TimToShape: Supporting Practice of Musical Instruments by Visualizing Timbre with 2D Shapes based on Crossmodal Correspondences

Kota Arai
arai0922@cyber.t.u-tokyo.ac.jp
The University of Tokyo
Bunkyo, Tokyo, Japan

Yutaro Hirao
hirao@cyber.t.u-tokyo.ac.jp
The University of Tokyo
Bunkyo, Tokyo, Japan

Takuji Narumi
narumi@cyber.t.u-tokyo.ac.jp
The University of Tokyo
Bunkyo, Tokyo, Japan

Tomohiko Nakamura
tomohiko.nakamura.jp@ieee.org
The University of Tokyo
Bunkyo, Tokyo, Japan

Shinnosuke Takamichi
shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
The University of Tokyo
Bunkyo, Tokyo, Japan

Shigeo Yoshida*
shigeo.yoshida@sinicx.com
OMRON SINIC X Corporation
Bunkyo, Tokyo, Japan

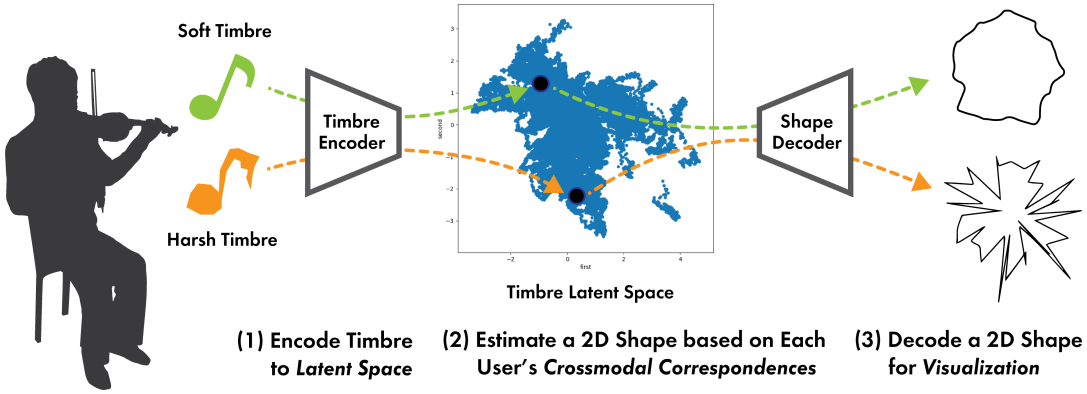

Figure 1: TimToShape estimates and visualizes 2D shapes corresponding to the timbre played by the musical-instrument learner based on their timbre–shape correspondences (crossmodal correspondences) to facilitate an intuitive understanding of how to adjust the performance to play the desired timbre. TimToShape uses the variational autoencoder (VAE) to encode the timbre to the latent space, and generates a 2D shape corresponding to the position of that timbre in the latent space by linear interpolation of shapes that the users previously answered as corresponding to few timbres.

ABSTRACT

Timbre is high-dimensional and sensuous, making it difficult for musical-instrument learners to improve their timbre. Although some systems exist to improve timbre, they require expert labeling for timbre evaluation; however, solely visualizing the results of unsupervised learning lacks the intuitiveness of feedback because human perception is not considered. Therefore, we employ *crossmodal correspondences* for intuitive visualization of the timbre. We designed *TimToShape*, a system that visualizes timbre with 2D shapes based on the user's input of timbre–shape correspondences. TimToShape generates a shape morphed by linear interpolation according to the timbre's position in the latent space, which is obtained by unsupervised learning with a variational autoencoder (VAE). We confirmed that people perceived shapes generated by TimToShape to correspond more to timbre than randomly generated shapes. Furthermore, a user study of six violin players revealed that TimToShape was well-received in terms of visual clarity and interpretability.

