Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18932-3

SingDistVis: interactive Overview+Detail visualization for F0 trajectories of numerous singers singing the same song

Takayuki Itoh[1] · Tomoyasu Nakano[2] · Satoru Fukayama[3] · Masahiro Hamasaki[2] · Masataka Goto[2]

Received: 8 December 2022 / Revised: 22 February 2024 / Accepted: 13 March 2024

Abstract
This paper describes SingDistVis, an information visualization technique for fundamental frequency (F0) trajectories of large-scale singing data where numerous singers sing the same song. SingDistVis allows to explore F0 trajectories interactively by combining two views: OverallView and DetailedView. OverallView visualizes a distribution of the F0 trajectories of the song in a time-frequency heatmap. When a user specifies an interesting part, DetailedView zooms in on the specified part and visualizes singing assessment (rating) results. Here, it displays high-rated singings in red and low-rated singings in blue. When the user clicks on a particular singing, the audio source is played and its F0 trajectory through the song is displayed in OverallView. We selected heatmap-based visualization for OverallView to provide an overview of a large-scale F0 dataset, and polyline-based visualization for DetailedView to provide a more precise representation of a small number of particular F0 trajectories. This paper introduces a subjective experiment using 1,000 singing voices to determine suitable visualization parameters. Then, this paper presents user evaluations where we asked participants to compare visualization results of four types of Overview+Detail designs and concluded that the presented design archived better evaluations than other designs in all the seven questions. Finally, this paper describes a user experiment in which eight participants compare SingDistVis with a baseline implementation in exploring interested singing voices and concludes that the proposed SingDistVis archived better evaluations in nine of the questions.

Keywords F0 trajectories · Overview+Detail visualization · Visualization for singings · Music visualization · Singing voice · Singing assessment · Singing information processing · Music information retrieval

1 Introduction

Many people sing the same songs and upload them on video sharing and social karaoke services, such as Niconico, YouTube, and Smule. Collections of the same songs performed

✉ Takayuki Itoh
itot@is.ocha.ac.jp

Extended author information available on the last page of the article

by different singers [11] increase the viewing pleasure of original songs and lead to the discovery of new favorite content and artists. Furthermore, because voice timbre and singing style are important factors that characterize the individuality and similarity of singings, the analysis of large-scale data of various singings of the same song is important for evaluating and comprehending the characteristics of human singing voice perception and singing styles. It is especially interesting to discover particular parts of the melody of a song that are difficult only for novice singers or even for expert singers. Or, it is also interesting to observe how user-interested expert singers have characteristics of their singing.

Since singing voice plays a central role in popular music [14], many singing-related technologies and singing interfaces technologies that are indispensable for singing analysis/synthesis have been actively studied in recent years [8, 9]. Focusing on visualization technique,

1. visualization of time-series acoustic data for one or a few singing voices [13, 17, 23, 25, 31, 32, 43], and
2. visualization of relationships between multiple songs [11, 19, 36]

have been studied.

In addition to the above two visualization research directions, along with the popularization of social media and user-generated content, a new research direction for singing visualization can be considered, i.e., the simultaneous visualization of numerous time-series acoustic data. Since many amateur singers have submitted their singing voices covering famous songs to video-sharing or social karaoke platforms, large-scale recordings covering the same song have accumulated and are useful for singing research. By leveraging such covers, visualization of the tendency of the acoustic features of numerous singing voices and detailed comparison of the voices can be useful for singing analysis and training.

However, to the best of our knowledge, the simultaneous visualization of numerous time-series acoustic data has not been studied except in the literature by Tsuzuki et al. [40]. Although they proposed utilizing a heatmap to show the fundamental frequency (F0, a physical quantity that corresponds to pitch) distribution for 4,941 singers, the visualization technique is only for a short-time local segment, and its efficiency has not been evaluated.

To summarize, there have been no studies for visualization of a large number of singings that satisfy the followings:

- Display the distribution and transition of F0 of all the singers through the song and assist the discovery of user-interested phrases of the melody.
- Display the detailed transition of F0 in a short phrase and assist the comparative observation between novice and expert singers.

In terms of more general purposes, visualizing large-scale data has been utilized for information retrieval [2, 12] and data analysis [4, 28]. The available techniques for multiple time-series are the polyline charts and scatterplots [41], brightness and hue [15, 33], line thickness [21], and density-based polylines [30, 45]. Such general-purpose visualization techniques can be applied to the visualization of the transition of F0; however, we need to carefully select the visualization techniques. Although the F0 trajectory of one singing voice (or several singing voices) is typically drawn by a line chart, a line chart including numerous F0 trajectories (numerous polylines) causes the visual cluttering problem and the data cannot be read, as shown in Fig. 1.

Based on the above background, this paper proposes SingDistVis, a technique for visualizing numerous F0 trajectories singing the same song. To solve the problem demonstrated in Fig. 1, we used the straightforward but powerful "Overview+Detail" architecture [6] to build SingDistVis, which combines the following two views:

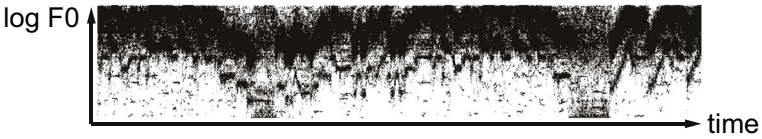

Fig. 1 Visual cluttering problem: Polylines are unsuitable for simultaneous visualization of numerous F0 trajectories

OverallView Two heatmap views are displayed: the F0 distribution of the entire song and the F0 variance at each time. The F0 distribution heatmap can be used to capture singing tendency while avoiding the visual cluttering problem, and the F0 variance heatmap can be used to find parts where high-rated and/or low-rated[1] singing voices are singing in various ways. From high variance parts for high-rated singing voices, users may discover unique singing based on the performers' individuality or difficult parts even for skilled singers. In contrast, from high variance parts for low-rated singing voices, users may discover difficult parts for beginners. Then, the users can interactively specify a partial section.

DetailedView This view draws polylines to visualize the focused local F0 trajectories with high time-frequency resolution. To avoid the visual cluttering problem, the number of polylines can be limited. Furthermore, each singing voice's singing assessment result is color-coded, which can help decide which F0 trajectory to focus on. Users can understand the tendency of the selected singer while comparing it with other singers during the whole song by overlaying the polyline of the singer on the OverallView.

The main contributions of this paper are as follows:

- This paper proposes the unique visualization design that applies a heatmap for OverallView and a polyline chart for DetailedView.
- This paper shows the appropriateness of the visual design by subjective user evaluations shown in Section 5.2.
- Through the user experiments presented in Section 6, this paper demonstrates that the interaction mechanism between the above two views is effective for comparing a large number of singings and discoverying interesting phrases.

This paper is an extended version of a domestic conference paper [16]. This paper introduces an extended implementation and additional user evaluations in addition to the contents presented by the domestic conference paper.

2 Related work

This section introduces related studies including visualization for singings and general-purpose information visualization techniques for time-series datasets such as F0 trajectories.

2.1 Visualization for singings

Visualization has been applied to the analysis and understanding of music and other multimedia content. Visualization is especially useful for the overview of a large number of

[1] Singing voices are manually rated in this study. A detailed example is shown in Section 4.

works and is actually applied to represent their distributions such as movie emotion map [7]. Also, various two-dimensional time-frequency representations have been proposed for acoustic applications [29]. The paper described that "The distinctive two-dimensional time-frequency visualization can produce better features for audio detection and classification tasks" [29].

