A Music Exploration Interface Based on Vocal Timbre and Pitch in Popular Music

Tomoyasu Nakano[1], Momoka Sasaki[2], Mayuko Kishi[2], Masahiro Hamasaki[1], Masataka Goto[1], and Yoshinori Hijikata[2] *

[1] National Institute of Advanced Industrial Science and Technology (AIST)
[t.nakano, masahiro.hamasaki, m.goto]@aist.go.jp
[2] School of Business Administration, Kwansei Gakuin University
contact@soc-research.org

Abstract. This paper proposes an interface that enables music exploration focusing on two factors of singing voices, vocal timbre and pitch, that are useful in finding singing voices that match users' preferences. The proposed interface uses a two-dimensional color map to visualize songs being explored and locates them according to timbre or pitch similarities of their singing voices. Since similar songs are located closely on the map, users can visually find singing voices similar to their favorite singing voices. In addition to the location, the interface uses the color of each song on the map to visualize an additional factor related to characteristics of singing voices, such as acoustic features or words describing singing voices (*e.g.,* "Clear"). Prior to developing the interface, we conducted a questionnaire survey with 20 participants and confirmed that both vocal timbre and pitch are important when listening to music. The proposed interface was implemented with 102 songs, and a user study was conducted with 60 participants.

Keywords: Music information retrieval, vocal timbre, pitch histogram, singing descriptors, music exploration interface

1 Introduction

Since vocals are one of the major parts in music [1], music information retrieval (MIR) technologies focusing on singing voices are beneficial to a wide range of users [2]. In fact, MIR methods and interfaces that focus on various factors of singing voices — such as vocal timbre [3–6], vocal range profile (*i.e.,* pitch and intensity) [7], lyrics [5,8–11], singing style [12–14], and gender [15] — have been proposed for the purpose of listening to songs or the purpose of finding songs to sing. In order to provide a new direction for such a series of academic studies, this paper proposes a novel interface that enables exploratory music retrieval focusing on multiple factors of singing voices.

* This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JSPS KAKENHI Grant Number JP21H04917.

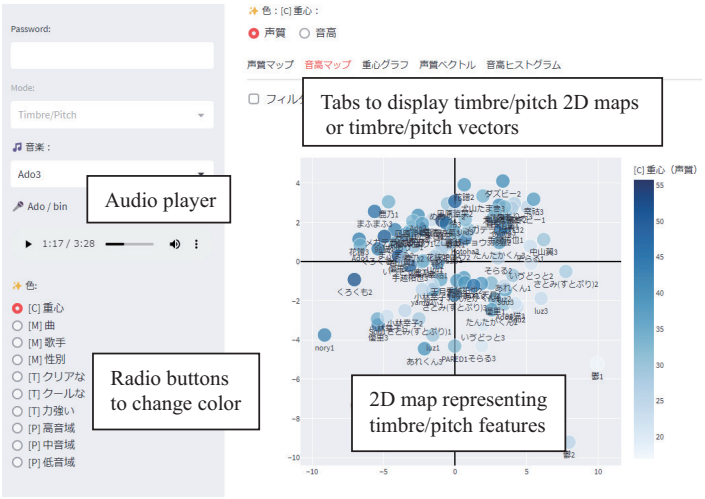

Fig. 1. Screenshot of the proposed interface.

Since singing voices have various factors, a music exploration interface that allows switching the visualized factors to be focused on is convenient for uses with different purposes. For users who are interested in finding singers having a similar vocal timbre, visualizing the vocal timbre is useful, and for users who are interested in finding songs having similar vocal pitch distribution, visualizing the factor of vocal pitch is useful. The factors to be focused on thus depend on the purpose of the exploration.

We target vocal timbre and pitch for our interface. We consider these two factors to be effective in music exploration for two reasons. First, as a result of our survey explained later in which participants were asked to describe their favorite singing voices, many of the answers described vocal timbre and pitch. Second, it is helpful for users who want to find songs with their favorite singing voices to use or combine vocal timbre and pitch. Recently, there has been a culture in which a lot of people enjoy singing existing songs as cover versions and share their cover songs online. Users who enjoy such songs could find and enjoy songs having their favorite singing voices even if they do not know those songs or singers.