*Secondary affiliation: The University of Tokyo

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03.
https://doi.org/10.1145/3581641.3584053

CCS CONCEPTS

• **Applied computing** → *Sound and music computing*; • **Human-centered computing** → *Visualization techniques.*

KEYWORDS

crossmodal correspondences, timbre–shape correspondences, timbre, musical instrumental practice, variational autoencoder

ACM Reference Format:
Kota Arai, Yutaro Hirao, Takuji Narumi, Tomohiko Nakamura, Shinnosuke Takamichi, and Shigeo Yoshida. 2023. TimToShape: Supporting Practice of Musical Instruments by Visualizing Timbre with 2D Shapes based on Crossmodal Correspondences. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3581641.3584053

1 INTRODUCTION

Music players can convey information and feelings to their audience by adequately changing the sound timbre. However, because timbre is a high-dimensional and sensuous concept, it is difficult for learners to develop the skills necessary to play the desired timbre.

Although several methods have been proposed for supporting the acquisition of timbre-related skills [14, 39], they frequently require expert labeling for timbre evaluation. There is also a method using unsupervised learning that does not require manual labeling [18]. However, feedback solely based on the results of unsupervised learning is not intuitive enough to understand the relationship between timbre and visual feedback. We presume that this could lead to learners having difficulty bringing their timbre closer to the goal timbre because visualization that is not intuitive hinders mastery [40]. Moreover, learners might become dependent on the feedback and are unable to maintain performance when it is no longer available because a practice with feedback in which auditory and visual information is incongruent results in worse retention scores than when it is congruent [2].

In this paper, we focused on *crossmodal correspondences* (in particular, *timbre–shape correspondences*) to provide intuitive visual feedback on timbre to musical instrument learners. Crossmodal correspondence is a non-arbitrary associative relationship between different modalities [43]. In auditory–visual relationships, the relationship between timbres and visual 2D shapes has been reported (e.g., harsh timbres with sharp angular shapes and soft timbres with curved shapes) [1, 34]. By using this kind of correspondence, we assume it is possible to present feedback that is easy for learners to understand the relationship between timbres and feedback, making it easier to adjust the playing timbre closer to the goal. In addition, we assumed that using crossmodal correspondences makes it easier for learners to recall the feedback from the timbre being played, even after the system's feedback is gone, and is less likely to cause performance degradation.

To leverage timbre–shape correspondences, we must consider an *intelligent* method to map the shape for any timbre. Because there may be individual differences in the correspondences between the timbres and shapes [6], the mapping must be personalized for each user. However, asking users to indicate the shape they feel corresponds to timbre for all timbres is impossible.

Therefore, we developed *TimToShape*, a system to generate and visualize 2D shapes corresponding to the current timbre in real-time based on the correspondences between some timbres and shapes answered by the user in advance (Fig. 1). TimToShape first acquires the latent space of timbre by unsupervised learning with variable autoencoder (VAE) [20] and then selects the optimal points from the latent space to estimate the user's timbre–shape correspondences. Users are asked to indicate the shapes they feel correspond to the timbres at those points, and TimToShape estimates the shape for any timbre in the latent space with linear interpolation of the answered shapes.

We evaluated our system using the violin as a case study because of the difficulty and importance of improving techniques related to timbre quality. Bowed string instruments, which produce sound by rubbing strings with a bow, have difficulty producing beautiful timbre. In addition, the timbre quality of violins is crucial for the overall evaluation of musical performance [11].

We conducted two user studies to verify the validity of TimToShape. The first study used crowdsourcing (n=75) to validate that the proposed system can generate shapes corresponding to any violin timbre presented. The shapes generated by our proposed system were more likely to be felt corresponding to timbres than the randomly generated shapes. The second study was conducted with six violin players to investigate the effects of the shape visual feedback generated by our system on violin practice. Violin players well received our system in terms of visual clarity, ease of understanding the relationship between the timbre and feedback, and ease of recalling the feedback even after it is no longer available, compared with the feedback system, which simply visualizes the position of the timbre in the latent space like a map.