As examples of music visualization, Rau et al. [35] presented a system for the assistance of composition processes by visualizing melodies applying using piano roll, cycle plot, node-link chart, and similarity-based scatterplot. Carter-Enyi et al. [5] verified the usability of a famous visualization technique that connects similar parts of a pair of music by drawing arcs. There have also been several survey papers on music visualization [18, 22].

Studies on visualization of singings, focusing on transitions of F0, have a long history. Numerous studies [13, 25, 32] have focused on the visualization of F0 trajectory in a song to improve the singing ability. There have been studies that summarize F0 trajectories into note units and visualize pitch deviation [31, 43]. Furthermore, visualizations using F0 and volume as two axes [23] and using F0 and F0 difference as two axes [17] have been reported to represent the differences in singing styles. However, these studies just visualized one or a few singing voices.

Meanwhile, several studies have focused on the visualization of relationships between multiple singings. Hamasaki et al. proposed Songrium RelayPlay [11], which automatically connects each song and plays them back as a single song while visualizing the number of times each song is played. Tsuzuki et al. proposed Unisoner [40], a system that can create a virtual chorus by playing multiple songs simultaneously. They also proposed a technique for improving the accuracy of F0 estimation by using 4,941 F0s as well as a visualization of F0 distribution locally as a heatmap. These techniques just visualize a small number of singings or a short part of a song; in other words, these studies have not visualized the F0 distribution of a large number of singings through a whole song.

Recent studies archived visualization of a large number of singings. Shen et al. [36] visualized the distribution of features of singings in a latent space as scatterplots. Similar representations applying dimension reduction and scatterplots to the high-dimensional features have been applied to a large number of songs or singings for the purpose of singer identification [44], singer diarization [38] and singer timbre evaluation [39]. However, it is difficult to represent the transitions of F0 of a large number of singings by such visual representations.

Table 1 summarizes the limitations of the related work and the main advantages of SingDistVis.

2.2 Information visualization for time-series datasets

F0 trajectories can be viewed as a time-series dataset. The following are typical visualizations that can be applied to such time-series dataset with K samples each of which has real values at time t:

- A: Polyline charts or scatterplots with T timesteps on one axis and real values on the other [34, 41, 42].
- B: Density heatmaps encoding the density of 2D (time-frequency) data into color or brightness [20, 42].
- C: A matrix with T timesteps on one axis and the K samples align along the other axis, and the real values are represented as brightness and hue [15, 33] or line thickness [21].

Table 1 Comparison of studies on visualization of singings

Reference	Features and limitations
[13, 17, 23, 25, 31, 32, 43]	Visualization of F0 trajectory in a song.
	Just one or a few singings are visualized.
[11, 40]	Visualization of F0 trajectory in a song.
	A small number of singings are a short part of a song is visualized
[36, 38, 39, 44]	Visualization of distribution of a large number of singings by scatterplots
	Transitions of F0 is not directly visualized.
SingDistVis	Visualization of distribution of F0 trajectories as a heatmap.
	A large number of singings through a whole song is visualized.
	Detailed F0 trajectories in a user-interested part is also finely visualized

Polyline charts (A) are more suitable for continuity representation than scatterplots (A) and the limited number of polylines can be useful to display them in detail. In addition, density heatmaps (B) can be used to visualize charts with large K. Finally, in the case of (C), line thickness is not adequate for large K and the color is not suitable for distinguishing changes in values [26]. We adopted the density heatmap (B) for OverallView and the polyline chart (A) for DetailedView.

Accordingly, SingDistVis can provide singing skill visualization by color-coded polylines.

3 SingDistVis: visualization of the numerous F0 trajectories

This section first describes the system design of the presented system, SingDistVis, and then presents the implementation details of two views of SingDistVis: OverallView and DetailedView.

3.1 Objective and system design

As described in Section 1, this study aims to the visualization of F0 trajectories of a large number of singings satisfying the following requirements:

$R1$: Discovery of particular parts of a song that are difficult only for novice singers or even expert singers.
$R2$: Understanding of differences in singing between novice and expert singers at the particular parts of the song.
$R3$: Observation of how user-interested novice/expert singers have characteristics of their singing.

SingDistVis, the system for visualization of a large number of singings, has been developed to satisfy the above requirements. This section describes the system overview and processing flow of SingDistVis. Figure 2 depicts the operation scenario in which users specify the area of interest in the OverallView and select one singing voice in the DetailedView. By using this technique, users can briefly observe the F0 distribution using OverallView. The users can specify the area of interest in OverallView by also using the F0 variance heatmap and then displaying the polylines in the area in DetailedView. They can specify a specific singing in DetailedView to listen to it and observe its F0 trajectory.

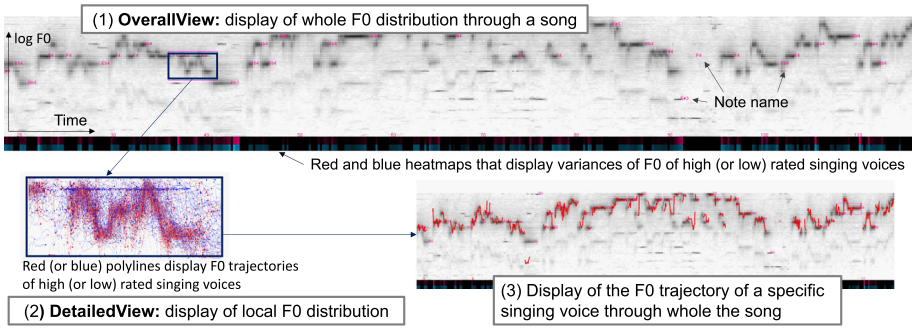

Fig. 2 Snapshots and operating procedures of SingDistVis. (1) OverallView: The F0 distribution for numerous singing voices of the entire song and the F0 variance at each time are each visualized as a heatmap. (2) DetailedView: F0s in the rectangular area specified in the OverallView are visualized as colored polylines. The color of each polyline corresponds to the singing assessment result, red is for high, and blue is for the low rating. (3) Superimposed display in OverallView of an F0 trajectory selected in the DetailedView

3.2 Supposed data structure

Let S be a set of K singing voices as follows:

$$S = \{s_1, ..., s_k, ..., s_K\}, \ \ s_k = \{p_{k1}, ..., p_{kt}, ..., p_{kT}, e_k\}, \tag{1}$$

where s_k denotes the F0 series of the k-th singing voice, p_{kt} denotes its log F0 at time t, and T is the total number of sampled timesteps. The value of p_{kt} is set to zero if silent. e_k denotes the assessment value for the k-th singing, which is based on the 5-point Likert scale (as interval scale) for evaluation of the singing skill. The F0 series among S are aligned automatically because they all need to be the same length and time synchronized.

Figure 3 illustrates the supposed structure of the input datasets. The k-th singing has an assesment value e_k, and a sequence of F0 values p_{k1} to p_{kT}. These values are consumed while generating both OverallView and DetailedView.

3.3 OverallView: heatmap representation of the entire song

OverallView visualizes the overall distribution of F0 trajectories by a heatmap. Users can specify a rectangular region in the heatmap so that users can explore the details of the F0 trajectories. Also, OverallView features another thin heatmap that represents the F0 variance. This section presents the implementation details of these functions.