We therefore developed a music exploration interface that visualizes the two factors, vocal timbre and pitch, and enables users to switch the visualized factors to find songs having their favorite singing voices. A screenshot of our interface is shown in Figure 1. On the right side, each song is depicted as a circular dot on a two-dimensional color map representing the similarity of vocal timbre or vocal pitch factors, which can be interactively switched by a user. Since similar songs are located closely on the map, the user can easily find a song having a vocal timbre similar to that of the user's favorite singer on the map focusing on the vocal timbre similarity. The user can see the song title and singer name by mousing over a song. Each song has an identifier (ID) based on the singer name (e.g., *Ado3* means the third song of singer *Ado* in our dataset used for the interface), and the user can play back a song by specifying its ID from a pull-down menu on the left sidebar of the screen. The interface uses the color of the song to indicate one of the following: singer name, song title, singer gender, center of gravity of

the average mel spectrum, center of gravity of the pitch histogram, and singing descriptors (*e.g.,* "Clear"). This additional color helps users understand the characteristics of singing voices in finding their favorite songs, and the combination of the location and color on the two-dimensional map gives high flexibility in visualizing multiple factors of singing voices. To the best of our knowledge, a flexible music exploration interface that leverages both vocal timbre and pitch factors has not been proposed.

2 Related work

Related to this research are studies on music visualization interfaces for finding one's favorite singers or lyrics. Fujihara *et al.* [3] proposed VocalFinder, an interface that retrieves songs having similar singing voices by modeling vocal timbre and singing style using a Gaussian mixture model. Hamasaki *et al.* [15] proposed Songrium, a music browsing assistance interface that has a function to analyze and visualize singing voices. It uses a circle to visualize a song, and the color and size of the circle indicate the singer's gender and the number of play counts, respectively. Sasaki *et al.* [8] proposed LyricsRadar, an interface that estimates topic distributions from lyrics text using latent Dirichlet allocation and locates lyrics on a two-dimensional map using t-SNE [16]. Tsukuda *et al.* [10] proposed Lyric Jumper, an interface that visualizes lyric topic distributions for each singer as a donut chart to let users find singers with similar topics. Watanabe *et al.* [11] proposed Query-by-Blending, an interface that enables users to find songs by a query combining lyrics, song acoustic signals, and artists.

Map-based music browsing interfaces that locate songs on a two- or three-dimensional map have also been proposed [17]. In addition, as MIR methods targeting pitch, Tzanetakis *et al.* [18] used a pitch histogram to automatically classify music genres. Moreover, to recommend songs appropriate for the user's singing ability, a feature called vocal range profile (VRP) has also been studied (*e.g.,* [7]). The VRP indicates the range of intensity for each pitch that a singer can sing.

Words that describe singing voices help determine vocal characteristics that people are likely to pay attention to when listening to songs. There have been studies on emotional expressions of singing voices [19, 20]. Scherer *et al.* [20] studied the correlation of acoustic features to "anger", "fear", "tenderness", "joy", "sadness", and "pride" when eight professional opera singers sang musical scales. There have also been studies that determined a set of words that express impressions of singing voices and annotated them to songs [21, 22] for their automatic estimation. Kanato *et al.* [21] defined a set of 47 impression words of singing voices. The factor analysis revealed three factors, "power," "politeness," and "brightness," as well as 12 words (*e.g.,* "clear" and "cute") that comprise the singing impression scale. Kim *et al.* [22] defined 70 vocabulary words to describe solo singers. From results of five semi-experts' annotations of actual songs using those 70 vocabulary words, 42 vocabulary words were obtained and classified into five categories: pitch (range), timbre, gender, genre, and technique.

Compared with the above studies, the key contribution of this paper is to develop a music exploration interface that uses a combination of vocal timbre and pitch. Another contribution is that we show the appropriateness of using vocal timbre and pitch as factors in music exploration by conducting a questionnaire survey.

3 Survey of preference for singing voice when listening to music

Prior to the interface development, a questionnaire survey was conducted with 20 participants, males and females in their twenties. The purpose of the survey was to investigate users' impressions of vocals and their needs when listening to music. Although there were previous studies [21, 22] that defined words to describe singing voices in popular music, they did not focus on preference for singing voice when listening to music.