In summary, the main contributions of this paper are as follows:

(1) Introduced the concept of visualizing timbres by 2D shapes based on timbre–shape correspondences of each user to support the practice of musical instruments.
(2) Developed TimToShape, which encodes the timbre to the latent space using the VAE and generates a shape corresponding to the timbre with linear interpolation of some shapes that the user answered as corresponding to some timbres in the latent space.
(3) Reported the evaluation results and reflections on the shape estimation method of TimToShape with online user studies and on using TimToShape for practicing musical instruments with in-person user studies.

2 RELATED WORK

This section reviews prior works on the computer-supported practice of musical instruments, especially the feedback methods analyzing the played sounds by learners. We also review crossmodal correspondences, especially audiovisual correspondences, because our idea of visualization is based on these kinds of human sensory attributes among the different modalities. In addition, we review the existing methods of sound visualization methods based on perceptual dimensions to situate our approach.

2.1 Computer-supported Practice of Musical Instruments

Several studies have proposed using technology to support the individual practice of playing musical instruments. Particularly, numerous systems exist that support the acquisition of skills related to pitch and rhythm [3, 5, 26, 31]. This is because they can be easily evaluated quantitatively since they are associated with some physical counterparts; pitch with fundamental frequency (the reciprocal of the period of a harmonic sound) and rhythm with note onset (the time when a note is played).

In contrast, timbre is high-dimensional and sensuous concept compared to pitch and rhythm. Although timbre can be associated physically with the spectral envelope of sound, the spectral envelope contains so many features that it is difficult to evaluate timbre quantitatively. Despite the difficulty in evaluating timbre, various systems have been developed to support the acquisition of skills related to timbre. Examples of evaluating the parameters of performance movements rather than evaluating the timbre itself include those that directly sense and visualize the parameters of performance movements by motion tracking [25, 44], and those that indirectly estimate the parameters of performance movements from the sounds played by support vector machine (SVM) [24], hidden Markov Model [37], and long short-term memory (LSTM) [35, 36]. However, it is difficult to measure the sensitive parameters by these examples, such as the amount of force, with high precision, and they also lack the versatility to be applied to any kind of musical instrument.

Several studies have made it possible to evaluate the timbre itself by extracting several features from sound. For example, Picas et al. proposed a method to evaluate timbre using scores by selecting spectral features with descriptive power [39]. Giraldo et al. proposed a method to evaluate the timbre of a violin based on semantic axes, such as dark–light, using machine learning [14]. Knight et al. proposed a method to rate the quality of trumpet timbres using SVM [21]. These studies differ from our work in that they focus on building a model that can be evaluated similarly based on evaluations by skilled players for each instrument. There were no evaluations of the improvement in timbre that could be achieved by using these models as feedback. Moreover, these examples still have the same drawback of versatility because they require labeling by experts or careful feature selection for each instrument.

In contrast, Kimura et al. proposed a practice support system that analyzes and visualizes timbres, and can be easily applied to various instruments without the need for labeling by experts [18]. They used an unsupervised learning method using the VAE to obtain the latent space of timbres as a 2D space and they proposed the visual feedback system which visualizes the position of the played timbre in the obtained latent space like a map. Although our work is based on Kimura et al.'s VAE-based method for analyzing timbre, we aim to design intuitive visual feedback for timbre to facilitate performance adjustment and reduce dependence on feedback during the practice of musical instruments.

2.2 Crossmodal Correspondences of Sound and Vision

There is a natural and intuitive connection between seemingly unrelated pieces of information, such as associating sound with brightness [27] and shape with sound [22]. Such non-arbitrary associative relationships between different modalities are called "crossmodal correspondences" [43]. The Bouba-Kiki effect is a well-known example of crossmodal correspondences. For most people, the word "Bouba" is associated with a curved shape, and the word "Kiki" is associated with a sharp angular shape [38].