3.3.1 Generating heatmap

The left part of Fig. 4 illustrates the generation of the heatmap in OverallView. The density distribution of the numerous F0 trajectories is represented by a rectangular heatmap R divided into $N \times M$ subranges. Here, the horizontal axis of the heatmap represents the time axis along which the song duration is divided into N parts, and the vertical axis represents the frequency axis along which the target range of the log F0 is divided into M parts. We first compute a discrete time-frequency density function $R(n, m)$ where n and m denote the horizontal and vertical position of a subrange, and then compute a F0 distribution heatmap represented by intensity $I(n, m)$.

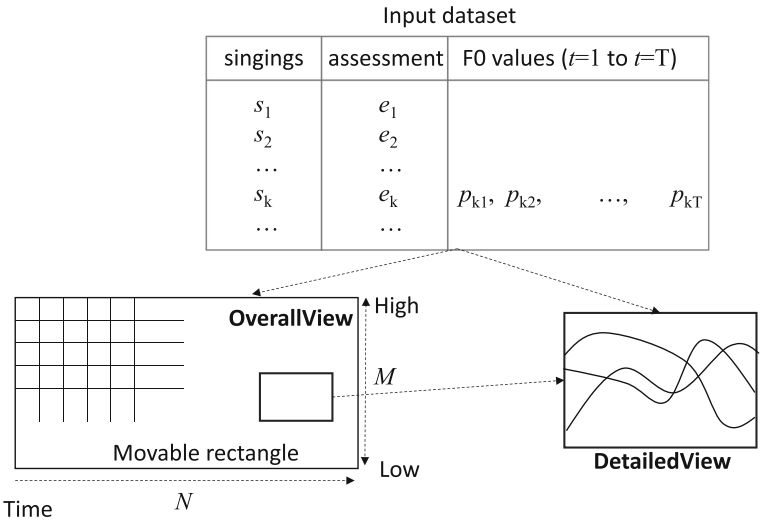

Fig. 3 Supposed structure of an input dataset. Each singing has an assesment value and a sequence of F0 values

$R(n, m)$ is computed by counting the number of polylines contained in the subrange of (n, m). Take note of the description: $t_{\text{start}} = t_1$, $t_{\text{end}} = t_{N+1}$, $p_{\min} = f_1$, $p_{\max} = f_{M+1}$. We calculate to which rectangular region of the above lattice structure p_{kl} belongs. In other words, we specify the values u and v that satisfy $t_u < k < t_{u+1}$ and $f_v < p_{kl} < f_{v+1}$, where u denotes the position of the rectangular region from the left and v denotes the position of the rectangular region from the bottom. By using the $R(n, m)$, $I(n, m)$ is then computed as follows:

$$I(n, m) = 1.0 - \alpha \cdot R(n, m)^{\gamma}, \tag{2}$$

Here, α and γ are adjustable variables affecting brightness and contrast, respectively, as shown in Fig. 5. The larger α is, the lower the brightness is in the area where r_{uv} is high.

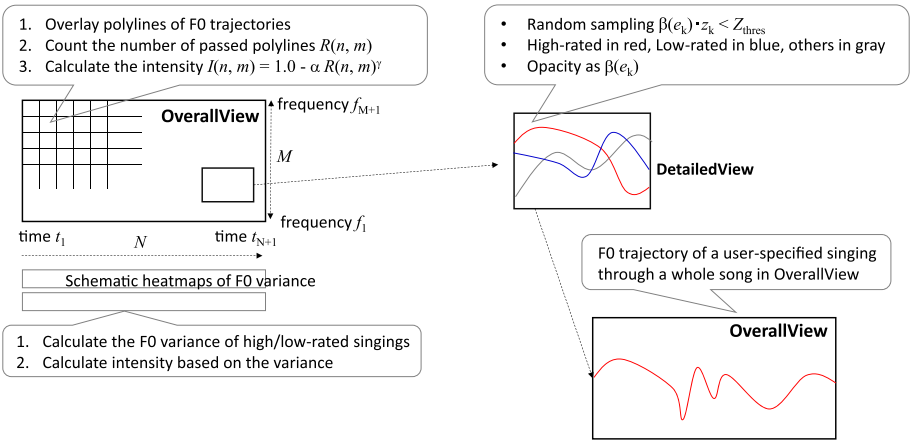

Fig. 4 Processing flow of the generation of OverallView and DetailedView

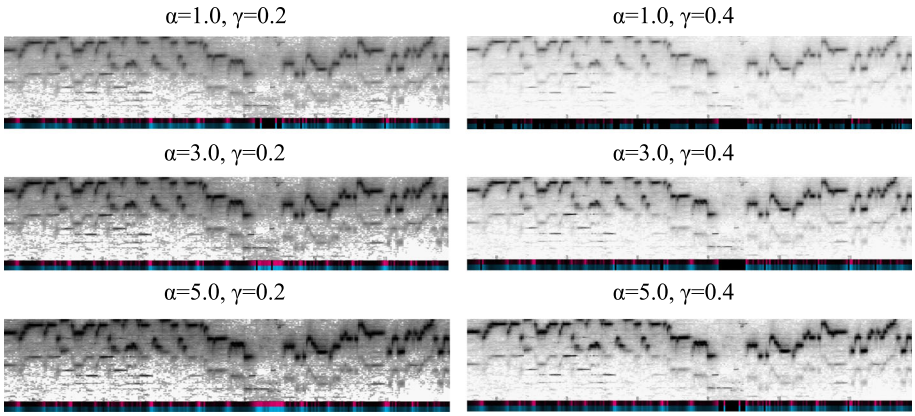

Fig. 5 Visualization examples under several settings of α and γ

Meanwhile, the greater the value of γ, the greater the contrast between the areas where r_{uv} is large. Our implementation provides a user interface to interactively adjust these variables.

3.3.2 Manipulating rectangular regions on a heatmap

A rectangular region depicting a certain time range and the F0 range is superimposed in OverallView. This rectangular region can be switched between "movable mode" and "stop mode" by clicking the mouse. Shift operations are applied in movable mode. The F0 trajectories that correspond to the inside of the rectangular region are displayed as polylines in DetailedView in stop mode.

3.3.3 Schematic representation of F0 variance

OverallView also displays the F0 variance at each timestep at the bottom with a black background. The variance values for the high-/low-rated singing voices are visualized using a red/blue heatmap, respectively. A 5-point Likert scale (as an interval scale) was used to assign ratings. A singing voice that receives a 4 or 5 rating is considered high-rated; a 1 or 2 rating is considered low-rated.

The above global heatmaps of the F0 distribution and variance can be used as useful hints to select a local area (e.g., a phrase) and to show it in detail. The user can select its area by using the mouse to move/resize a rectangular shape that is placed on the F0 distribution heatmap. In the following DetailedView, F0 trajectories that correspond to the interior of the specified area are presented as polylines.

3.4 DetailedView: polyline representation of the local area of user's interest

The F0 trajectories that pass the rectangular region of OverallView are drawn as a set of polylines in DetailedView. This view features the following functions for drawing polylines from the F0 trajectories of a large number of singings.

- F0 trajectories inside the recutangular region described in Section 3.3.2 are drawn.