The questionnaire consisted of the following three sections:

S1: a section to measure the participants' musical ability based on the Goldsmiths Musical Sophistication Index (Gold-MSI) [23],
S2: a section in which participants were asked to write freely about their favorite singing voices, favorite artists/songs, and the reasons for their favorites, and
S3: a section in which participants were asked to rate 3 vocal aspects (pitch height, pitch range, and timbre on a 7-point scale.

3.1 S1: Musical sophistication of participants

In the Gold-MSI, participants answer questions such as "I am able to hit the right notes when I sing along with a recording." on a 7-point scale from 1 (Completely Disagree) to 7 (Completely Agree). In this paper, we asked participants to answer questions on "Active Engagement" and "General Sophistication" because we thought they are relevant to music appreciation in general. Since the survey focused on singing, participants were also asked to answer questions about "Singing Abilities."

The scores for "Active Engagement," "Singing Abilities," and "General Sophistication" are shown in Figure 2. Each score was obtained by averaging the raw score values for the relevant questions for comparisons independent of the number of questions. In the scatter plot, differences in the Singing Abilities scores are represented by different colors. The results show that all median values of the scores were around 4. The correlation between Active Engagement and General Sophistication was high at 0.90, and their medians were slightly below 4, indicating a slightly lower score distribution. On the other hand, the median for Singing Abilities was slightly above 4, with a balanced distribution of high and low scores. The correlation between Singing Abilities and Active Engagement was 0.59, and that between Singing Abilities and General Sophistication was 0.74. These results indicate that the participants had an average interest and ability in music, and the high correlation between Active Engagement and General Sophistication indicated a certain degree of reliability in their answers.

3.2 S2: Preference for singing voice when listening to music

Here, analysis focused on answers to the following two open-ended questions.

Q1 "Please describe as many characteristics as possible of your favorite singing voice when you listen to music."
Q2 "Please describe the artist whose voice you like to listen to. Please also describe what you like about that artist's voice."

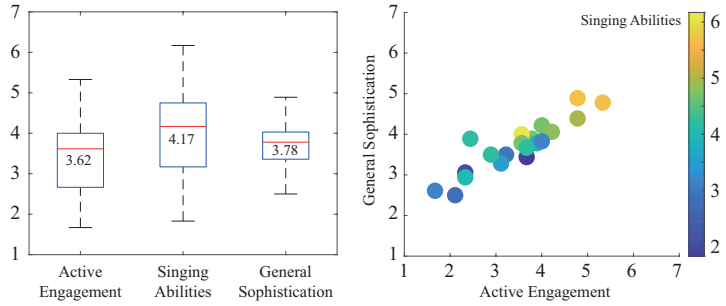

Fig. 2. Distribution of scores for Active Engagement, Singing Abilities, and General Sophistication in Gold-MSI. In the scatter plot, differences in Singing Abilities scores are indicated by different colors. The higher the scores, the more sophisticated with regard to those factors.

For the answers to Q1 and Q2, the words used by participants to describe the singing voice are shown below. These words are hereafter referred to as "singing descriptors." The number of people who used them is also shown in parentheses. According to Kim *et al.* [22], each singing descriptor was classified into four categories: pitch, timbre (voice quality, singing style, emotion[3]), gender, and singing ability (singing technique).

- **Pitch**: High-pitched (10), Low-pitched (6), Not too high (1), Mid-low range (1), Not too low (1), Very low (1), Wide range (1)
- **Timbre (voice quality, singing style, emotion)**[4]: Clear / Transparent (14), Beautiful (6), Unique (6), Tender (4), Powerful (4), Calm (4), Fluffy / Airy / Floating (3), Cute (3), Comfortable (3), Cool (2), Sexy (2), Sweet (2), Cheerful / Energizing (2), Likable (2), Rough (2), Soft (2), Deep (2), Delicate (2), Distinctive (2), Healing (2)
- **Gender**: Female (3), Male (2), Neutral (1)
- **Singing ability (singing technique)**: Expressive (3), Falsetto (2), Large inflection (2), Vibrato (2), Strong (2), Accurate pitch control (2), Long tones without hoarseness (1), Comfortable high tone (1), Breathy (1), Head voice (1), Sound on inhalation (1), Precise control (1), Not labored (1), Emotional variation (1), Steady (1), Kobushi[5] (1), Growl (1), Sing out from the stomach (1)

The above results show that singing descriptors related to pitch and timbre were frequently used. Pitch-related "High-pitched" and "Low-pitched" were included in 10 and 6 answers, respectively. Timbre-related "Clear / Transparent" and "Beautiful" were included in 14 and 6 answers, respectively. On the other hand, singing descriptors related to gender and singing ability (singing technique) were not frequently used. These results suggest that pitch and timbre are important in describing favorite singing voices.