Various examples of crossmodal correspondences between sound and vision have been reported [8, 13, 29, 29, 33, 34]. As an example of the correspondence between timbre and visual shapes, Parise et al. reported that sine waves were more likely to be associated with curved shapes, and square waves were more likely to be associated with sharp angular shapes [34]. Adeli et al. examined the relationship between visual shape and instrumental timbres [1]. Harsh timbres, such as triangles and cymbals, were associated with a sharp angular shape. Soft timbres, such as piano and marimba, were associated with a curved shape. Timbres with harshness and softness combined were associated with a mixture of two previous shapes. Gurman et al. replicated the Adeli et al. experiment using more timbres and shapes [16]. They found the same timbre–shape correspondences as Adeli et al. and more timbre–shape correspondences (hollowness of sound and shape, and distorted sound with fuzzy shape). These studies have shown that numerous people are consistent in the types of shapes they perceive as congruent with timbre. By leveraging these correspondences between timbres and shapes, our system aims to provide intuitive visual feedback regarding timbres. While the studies presented above have examined the relationship between discretely different timbres and discretely different shapes, we need to determine the corresponding shapes for continuously varying timbres. Therefore, our system estimates the correspondence of a shape to an arbitrary timbre based on the correspondence of the shape to several timbres.

Although the crossmodal correspondence is a characteristic shared by many people, there are individual differences due to cultural differences, experiences, and other factors [6], because crossmodal correspondences are said to be based on experiences acquired in daily life [10, 43]. Therefore, rather than visualizing a fixed correspondence between timbre and shape, TimToShape was designed to allow learners to customize the correspondences.

2.3 Sound Visualization Based on Perceptual Dimensions

Various studies have proposed the methods to visualize sound based on its perceptual dimensions. Most of their purpose is to enable users to quickly find the desired sound from a library of sounds. For example, Music Icons [23] mapped a flower-like icon to a single music file based on user perceptions of tempo, timbre, etc. ThumnailDJ [7] also presented a single thumbnail icon for a single music file based on the music genre and aggressiveness. While these are examples of visualizing impressions throughout the music, our goal is to provide real-time visualization of timbre, so we need to visualize timbre for each moment.

Several methods have also been proposed for a visualization based on perceptual dimensions of timbre that can be defined for the short momentary timbres. Giannakis proposed Sound Mosaics [12], which represent sharpness, compactness, and dissonance of sound using coarseness, granularity, and repetitiveness of textural patterns. Grill et al. [15] proposed a method for visualizing sound in the form of texture based on the impression of sound on axes such as ordered-chaotic, smooth-coarse, etc., mapping each to the regularity of elements, the jaggedness of element outline, etc. These examples are similar to our visualization method in some respects. However, they artificially fixed a single function that maps certain elements of sound to particular visual parameters. In contrast, we provided a flexible mapping function according to the sensory characteristics of each user.

3 THEORY & IMPLEMENTATION

The easiest way to estimate the correspondence between timbres and shapes for each user is to ask the user to indicate the shape that they feel corresponds to timbre for all timbres. However, the timbre of a musical instrument (even if it is the same instrument and played at the same pitch) changes continuously in various ways, depending on how it is played. It is theoretically impossible to ask the user to answer the shapes that they feel correspond to "all" timbres. Therefore, in this paper, this would be the problem of estimating the shape the user may perceive as corresponding to any given timbres based on the user's correspondence of shape to some "representative" timbres.

3.1 Problem Formulation

Let $T \subseteq \mathbb{R}^N$ be the feature space of timbres, and $\boldsymbol{f}_{timbre}$ be the map from raw timbre data to the feature vector of timbre. Similarly, let $S \subseteq \mathbb{R}^M$ be the feature space of shapes, and $\boldsymbol{f}_{shape}$ be the map from raw shape data to the feature vector of shape.