- The appropriateness of the timing is visualized by highlighting the start and end of each F0 trajectory (i.e., onset and offset). DetailedView draws a point of a specific size at p_{kl} if $p_{k(l-1)} = 0$ or $p_{k(l+1)} = 0$.
- The number of polylines to be drawn is controlled by a sampling method described below.
- The color and opacity of polylines are determined by the assessment of the singing.
- OverallView displays the F0 of the user-selected singing as a polyline when a particular polyline in DetailedView is selected.

The right part of Fig. 4 illustrates the above functions in DetailedView.

To avoid visual cluttering, DetailedView enables controlling the number of polylines to draw simultaneously by using random sampling defined as follows:

$$\beta(e_k) \cdot z_k > Z_{\text{thres}} \tag{3}$$

where z_k is a uniform random value ranging over $0.0 \leq z_k \leq 1.0$, and Z_{thres} is an adjustable threshold value used to control the number of polylines. Our implementation provides a user interface to interactively adjust this threshold. $\beta(e_k)$ is a coefficient according to the singing assessment e_k. In our current implementation, $\beta(e_k)$ is defined as

- $\beta(e_k) = 1.0$ when $e_k = 2$ or 5,
- $\beta(e_k) = 0.8$ when $e_k = 4$,
- $\beta(e_k) = 0.5$ when $e_k = 1$ or 3.

to emphasize low-/high-rated singing voices. Since it turns out that the F0 of singing voices with $e_k = 1$ tends to be too much deviated and is not useful to be emphasized, we deemphasize it and emphasize $e_k = 2$ instead.

The number of polylines can be adjusted so that only lines of low-rated singings are drawn more frequently, or only lines of high-rated singings are drawn more frequently. The number of polylines can be adjusted so that only lines of low-rated (or high-rated) singings are drawn more frequently. When the rectangular region of OverallView is set to stop mode or the value of Z_{thres} is changed, sampling is performed in our implementation.

Figure 6 shows visualization examples on different values of Z_{thres}. Here, F0 trajectories of high-rated singing voices ($e_k = 4$ or $e_k = 5$) are drawn in red; low-rated voices ($e_k = 1$

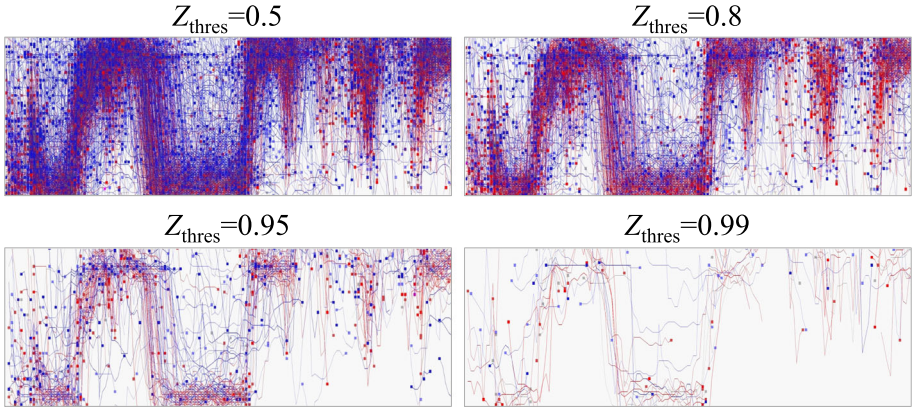

Fig. 6 Controlling the number of polylines by changing Z_{thres} (red/blue indicates high/low-rated singing voices)

or $e_k = 2$) are drawn in blue; other voices ($e_k = 3$) are drawn in gray. Opacity of polylines is defined as $\beta(e_k)$.

Users can select a particular F0 by clicking a polyline or clicking a voice's identification code shown in the text area. The selected F0 polyline can also be superimposed on the F0 distribution heatmap in OverallView to compare it with other singing voices throughout the song.

Also, they can select a singing using a dialog window that displays the list of singings drawn in DetailedView ordered by the distance with the average F0 or minimum distance with another singing. When a user selects a singing, the whole transition of F0 is displayed as a polyline in OverallView, and the audio source of the selected singing is played.

4 Examples

4.1 Specification of user-interested regions on OverallView

Figure 7 depicts an example of OverallView and six examples of DetailedView derived from various positions of the rectangular region on OverallView. In OverallView (upper part of Fig. 7), the darker the F0 distribution heatmap, the more singers sing at the same pitch. The red and blue heatmaps at the bottom (i.e., the F0 variance for the high-/low-rated singing voices) can help users choose which regions are of interest. A region with high variance, for example, may tend to be sung in different ways and let users discover interesting good singing voices with unique styles. Or, such a high variance region may simply be difficult and wrongly sung by singers. Users might want to practice its region by singing repeatedly.

4.2 Detailed observation of the singing evaluation and F0 trajectory on DetailedView

Figure 7(1) and (2) show long tones after a descending scale. In Fig. 7(1) with a red and blue high-variance region in OverallView, red polylines form a thick band with a wide range in F0 even among high-rated singers. Skilled singers might want to practice in such a region. On the other hand, in Fig. 7(2) with a blue high-variance region in OverallView, low-rated singers tend to miss the F0s more widely than high-rated singers.

Figure 7(3), (4), and (5) depict phrases in which the scale moves significantly up and down. These examples demonstrate that the sung F0s differed significantly between high-and

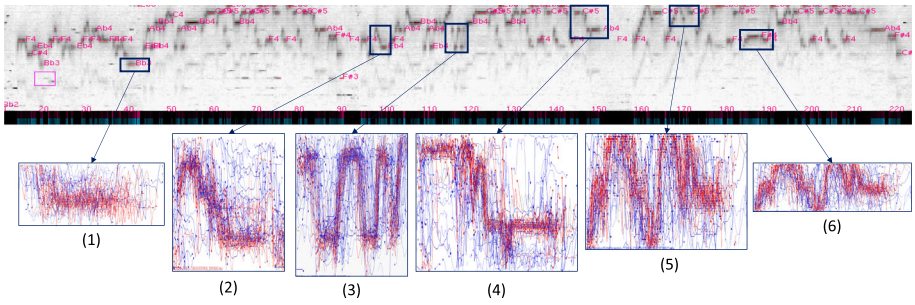

Fig. 7 An example of OverallView and six examples of DetailedView. In DetailedView, dark red indicates high singing assessment value ($e_k = 5$) and dark blue indicates low singing assessment value ($e_k = 2$)

low-rated singers (i.e., the F0s of low-rated singers show wider deviation). Figure 7(5) and (6) show a wide range of F0 at the end of the phrase, even for high-rated singers.

4.3 Superimposed polylines on OverallView

Figure 8 shows examples of specific polylines being overlaid in OverallView after a user selects them by clicking operations on Fig. 7(1).

The polyline shown in Fig. 8(1) was selected from the area in which the red polylines were concentrated in Fig. 7(1). Throughout the song, the singer corresponding to this polyline sang with a good F0. Most high-rated singers had similar polylines (F0 trajectories). As shown in Fig. 8(2), however, a few high-rated singers had different polylines from most of the singers. We listened to singing voices corresponding to such polylines and found that the main reasons for such deviations include unique singing styles as well as errors in the automatic temporal alignment and F0 estimation.

Both polylines shown in Fig. 8(3) and (4) correspond to low-rated singing voices. Although the polyline in Fig. 8(4) was significantly deviated, the polyline in Fig. 8(3) was not much deviated. In the latter case, even though the F0 was not deviated, other vocal properties that do not appear in the F0 trajectory, such as vocal timbre, are considered reasons for the low-rated voice.