[3] Emotion evoked by singing voices was classified here since it could be related to timbre.
[4] Since there were too many singing descriptors answered for the timbre category, only singing descriptors answered by two or more participants are shown.
[5] A singing technique that uses two or more notes in a single syllable, as in melisma.

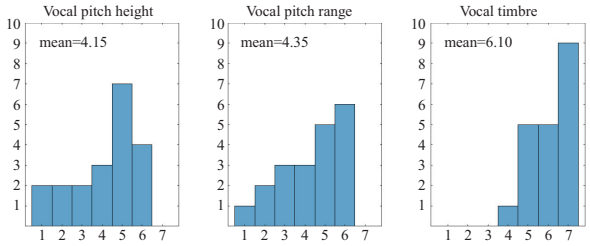

Fig. 3. S3: Answers to questions related to vocal timbre and pitch on a 7-point Likert scale. The higher the score value, the more aware of the factor when listening to music.

3.3 S3: Factors of singing to be aware of when listening to music

Using a 7-point Likert scale, we asked participants to rate three vocal factors — pitch height, pitch range, and timbre — as vocal elements they are aware of when listening to music, without limiting themselves to specific songs.

The numbers of participants who answered each of the rating points (scores) are shown in Figure 3. The average scores for vocal pitch height and vocal pitch range were 4.15 and 4.35, respectively, indicating that the degree of awareness was higher than 4. The average score for vocal timbre was 6.1, which was also high.

3.4 Discussion

The results of this survey suggest that vocal pitch and timbre play important roles in determining a favorite singer's voice when participants with an average musical sophistication listen to music. This is also supported by previous studies in which pitch and timbre categories were used as vocal tags defined by Kim *et al.* [22] and music tags defined by Turnbull *et al.* [24]. We therefore believe that developing a music exploration interface focusing on vocal pitch and timbre is worthwhile and effective.

4 Interface

In order to visualize the similarity of vocal timbre and pitch and to enable exploratory search, we implement the interface (Fig. 1) as a map-based interface [17], which has been proposed widely in the past. The proposed interface estimates the timbre and pitch feature vectors of vocals from audio signals of each song and uses them to locate each song as a single circular point on a two-dimensional color map.

4.1 Data and back-end processing

The songs used for interface development were 51 songs for 17 female singers (3 songs for each singer), and 51 songs for 17 male singers (3 songs for each singer), for a total of 102 songs. There are multiple renditions of songs by different singers available in the 102 songs (2 or 3 renditions per song). All of these were Japanese popular music and had at least 10,000 views on YouTube as of December 2022.

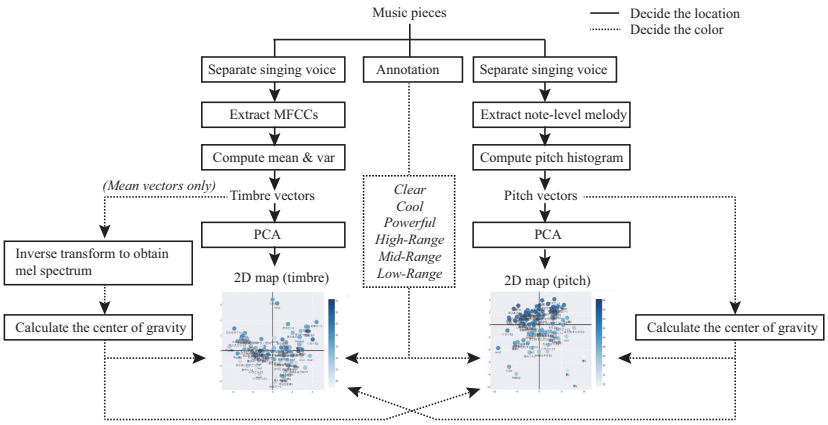

Fig. 4. Overview of the back-end processing of the proposed interface.

