and 400 Hz generates warmness.

Similarly, b_{j1} (125–250 Hz) for *thickness* as well as b_{j1} and b_{j2} (500–1,000 Hz) for *muddiness* and *mellowness* were positive values. Multiple books [5–7] also report that these words are related to a low pitch range, so our result matches this description.

Softness was also related to a low pitch range; b_{j1} and b_{j2} for this word were positive values, and b_{j1} was greater than b_{j2}. This result matches published descriptions [5–7] that softness can be enhanced by boosting 275 Hz.

On the other hand, *showiness*, *brightness*, *clearness*, and *presence* were related to a high pitch range. For these words, b_{j3} (2–4 kHz) and b_{j4} (8–16 kHz) were positive values. Also, b_{j4} was greater than b_{j3} for *presence* and *clearness*. Although Book [6] says that boosting a middle pitch range such as 4 to 6 kHz enhances the presence, our result implies that a higher pitch range should also be boosted to enhance the presence.

The parameter vectors for *mellowness* and *softness* were very close. This is because it was difficult to distinguish these two words. In fact, many people answered in an interview that they did not understand the difference between mellowness and softness.

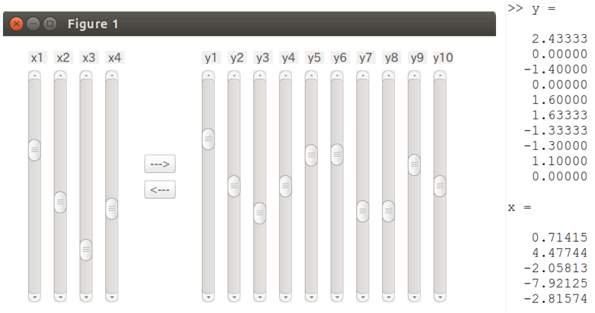
The coefficients of determination were higher than 0.6 for all sound expression words except *muddiness*. During the interview, some participants said that it was difficult to evaluate the difference of muddiness between two sounds even if they identify the difference of the timbres.

The standard deviations of the evaluations among the participants were between 1.8 and 2.2.

4. APPLICATION TO VERBAL GEQ CONTROLLER

Above, the sound evaluation y_j with respect to the sound expression word j is approximated by $y_j \approx \boldsymbol{b}_j^\top \boldsymbol{x}$. This can be described using a matrix, so that $\boldsymbol{y} \approx B\boldsymbol{x}$, where $\boldsymbol{y} = (y_1, \cdots, y_{10})^\top$ and $B = (b_{jk})$. That means that, given a vector $\boldsymbol{x}$ from the GEQ setting space, the corresponding vector $\boldsymbol{y}$ in the verbal evaluation space can be predicted. Inversely, given a vector $\boldsymbol{y}$ from the verbal evaluation space, the corresponding vector $\boldsymbol{x}$ in the GEQ setting space can be predicted by $\boldsymbol{x} \approx B^+\boldsymbol{y}$, where B^+ is the pseudo-inverse matrix of B.

Based on this idea, we implemented a prototpe of a verbal GEQ controller, which estimates a GEQ setting from a verbal evaluation vector, using Octave 4.0. An example of using this prototype is shown in Figure 1. A set of four vertical sliders in the left-hand area represents a GEQ setting vector, while a set of 10 vertical sliders in the right-hand area represents a verbal evaluation vector. Once the user inputs a verbal evaluation vector with the 10 sliders in the right-hand area and presses the "<---" button, the system calculates the corresponding GEQ setting vector and repre-



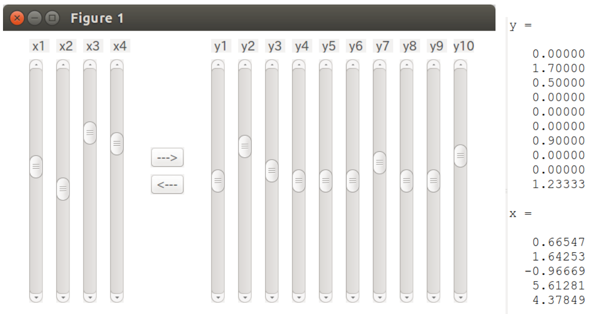
```
>> y =

  2.43333
  0.00000
 -1.40000
  0.00000
  1.60000
  1.63333
 -1.33333
 -1.30000
  1.10000
  0.00000

x =

  0.71415
  4.47744
 -2.05813
 -7.92125
 -2.81574
```