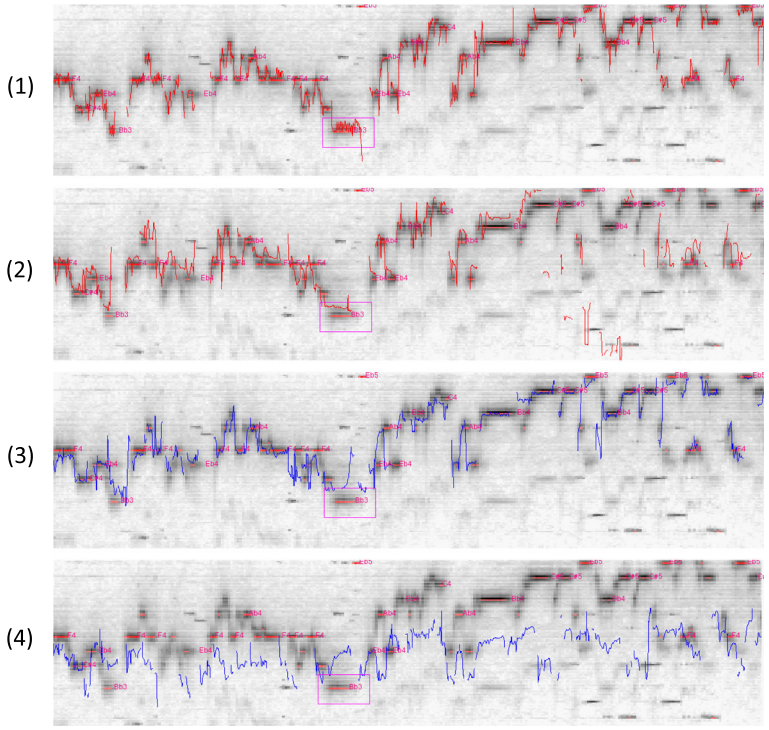

Fig. 8 F0 trajectories of some singers displayed as polylines. Here (1) and (2) show examples of high-rated singers, while (3) and (4) are examples of low-rated singers. The figures depict that singers in (1) and (3) had similar F0 trajectories as many other singers, while (2) and (4) had different ones from others

4.4 Potential application scenario

The followings are examples of typical application scenarios of SingDistVis.

- **For vocal coaches:** The OverallView can display the F0 variance of the high-/low-rated singing voices using the red/blue heatmap. It helps vocal coaches understand which parts of the melody need training for both novice and expert singers, or which parts need training especially for novice singers. For the former, the vocal coaches can use the DetailedView to observe how both novice and expert singers cannot sing the accurate F0, as demonstrated in Fig. 7(1). For the latter, the DetailedView can represent how novice and expert singers sing the same melody differently, as demonstrated in Fig. 7(2). The vocal coaches can then plan different vocal training for novice and expert singers.
- **For novice singers:** The OverallView can display the F0 polyline of the particular singer superimposed on the F0 distribution heatmap, as shown in Fig. 8. A novice singer can use this to compare his/her own singing with others throughout the song, and find which parts of the melody he/she sings differently. Then, by specifying a notable part of the song, the singer can use the DetailedView to compare his/her own singing in detail with high-rated singing voices, as shown in Fig. 7. The singer can thus understand how the high-rated singers are different from him/her.

5 Subjective evaluations

We conducted subjective evaluations to investigate suitable visualization parameters (α, γ, and Z_{thres}) as introduced in Section 5.1, and to prove the effectiveness of the visualization design as introduced in Section 5.2.

In the evaluations, we constructed a dataset consisting of randomly sampled 1,000 singing voices from the unaccompanied acoustic data of 231,278 singing voices singing "Let It Go" in DAMP [1] called "Let It Go Vocal Performances." Here, F0 trajectory of each singing was estimated automatically from the acoustic data using pYIN [24], and all F0 trajectories were aligned automatically by maximizing the cross-correlation. The size of the F0 distribution heatmap in the OverallView was set to $N = 1000$, $M = 480$, and the target frequency was set to four octaves ranging from 110 Hz to 1760 Hz (A1 to A5 in the musical scale).

To assess the singing skill of all 1,000 singing voices, we hired an evaluator who had 12 years of piano experience and asked her to rate each singing voice on a 5-point Likert scale (as an interval scale) as e_k based on accuracy, musicality, and vocal quality by listening to its chorus section. She used the original version sung by Idina Menzel as a reference during the evaluation process.

We invited 18 participants who are all undergraduate students majoring in computer science for two evaluations introduced in Sections 5.1 and 5.2.

5.1 Selection of parameter values

There have been several studies that demonstrate the importance of setting appropriate parameters to improve the visualization results [27, 37, 42]. We therefore generated multiple visualization images by adjusting α, γ (See (2)) and Z_{thres} (See (3)).

5.1.1 Selection of heatmap parameter values of OverallView

To investigate visualization parameters of the F0 distribution heatmap in OverallView (α and γ), we prepared twenty different images with the combination of $\alpha = \{1.0, 2.0, 3.0, 4.0, 5.0\}$ and $\gamma = \{0.1, 0.2, 0.3, 0.4\}$. The twenty images on one screen were presented simultaneously to each participant. All images are arranged in the order of α increasing from left to right and γ increasing from bottom to top (i.e., four rows and five columns arrangement).

From the 20 images, we asked the participants to select three images that best corresponded to the following two questions:

$Q1$: Which image would the participant most like to use for the overall display of the singing?
$Q2$: Which image is suitable to zoom in and select the part you want to check?

To reduce the influence of the location in the song, three song excerpts (each of 80 seconds) were applied to the above evaluations.

Figure 9 shows the aggregated results of the participants' selections. This result shows that the combination of parameter values $\alpha = 3.0$ and $\gamma = 0.3$ was one of the most preferable combinations for both Q1 and Q2. Therefore, we determined these were a set of suitable parameters, and used them in the evaluation experiments shown in Section 5.2.

5.1.2 Parameter for the number of polylines in DetailedView

To investigate the visualization parameter for DetailedView (Z_{thres}), we generated ten different images according to the value of $Z_{\text{thres}} = \{0.5, 0.7, 0.8, 0.85, 0.9, 0.93, 0.95, 0.97, 0.98, 0.99\}$. The generated ten images were also presented to participants with a single screen. All images were arranged in five rows and two columns in the order of Z_{thres} increasing from the upper left to the lower right in the screen.

From the ten images, we asked the participants to select two images that best corresponded to the following questions:

$Q3$: Which image would the participant feel easy to find out whether a high-rated singer is singing with an accurate pitch, a singer who is high-rated but dares to shift the pitch, or a singer who has an accurate pitch but is low-rated?
$Q4$: Which image would be most useful in determining whether a singer who sings differently from the majority (the other singers) is purposefully managing their pitch as a singing method, or whether the out-of-tune is unintentionally due to a lack of singing skill or a limited bit of training?
$Q5$: Which image would the participant feel easy to analyze while comparing the singing assessment result with the accuracy of the pitch?