An overview of the back-end processing is shown in Figure 4. First, the singing voices were separated from all 102 songs using Hybrid Demucs [25]. To estimate pitch histograms, note-level pitch sequence was estimated by using Omnizart [26]. The pitch histograms were standardized by song to eliminate the effect of song length and then standardized by dimension and referred to as pitch vectors.

Timbre features were obtained by calculating the mean and variance of each dimension of MFCCs from the separated singing voice.To calculate MFCCs, STFT was calculated for a music signal with a sampling frequency of 22,050 Hz, with a window length of 2048 and a shift width of 512. The number of mel frequency bins was 128 and the MFCC dimension was 12, excluding DC components. Here, the vocal activity segments were determined by utilizing the note-level pitch information from Omnizart, and only the MFCCs for those vocal segments were used to calculate the mean and variance. Finally, the mean and variance of the MFCCs for all 102 songs were standardized by dimension and referred to as timbre vectors.

Finally, principal component analysis was performed on these timbre and pitch vectors, and we located them in a two-dimensional timbre map and a two-dimensional pitch map. The performance of Hybrid Demucs was high enough, but even if there were errors, they were unlikely to affect the histograms and mean vectors.

4.2 Annotate singing descriptors

For the purpose of improving the user's understanding of the map, singing descriptors from human annotation are also used for coloring. To determine appropriate singing descriptors for each song, the 102 songs were tagged by six annotators, three male and three female. Three annotators per song were assigned to tag the singing voice, and at least one of the three was of a gender different from that of the singer of the song.

The singing descriptors used in this paper were determined based on previous studies [21, 22, 27] in which inter-annotator agreement, intelligibility, or synonymity were taken into account. First, 33 descriptors were selected from the tags used in the KVT dataset [22], 3 descriptors related to pitch range and 30 descriptors related to timbre.

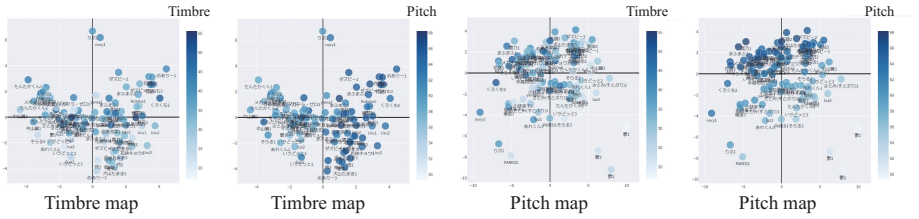

Fig. 5. Timbre maps and pitch maps colored using either the center of gravity of the timbre vector or that of the pitch vector.

Fig. 6. Timbre maps colored based on gender and three singing descriptors, “Clear,” “Cool,” and “Powerful.” Continuous coloring for each of the three descriptors depends on the number of annotators who labeled it.

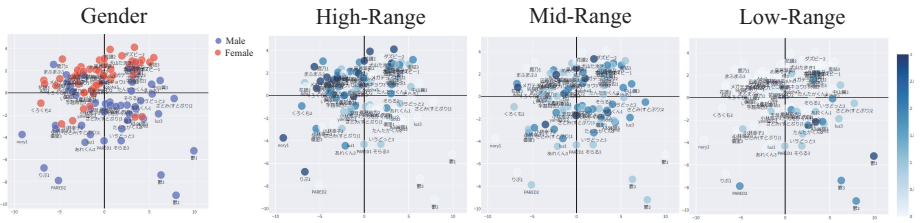

Fig. 7. Pitch maps colored based on gender or three singing descriptors, “High-Range,” “Mid-Range,” and “Low-Range.” Continuous coloring for each of the three descriptors depends on the number of annotators who labeled it.

Then nine descriptors were added, including seven descriptors — Powerful, Nasal, Calm, Weak, Sexy, Resonant, and Dosu (Threatening / Frightening) — that were selected from previous studies on singing impression [21] and speech timbre [27], and two descriptors — Beautiful and Cool — that were from previous studies of singing impression [21] and were related to 40 other descriptors. As a result, a total of 42 different descriptors were determined as singing descriptors labeled by the annotators.