Figure 1: An example of our prototype of verbal GEQ controller (Left: GUI, Right: specified $\boldsymbol{y}$ and estimated $\boldsymbol{x}$)



```
y =

  0.00000
  1.70000
  0.50000
  0.00000
  0.00000
  0.00000
  0.90000
  0.00000
  0.00000
  1.23333

x =

  0.66547
  1.64253
 -0.96669
  5.61281
  4.37849
```

Figure 2: Another example of our prototype of verbal GEQ controller (Left: GUI, Right: specified $\boldsymbol{y}$ and estimated $\boldsymbol{x}$)

sents it in the four sliders in the left-hand area (its inverse calculation is also supported). In this example, the slider for *warmness* was set to the highest value, those for *mellowness*, *softness*, and *thickness* were set to non-highest but positive values, while those for *showiness*, *brightness*, and *lightness* were set to negative values. As a result, we obtained a 4.47-dB boost for 125–250 Hz, a 2.06-dB cut for 500–1,000 Hz, a 7.92-dB cut for 2–4 kHz, and a 2.82-dB cut for 4–8 kHz. This low-boost setting was obtained because, according to our investigation, *warmness*, *mellowness*, *softness*, and *thickness* are enhanced when a low pitch range is enhanced.

Another example is shown in Figure 2. In this example, the sliders for *presence* and *clearness* were set to positive values, and those for *showiness* and *brightness* were set to small but positive values. As a result, we obtained a 1.64-dB boost for 125–250 Hz, a 0.97-dB cut for 500–1,000 Hz, a 5.61-dB boost for 2–4 kHz, and a 4.38-dB boost for 8–16 kHz. This is because these sound expression words are related to a high pitch range.

5. CONCLUSION

Our goal is to enable non-professional musicians to equalize singing voices by describing the desired timbre using verbal expressions. To achieve this, we investigated the mapping between the GEQ settings and verbal evaluations of various equalized singing voices using multivariate linear regression. As a result, we obtained a matrix for transforming a vector from the GEQ setting space to a vector in the verbal evaluation space and vice versa, and this transformation matrix matches findings described in know-how books for hobby musicians. We also implemented a prototype of a system that estimates a GEQ setting from a verbal evaluation vector.

We have some remaining issues. First, we should investigate other singing voices (particularly female voices). Second, a 10-dimensional verbal evaluation vector is still too high dimensional for non-professional musicians to specify, so we should consider dimensionality reduction of this vector by applying a technique such as principal component analysis (PCA). Also, we should evaluate how close GEQ settings estimated from verbal evaluation vectors are to the timbre desired by users.

6. ACKNOWLEDGMENTS

This work was supported by JSPS Kakenhi Grant No. JP17H00749.

7. REFERENCES

[1] Gerrit Bloothooft and Reinier Plomp: The Timbre of Sung Vowels, Journal of Acoustic Society of America, vol. 84, no. 3, pp. 847–860, 1988.

[2] Katsuya Yamauchi, Yasunao Kai, and Shin-ichiro Iwamiya: The effects of materials of a flute ’s crown and a cello’s endpin on the timbre of musical instruments, Acoustic Science and Technology, vol. 22, no. 1, pp. 47– 48, 2001.

[3] Jan Stepanek: Musical Sound Timbre: Verbal Description and Dimensions, Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06), pp.121–126, 2006.

[4] Enrique Perez Gonzalez and Joshua Reiss: Automatic equalization of multi-channel audio using cross-adaptive methods, Proceedings of the 127th AES Convention, 2009.

[5] Tomoyuki Sumi: *Sugu ni tsukaeru EQ receipe — DAW user hikkei no gakki-betsu setting shu* (EQ receipes useful soon — DAW users ’bible of EQ setting collection by instruments), Ritto Music, 2011. (in Japanese)

[6] Tomoyuki Sumi: *Engineer ga oshieru vocal effect technique 99* (99 vocal effect techniques that an engineer teaches, Ritto Music, 2012. (in Japanese)

[7] Yoshiro Kuzumaki: *Engineer ga oshieru mix technique 99* (99 mix techniques that an engineer teaches, Ritto Music, 2009. (in Japanese)