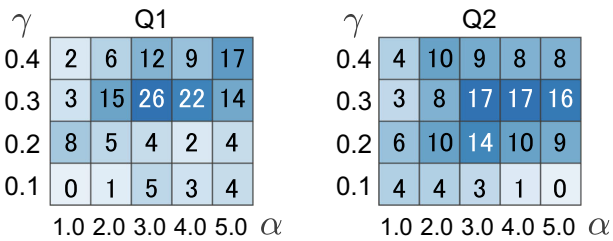

Fig. 9 Distributions of answers for Q1 and Q2

$Q6$: Which image would the participant feel likely to find a singer the participant would like to focus on?

To reduce the influence depending on the phrase, four randomly selected local areas were presented.

Figure 10 shows the aggregated results of the participants' selections. This result denotes the image with $Z_{\text{thres}} = 0.95$ was the most answered for three questions Q3 to Q5, and also the most answered by totaling all the questions. Based on the result, we adopted $Z_{\text{thres}} = 0.95$ in the experiments described in Section 5.2.

5.2 Validation of visualization design

This section introduces user evaluations that validate the appropriateness of the proposed visual design.

5.2.1 Evaluation of combination of OverallView and DetailedView

To validate our proposed Overview+Detail design, four combinations of heatmaps (H) and polylines (L) for OverallView (O) and DetailedView (D) were generated as images to use evaluation. In other words, we prepared the following four implementations:

- OH-DH (heatmaps for both OverallView and DetailedView)
- OL-DL (polylines for both OverallView and DetailedView)
- OL-DH (polylines for OverallView and heatmaps for DetailedView)
- OH-DL (heatmaps for OverallView and polylines for DetailedView)

Here, OH-DL is our proposed design while others are comparative designs. We generated four images using each of these implementations.

Then, we asked participants to comparatively evaluate the four images. The combinations of questions Q7 to Q11 described below and the generated images were presented randomly to the participants. The participants evaluated by answering the following questions (the mark * means negative questions) on a 7-point Likert scale (as an interval scale). The questions Q7 to Q10 are based on the System Usability Scale (SUS) [3] developed for system usability study.

$Q7$: I would like to use it frequently.
$Q8*$: I found it unnecessarily complex.
$Q9$: I thought it was easy to use.
$Q10*$: I needed technical support to use it.
$Q11$: I thought it is suitable for simultaneous visualization of numerous singing voices.

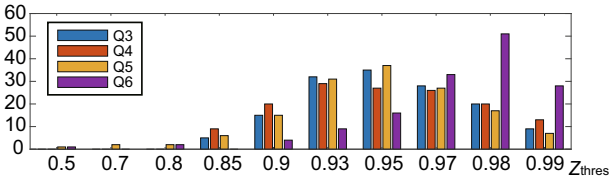

Fig. 10 Distributions of answers for Q3 to Q6

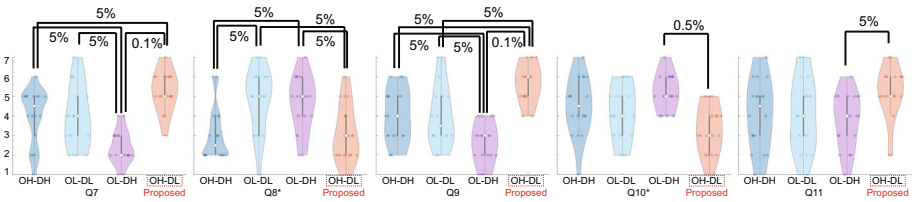

Fig. 11 Distribution of answers for Q7 to Q11 and results of hypothesis testing. Numbers denote the significance levels

Figure 11 shows the distribution of the answers. The Wilcoxon's rank-sum test was performed and the pairs for which the null hypothesis was rejected are shown in the figure along with their significance levels. This figure shows that the proposed design (OH-DL) had higher values for positive questions (Q7, Q9, and Q11) and lower values for negative questions (Q8* and Q10*). The exception was for Q8*, where OH-DH was also highly preferred.

5.2.2 Evaluation of individual visualization

Furthermore, we generated four images from the following implementation:

- OL (polylines for OverallView)
- OH (heatmaps for OverallView)
- DL (polylines for DetailedView)
- DH (heatmaps for DetailedView)

where OH and DL are used in our proposed designs.

Then, we asked participants to comparatively evaluate the images with the following questions, Q12 and Q13, on a 7-point Likert scale (as an interval scale).

$Q12$: It is easy to find the parts where many singers are singing at the same pitch.
$Q13$: It is easy to find areas where participants can find large differences in singing styles between singers.

Figure 12 shows the distribution of the answers. We also present the results of Wilcoxon's rank-sum test. Q12 is a question that assumes a task using OverallView, and the proposed method (OH) was highly evaluated. Although DL was also well-regarded, many participants stated that judging the density of polylines was simple. Q13 is a question that assumes a task using DetailedView, and the proposed method (DL) was also highly evaluated.

6 User study

We conducted a user study to comparatively evaluate SingDistVis. In addition to the implementation of SingDistVis, we also prepared a simple baseline implementation that displays only the F0 trajectory of a user-selected singer as a polyline chart. In this user study, we employed the singing dataset from four songs labeled for singing quality for 100 singers per song [10] [2].

We invited eight participants who were all undergraduate students majoring in computer science. We provided the participants with an instruction manual and asked them to practice

[2] https://github.com/chitralekha18/SingEval

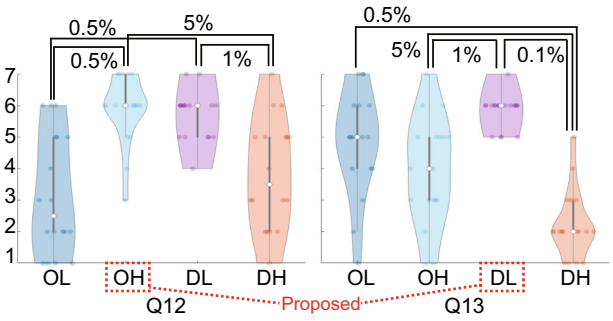

Fig. 12 Distribution of answers for Q12 to Q13

operating the software until they understood the operating procedures described therein. Then, we asked them to use either SingDistVis or the baseline implementation with one of the aforementioned song datasets, and then freely search for the following four types of singers:

$Type1$: High-rated singers who sang at the correct pitch.
$Type2$: High-rated singers who deliberately shifted the pitch.
$Type3$: Low-rated Singers who sang at the correct pitch.
$Type4$: Singers who are subjectively considered to be noteworthy.

Here, two of the four songs were randomly assigned to SingDistVis, and the other two songs were assigned to the baseline for each participant. They were allowed to take for a maximum of 30 minutes for each of SingDistVis and the baseline to select singers corresponded to Type 1 to Type 4.

After the task, we presented the following questions (the mark * means negative questions) to the participants and asked to answer them on a 7-point Likert scale (as an interval scale). Some questions are similar to Q1-Q13.

$Q14$: I like to use it for the overall display of the singing.
$Q15$: It is suitable to zoom in on the interested parts.
$Q16$: It is easy to find high-rated singers singing at the correct pitch, high-rated singers daring to shift the pitch, or low-rated singers singing at the correct pitch.
$Q17$: It is easy to discuss whether a singer who sings intentionally shifting the pitch as a singing technique, or whether the pitch is unintentionally off due to lack of skill or practice.
$Q18$: It is easy to analyze while comparing singing evaluation results and pitch accuracy.
$Q19$: It is easy to find singers who are subjectively felt to be noteworthy.
$Q20$: I would like to use it frequently.
$Q21*$: I found it unnecessarily complex.
$Q22$: I thought it was easy to use.
$Q23*$: I needed technical support to use it.
$Q24$: I thought it is suitable for simultaneous visualization of hundreds or thousands of singing voices.
$Q25$: It is easy to find the parts where many singers are singing at the same pitch.
$Q26$: It is easy to find areas where participants can find large differences in singing styles between singers.
$Q27$: It is easy to find high-rated singers who sang at the correct pitch.