Then, since using all the 42 descriptors gives too much information and is difficult, we used only the top three timbre descriptors — “Clear,” “Cool,” and “Powerful” — on the basis of how often they were annotated. As for the pitch descriptors, we used all the three descriptors for pitch ranges: “High-Range,” “Mid-Range,” and “Low-Range.”

4.3 Interaction

The user can select either the timbre map or the pitch map, and can change the color of the songs by using one of the following: singer name, song title, singer gender, center

of gravity of timbre vector, center of gravity of pitch vector, and singing descriptor (the number of annotators who assigned it). Discrete coloring is applied to the singer name, song title, and vocal gender, and continuous coloring (*i.e.,* gradation) is applied to the rest. The singer name, song title, and vocal gender are taken from metadata of the songs.

Figure 5 shows the timbre and pitch maps, each of which is colored using either the center of gravity of the timbre vector or that of the pitch vector. Figures 6 and 7 also show the timbre and pitch maps, respectively, colored using vocal gender and the corresponding singing descriptors. We can see that the horizontal axis of the timbre map is correlated with gender in the first map of Figure 6, the correlation coefficient was 0.80, as well as the center of gravity of the pitch vector in the second map of Figure 5 (0.62). It is also correlated with the number of annotators of "Clear" (0.57) in the second map of Figure 6, though the vertical axis of the timbre map is weakly correlated with the number of annotators of "Powerful" (0.34) in the fourth map of that figure.

The vertical axis of the pitch map is also correlated with gender in the first map of Figure 7 (0.51) as well as the center of gravity of the pitch vector in the fourth map of Figure 5 (0.84). It is also weakly correlated with the number of annotators of "High-Range" (0.38) and "Low-Range" (-0.48) in the second and fourth maps of Figure 7, though the horizontal axis of the pitch map is weakly correlated with the center of gravity of the timbre vector in the third map of Figure 5 (-0.37).

As shown in these examples, the proposed interface enables flexible changes in location and coloring with respect to timbre and pitch as well as related singing descriptors.

5 Evaluation

Since the proposed interface has functions for people who like music, we evaluated the effectiveness in terms of entertainment and knowledge discovery rather than efficiency and accuracy. Sixty participants, males and females in their teens or twenties, were assigned to the following groups, G1 through G3, each with 20 participants.

- **G1 (proposed)**: Using music exploration interface based on vocal timbre and pitch
- **G2**: Using music exploration interface based on pitch
- **G3**: Using music exploration interface based on timbre

G2 and G3 are comparison groups to evaluate the effectiveness of the proposed interface. Participants assigned to G2 could not use the timbre map, the center of gravity of the timbre vector, or the singing descriptors for timbre. Participants assigned to G3 could not use the pitch map, the center of gravity of the pitch vector, or the singing descriptors for pitch. Prior to the start of the experiment, the experimenter verbally explained the experiment procedure to the participants in Japanese. The experiment was conducted on a laptop computer, and participants played music using canal-type wired earphones. Participants were paid 1,800 JPY for their participation in the experiment (approximately 1 hour and 45 minutes).

Participants first completed a questionnaire that measured their level of interest in music and then they watched a video explaining the interface. The explanation was made as easy to understand as possible for participants who are not familiar with MIR, using as little technical terminology as possible. Next, while recording the screen operation, participants were asked to explore their favorite music until they got bored

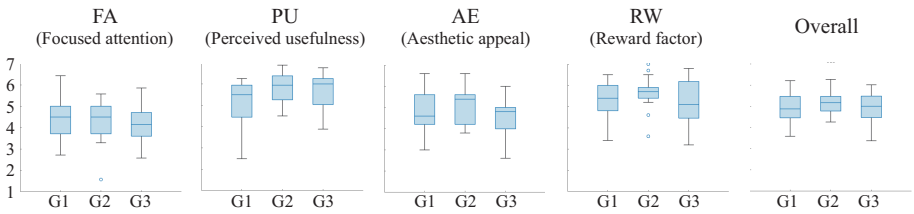

Fig. 8. Box plots of the scores in UES-LF.

within the duration of the experiment. After the experiment, each subject answered a questionnaire and was interviewed. In the post-experimental questionnaire, we assessed focused attention (FA), perceived usefulness (PU), aesthetic appeal (AE), and reward factor (RW) using questions from the User Engagement Scale (UES-LF) [28] on a 7-point scale. Participants also answered open-ended questions about the pros and cons of the interface. In addition, participants in G1 answered whether they felt that vocal timbre or pitch was more suitable for them when exploring the music.

5.1 Results

First of all, the data of three participants (two in G2 and one in G3) were filtered out because the data were inappropriate (*e.g.,* did not use the map). Using the data from G1 to G3 after filtering, each score in the UES-LF was calculated. The average screen recording time for the 57 participants was 28.8 minutes (ranged from 11.8 to 53.1 min).

As shown in Figure 8, where "Overall" is the overall engagement score, obtained by averaging the other four scores. A one-way ANOVA confirmed a significant difference only in PU at the 5% level ($p = 0.037$). The results of Bonferroni's multiple comparison test based on Wilcoxon's rank-sum test showed no significant differences in all combinations. This suggested that the type of interface did not affect user engagement.

Regarding the answers to the experimental questionnaire, 13 of the 20 participants in G1 answered that the vocal timbre feature was more suitable when searching for music, while 7 participants answered that the pitch feature was more suitable. This confirmed the need for our interface that allows searching from multiple factors since the vocal timbre feature works best for some users and the pitch feature works best for others. Moreover, in the interview, seven participants in G1 commented that the combination of vocal timbre and pitch facilitated their exploration. Some participants in G1 to G3 understood their own preferences for vocal timbre and pitch, while others found that they unexpectedly liked timbres and pitches that they had thought they did not like.

The top three functions mentioned as pros by all 57 of the participants were the timbre and pitch maps by 29 participants and the timbre and pitch vectors by 16 participants. In addition, 11 participants mentioned the design and usability of the interface, and 11 participants mentioned the identification or change of their preferences. On the other hand, since the design and usability were also mentioned as cons by 50 participants, it is necessary to improve the usability of the implementation in the future. Eight participants also commented that they did not understand the meaning of the axes of the two-dimensional color map and that the differences in color according to acoustic

features and singing descriptors did not match their own perception. Therefore, there is a possibility of developing a better interface to help users grasp the meaning of acoustic features and singing descriptors.

5.2 Discussion

The following can be considered as reasons for the small differences in UES-LF between G1 and G2 or between G1 and G3.

- Exploring music from the visualization of pitch and timbre was a novel experience for the participants. Even for G2 and G3, the participants may have felt that it was enough for them to find preferred songs from a new point of view. In fact, some participants understood their own preferences for pitch and timbre and discovered new or unexpected preferences during the use of the interface.
- This may be due to the doubling of the amount of information and manipulation. The interface has become more complex, which probably increased the time and effort required for participants to become familiar with the interface operation.

Three participants in G2 commented that while they felt the pitch information was effective, they also wanted information on vocal timbre. Therefore, some users are expected to be more satisfied with our interface that allows for both pitch and timbre.

6 Conclusion

In this paper we proposed a music exploration interface that flexibly visualizes vocal timbre and pitch as well as singing descriptors. The questionnaire survey results indicated that the vocal timbre and pitch can be utilized to explore music. In the present analysis based on the UES-LF, no significant differences were identified between the proposed interface and the comparison interfaces. However, the results of the questionnaire and interviews indicated that music exploration based on vocal timbre and pitch not only provides enjoyment and fun but also leads to the discovery of preferences regarding timbre and pitch. Future work will include building an interface that improves usability and taking into account the singer's singing style.

References

1. Demetriou, A., *et al.*: Vocals in Music Matter: the Relevance of Vocals in the Minds of Listeners. Proc. ISMIR 2018, pp. 514–520 (2018).
2. Humphrey, E.J., *et al.*: An Introduction to Signal Processing for Singing-Voice Analysis: High Notes in the Effort to Automate the Understanding of Vocals in Music. IEEE Signal Processing Magazine, vol.36, pp. 82–94 (2019).
3. Fujihara, H., *et al.*: A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval. IEEE TASLP, vol.18, no.3, pp. 638–648 (2010).
4. Nakano, T., *et al.*: Vocal Timbre Analysis Using Latent Dirichlet Allocation and Cross-Gender Vocal Timbre Similarity. Proc. ICASSP 2014, pp. 5239–5343 (2014).