$Q28$: It is easy to find high-rated singers who deliberately shifted the pitch.
$Q29$: It is easy to find low-rated singers who sang at the correct pitch.

Figure 13 shows the mean of the answers of eight participants. The results show that SingDistVis was preferable to baseline for many of the questions. In particular, a large difference was found in Q24 (simultaneous visualization of a large number of singings) and Q25 (finding areas where many singers sing at the same pitch), indicating that SingDistVis is working properly as intended in this study.

On the other hand, the baseline was preferred for Q19, Q21, Q22, and Q23, indicating that the practice time for this user study was short, or in other words, that a certain amount of practice time is required to master SingDistVis. Meanwhile, the two participants who supported SingDistVis in these questions had common properties: they love music but do not have expert-level listening skills for singing. This suggests that SingDistVis is particularly useful for users who are inquisitive about music and who would benefit from supplementing their individual ability with visualization in the singing-search tasks. We can also see that the baseline was more favorable in Q28. This suggests that some of the use scenarios of SingDistVis could be realized with a simpler user interface. We would like to attempt to improve the user interface from this perspective.

In the comments of the participants, the advantages of SingDistVis were mentioned as

- "easy to compare with other singers,"
- "easy to find similar singers," and
- "easy to find the parts where there are variations in pitch and where everyone sings at the same pitch."

On the other hand, "complicated," "too many functions," and "too much information," were also pointed out.

We also checked the singing selected by the participants. Songs selected using baseline had a lot of overlap among participants, while SingDistVis had little overlap. This suggests that SingDistVis can be used to explore a variety of songs according to the user's interests.

As summary, this user experiment demonstrated that the main strengths of SingDistVis were as follows:

- simultaneous visualization and comparison of a large number of singings,

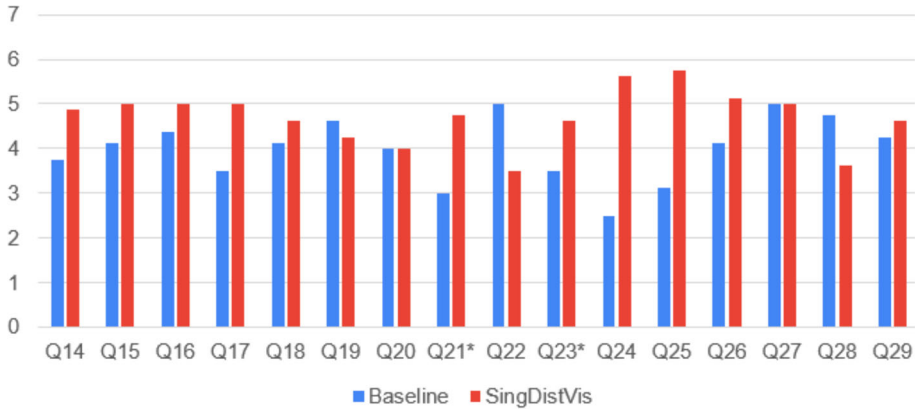

Fig. 13 The means of answers for Q14 to Q29

- easiness to find the parts where there are variations in pitch and where everyone sings at the same pitch, and
- exploration of a variety of singers according to the user's interests.

Meanwhile, we found that the complicated operations and functions might be disadvantages for some users; therefore, we would like to improve the current implementation by simplifying the functions.

7 Conclusions

We proposed a visualization technique for the F0 distribution of numerous singing voices based on the "Overview+Detail" architecture. We also presented the investigation of appropriate visualization parameters in the proposed design, as well as the assessment of its efficacy.

Our contributions are summarized as follows:

- We present the first simultaneous visualization of numerous time-series song data (not only for the short-time local part).
- We determined suitable visualization parameters for our Overview+Detail framework based on a subjective experiment using 1,000 singing voices.
- We confirmed that combining heatmap-based OverallView and polyline-based DetaledView was preferable over other combinations through subjective evaluation.

Though the subjective evaluation archived totally preferable results, we pointed out that SingDistVis requires a certain amount of practice time because of its many functions and complicated structures. As a future issue, we would like to simplify the implementation to reduce the time for practice without losing the advantages of the current implementation.

Funding This research is not supported by any funding schemes or governments.

Data Availability The text dataset describing F0 trajectries introduced in this paper is available on-demand. This dataset is generated from audio files included in an open dataset [1].

Declarations

Conflicts of interest The authors have declared that there is no conflict of interest exists.

References

1. WELCOME TO STANFORD'S DAMP: Stanford Digital Archive of Mobile Performances, a repository of geo-tagged mobile performances to facilitate the research of amateur practices. https://ccrma.stanford.edu/damp/
2. Ali M, Jones MW, Xie X, Williams M (2019) TimeCluster: dimension reduction applied to temporal data for visual analytics. Vis Comput 35:1013–1026