5. Nakano, T., *et al.*: Musical Similarity and Commonness Estimation Based on Probabilistic Generative Models of Musical Elements. IJSC, vol.10, no.1, pp. 27–52 (2016).
6. Nakano, T., *et al.*: Musical Typicality: How Many Similar Songs Exist?. Proc. ISMIR 2016, pp. 695–701 (2016).
7. Mao, K., *et al.*: Competence-Based Song Recommendation: Matching Songs to One's Singing Skill. IEEE Trans. on Multimedia, vol.17, no.3, pp. 396–408 (2015).
8. Sasaki, S., *et al.*: LyricsRadar: A Lyrics Retrieval System based on Latent Topics of Lyrics. Proc. ISMIR 2014, pp. 585–590 (2014).
9. Nakano, T., *et al.*: LyricListPlayer: A Consecutive-Query-by-Playback Interface for Retrieving Similar Word Sequences from Different Song Lyrics. Proc. SMC 2016, pp. 344–349 (2016).
10. Tsukuda, K., *et al.*: Lyric Jumper: A Lyrics-Based Music Exploratory Web Service by Modeling Lyrics Generative Process. Proc. ISMIR 2017, pp. 544–551 (2017).
11. Watanabe, K., *et al.*: Query-by-Blending: A Music Exploration System Blending Latent Vector Representations of Lyric Word, Song Audio, and Artist. Proc. ISMIR 2019, pp. 144–151 (2019).
12. Ohishi, Y., *et al.*: A Stochastic Representation of the Dynamics of Sung Melody. Proc. ISMIR 2007, pp. 371–372 (2007).
13. Yamamoto, Y., *et al.*: Analysis and Detection of Singing Techniques in Repertoires of J-POP Solo Singers. Proc. ISMIR 2022, pp. 384–391 (2022).
14. Yakura, H., *et al.*: Self-Supervised Contrastive Learning for Singing Voices. IEEE/ACM TASLP, vol.30, pp. 1614–1623 (2022).
15. Hamasaki, M., *et al.*: Songrium: A Music Browsing Assistance Service with Interactive Visualization and Exploration of a Web of Music. Proc. WWW 2014 (2014).
16. Van der Maaten, L., *et al.*: Visualizing Data using t-SNE. JMLR, vol.9, no.11 (2008).
17. Knees, P., *et al.*: Intelligent User Interfaces for Music Discovery. TISMIR, vol.3, no.1, pp. 165–179 (2020).
18. Tzanetakis, G., *et al.*: Pitch Histograms in Audio and Symbolic Music Information Retrieval. JNMR, vol.14, no.2, pp. 143–152 (2003).
19. Scherer, K.R.Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication, vol.40, pp. 227–256 (2003).
20. Scherer, K.R., *et al.*: The Expression of Emotion in the Singing Voice: Acoustic Patterns in Vocal Performance. J. Acoust. Soc. Am., vol.142, no.4, pp. 1805—1815 (2017).
21. Kanato, A., *et al.*: An Automatic Singing Impression Estimation Method Using Factor Analysis and Multiple Regression. Proc. Joint ICMC SMC 2014, pp. 1244–1251 (2014).
22. Kim, K.L., *et al.*: Semantic Tagging of Singing Voices in Popular Music Recordings. IEEE/ACM TASLP, vol.28, pp. 1656—1668 (2020).
23. Müllensiefen, D., *et al.*: The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. PLOS ONE, vol.9, no.2 (2014).
24. Turnbull, D., *et al.*: Semantic Annotation and Retrieval of Music and Sound Effects. IEEE/ACM TASLP, vol.16, no.2, pp. 467–476 (2008).
25. Défossez, A.Hybrid Spectrogram and Waveform Source Separation. Proc. MDX 2021, pp. 1–11 (2021).
26. Wu, Y.T., *et al.*: Omnizart: A General Toolbox for Automatic Music Transcription. JOSS, vol.6, no.68, p. 3391 (2021).
27. Kido, H. and Kasuya, H.: Representation of Voice Quality Features Associated with Talker Individuality. Proc. ICSLP 1998, pp. 1–4 (1998).
28. O'Brien, H.L., *et al.*: A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and new UES Short Form. Intl. J. of Human-Computer Studies, vol.112, pp. 28–39 (2018).