3. Brooke J (1996) SUS: A "quick and dirty" usability scale. In Jordan PW, Thomas B, McClelland IL, Weerdmeester B (eds) Usability evaluation in industry, chapter 21, pp. 189-194. Taylor & Francis, London
4. Buono P, Plaisant C, Simeone AL, Aris A, Shmueli G, Jank W (2007) Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting. In Proceedings of the International Conference on Information Visualization (IV 2007), pp 191–196
5. Carter-Enyi A, Rabinovitch G, Condit-Schultz N (2021) Visualizing intertextual form with arc diagrams: Contour and schema-based methods. In Proceedings of the 22nd International Society for Music Information Retrieval Conference(ISMIR 2021), pp 74–80
6. Cockburn A, Karlson A, Bederson BB (2009) A review of overview+detail, zooming, and focus+context interfaces. ACM Comput Surv 41(2):1–31
7. Cohen-Kalaf M, Lanir J, Bak P, Mokryn O (2022) Movie emotion map: an interactive tool for exploring movies according to their emotional signature. Multimed Tools Appl 81:14663–14684
8. Gómez E, Blaauw M, Bonada J, Chandna P, Cuesta H (2018) Deep learning for singing processing: achievements, challenges and impact on singers and listeners. CoRR arXiv:1807.03046
9. Goto M, Saitou T, Nakano T, Fujihara H (2010) Singing information processing based on singing voice modeling. In Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2010), pp 5506–5509
10. Gupta C, Li H, Wang Y (2020) Automatic leaderboard: Evaluation of singing quality without a standard reference. IEEE/ACM Trans Audio Speech Lang Process 28:13–26
11. Hamasaki M, Ishida K, Nakano T, Goto M (2020) Songrium RelayPlay: A web-based listening interface for continuously playing user-generated music videos of the same song with different singers. In Proceedings of the International Computer Music Conference 2021 (ICMC 2021), pp 426–429
12. Hochheiser H, Shneiderman B (2004) Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. Inf Vis 3(1):1–18
13. Hoppe D, Sadakata M, Desain P (2006) Development of real-time visual feedback assistance in singing training: A review. J Comput Assist Learn 22(12):308–316
14. Humphrey EJ, Reddy S, Seetharaman P, Kumar A, Bittner RM, Demetriou A, Gulati S, Jansson A, Jehan T, Lehner B, Kruspe A, Yang L (2019) An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music. IEEE Signal Process Mag 36(1):82–94
15. Imoto M, Itoh T (2010) A 3d visualization technique for large scale time-varying data. In Proceedings of the international conference on information visualization (IV 2010), pp 17–22
16. Itoh T, Nakano T, Fukayama S, Hamasaki M, Goto M (2021) SingDistVis: User interface for visualizing the tendency of singing from a large number of singings (in japanese). In Proceedings of the 29th workshop on interactive systems and software (WISS), pp 1–6,
17. Kako T, Ohishi Y, Kameoka H, Kashino K, Takeda K (2009) Automatic identification for singing style based on sung melodic contour characterized in phase plane. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2009), pp 393–398
18. Khulusi R, Kusnick J, Meinecke C, Gillmann C, Focht J, Jänicke S (2020) A survey on visualizations for musical data. Comput Graph Forum (CGF) 39(6):82–110
19. Knees P, Schedl M, Goto M (2020) Intelligent user interfaces for music discovery. Trans Int Soc Music Inf Retrieval 3(1):165–179
20. Kraus M, Angerbauer K, Buchmüller J, Schweitzer D, Keim DA, Sedlmair M, Fuchs J (2020) Assessing 2d and 3d heatmaps for comparative analysis: An empirical study. In Proceedings of the 2020 ACM CHI conference on human factors in computing systems (ACM CHI 2020), pp 1–14
21. Krstajic M, Bertini E, Keim DA (2011) CloudLines: Compact display of event episodes in multiple time-series. IEEE Trans Vis Comput Graph 17(12):2432–2439
22. Lima HB, Santos CGRD, Meiguins BS (2021) A survey of music visualization techniques. ACM Comput Surv 54(7):143
23. Lin KWE, Anderson H, Agus N, So C, Lui S (2014) Visualising singing style under common musical events using pitch-dynamics trajectories and modified TRACLUS clustering. In Proceedings of the 13th International Conference on Machine Learning and Applications (ICMLA'14), pp 237–242
24. Mauch M, Dixon S (2014) pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP 2014), pp 659–663
25. Mayor O, Bonada J, Loscos A (2009) Performance analysis and scoring of the singing voice. In Proceedings of the AES 35th international conference, pp 1–7
26. Mazza R (2009) Introduction to Information Visualization. Springer
27. Micallef L, Palmas G, Oulasvirta A, Weinkauf T (2017) Towards perceptual optimization of the visual design of scatterplots. IEEE Trans Vis Comput Graph 23(6):1588–1599

28. Miranda F, Lage M, Doraiswamy H, Mydlarz C, Salamon J, Lockerman Y, Freire J, Silva CT (2018) Time Lattice: A data structure for the interactive visual analysis of large time series. Comput Graph Forum 37(3):23–35
29. Mistry, YD, Birajdar GK, Khodke AM (2023) Time-frequency visual representation and texture features for audio applications: a comprehensive review, recent trends, and challenges. Multimed Tools Appl 82:36143–36177
30. Moritz D, Fisher D (2018) Visualizing a million time series with the density line chart. In arXiv:1808.06019
31. Moschos F, Georgaki A, Kouroupetroglou G (2016) FONASKEIN: An interactive software application for the practice of the singing voice. In Proceedings of the 13th Sound and Music Computing Conference (SMC 2016), pp 326–331
32. Nakano T, Goto M, Hiraga Y (2007) MiruSinger: A singing skill visualization interface using real-time feedback and music cd recordings as referential data. In Proceedings of the 9th IEEE International Symposium on Multimedia (ISM 2007) Workshops, pp 75–76
33. Oliveira G, Comba J, Torchelsen R, Padilha M, Silva C (2013) Visualizing running races through the multivariate time-series of multiple runners. In Proceedings of the Conference on Graphics, Patterns and Images (SIBGRAPI 2018), pp 99–106
34. Perin C, Vernier F, Fekete J-D (2013) Interactive horizon graphs: Improving the compact visualization of multiple time series. In Proceedings of the 2013 ACM SIGCHI conference on human factors in computing systems (ACM CHI 2013), pp 3217–3226
35. Rau S, Heyen F, Wagner S, Sedlmair M (2022) Visualization for ai-assisted composing. In Proceedings of the 23th International Society for Music Information Retrieval Conference (ISMIR 2022), pp 151–159
36. Shen J, Wang R, Shen H-W (2020) Visual exploration of latent space for traditional chinese music. Vis Inf 4(2):99–108
37. Smart S, Szafir DA (2019) Measuring the separability of shape, size, and color in scatterplots. In Proceedings of the 2019 ACM CHI conference on human factors in computing systems (ACM CHI 2019), p 669:1–14
38. Suda H, Saito D, Fukayama S, Nakano T, Goto M (2022) Singer diarization for polyphonic music with unison singing. IEEE/ACM Trans Audio, Speech, Lang Process 30:1531–1545
39. Sun X, Gao Y, Lin H, Liu H (2023) Tg-Critic: A timbre-guided model for reference-independent singing evaluation. In Proceedings of the 2023 IEEE international conference on acoustics, speech, and signal processing (IEEE ICASSP 2013)
40. Tsuzuki K, Nakano T, Goto M, Yamada T, Makino S (2014) Unisoner: An interactive interface for derivative chorus creation from various singing voices on the web. In Proceedings of the 40th International Computer Music Conference and11th Sound and Music Computing Conference (Joint ICMC SMC 2014Conference), pp. 790-797
41. Uchida Y, Itoh T (2009) A visualization and level-of-detail control techniquefor large scale time series data. In Proceedings of the international conference on information visualization (IV 2009), pp 80–85
42. Wang Y, Han F, Zhu L, Deussen O, Chen B (2018) Line graph or scatter plot? automatic selection of methods for visualizing trends in time series. IEEE Trans Vis Comput Graph 24(2):1141–1154
43. Weiß C, Schlecht SJ, Rosenzweig S, Müller M (2019) Towards measuring intonation quality of choir recordings: A case study on Bruckner's Locus iste. In Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019), pp 276–283
44. Zhang X, Wang J, Cheng N, Xiao J (2022) Singer identification for metaverse with timbral and middle-level perceptual features. In Proceedings of the 2022 International joint conference on neural networks (IJCNN)
45. Zhao Y, Wang Y, Zhang J, Fu C-W, Xu M, Moritz D (2022) KD-Box: Line-segment-based KD-tree for interactive exploration of large-scale time-series data. IEEE Trans Vis Comput Graph 28(1):890–900

Authors and Affiliations

Takayuki Itoh[1] · Tomoyasu Nakano[2] · Satoru Fukayama[3] · Masahiro Hamasaki[2] · Masataka Goto[2]

Tomoyasu Nakano
t.nakano@aist.go.jp

Satoru Fukayama
s.fukayama@aist.go.jp

Masahiro Hamasaki
masahiro.hamasaki@aist.go.jp

Masataka Goto
m.goto@aist.go.jp

1 Ochanomizu University, 2-1-1 Otsuka, Bunkyo 112-8610, Tokyo, Japan

2 National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba 305-8568, Ibaraki, Japan

3 National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto, 135-0064 Tokyo, Japan