Furthermore, let $\boldsymbol{t}_i \coloneqq \boldsymbol{f}_{timbre}(\mathrm{t_i})$ $(\in T)$ be the feature vector of raw timbre data $\mathrm{t_i}$, let $\boldsymbol{s}_i \coloneqq \boldsymbol{f}_{shape}(\mathrm{s_i})$ $(\in S)$ be the feature vector of shape $\mathrm{s_i}$, and let $\boldsymbol{f}_{user} \colon T \longrightarrow S$ be the user's mapping function from the feature space of timbres to that of shapes.

Then, the problem of "estimating the corresponding shape for any timbre from the shapes answered to some representative timbres" can be formulated as follows:

> Given some observed timbre–shape pairs $\{(\boldsymbol{t}_i,\ \boldsymbol{s}_i\ (= \boldsymbol{f}_{user}(\boldsymbol{t}_i))) \mid \boldsymbol{t}_i \in T, \boldsymbol{s}_i \in S\}$, estimate $\boldsymbol{s}^* = \boldsymbol{f}_{user}(\boldsymbol{t}^*)$ for some timbre $\boldsymbol{t}^* \in T$.

Assuming that the set of observation points $\{\boldsymbol{t}_i\}$ is an irregular grid (scattered data) and that there are no (or few) errors in the observation of $\{\boldsymbol{s}_i (= \boldsymbol{f}_{user}(\boldsymbol{t}_i))\}$, this problem can be seen as an irregular-grid multivariate interpolation (or extrapolation) problem.

Sections 3.2 and 3.3 explain what we used for the feature space of timbres and shapes, respectively, and in Section 3.4, we explain how we solved this interpolation (or extrapolation) problem.

3.2 Feature Space of Timbres

To map the timbres to the feature space T, in this paper, as in Kimura et al. [18], VAE is used.

Fig. 2 shows the architecture of the VAE. The input for the VAE is a log-mel-spectrogram (calculated using $n\_fft = 2,048$, $hop\_length = 1,024$, $n\_mels = 128$, resulting in an overall size of 128×64) of the timbre frame. The timbre frame here means the segment of the timbre to which a single shape is mapped. In this paper, 66,560 samples ($\fallingdotseq$1.5 s) at a sampling rate of 44,100 Hz are used as one frame, and when using TimToShape in real-time, frames are updated every 1,024 sample points (frames are overlapped). This frame length was determined to be appropriate because it is sufficiently long to allow users to listen and associate shapes to the "representative" timbres but not so long that it may impair responsiveness in real-time use. Then, the input is encoded into the two-dimensional latent space (this dimension number follows Kimura et al. [18]), and the latent vector $\boldsymbol{z}$ is decoded into the output of the same size as the input (128×64). The VAE model was implemented using an open-source framework called Keras[1].

The VAE encodes the input as a stochastic distribution rather than as a single point. This is accomplished by the encoder separately outputting the mean ($\boldsymbol{\mu}$) and variance ($\boldsymbol{\sigma}$) of the distribution of the latent variable $\boldsymbol{z}$ separately, and sampling $\boldsymbol{z}$ which follows a normal distribution $\mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\sigma})$ by combining $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ with a random vector following a standard normal distribution (this technique is called "reparameterization trick" [20]). However, because the feature vector of one timbre should be constant in our system, we employed $\boldsymbol{\mu}$ of $\boldsymbol{z}$ as the feature vector of the timbre.

Note that because the VAE is trained unsupervised, this method of acquiring a timbre feature space can be applied regardless of the type of timbre as long as a dataset of the timbre frames is available. We only need to compute the log-mel-spectrogram of each timbre frame in the dataset and use it to train the VAE.

3.3 Feature Space of Shapes

To be used for interpolation of shapes, S (feature space of shapes) must satisfy the following condition: for any two points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ in S, when parameters move on the line connecting the two points, the corresponding shape also changes smoothly from $\mathrm{s_i}$ to $\mathrm{s_j}$.

In this paper, we used the "frequency domain" of shape as this feature space S referring to Wada et al. [45]. A shape can be smoothly morphed between any two shapes using their frequency vectors. See Appendix A.1 for a detailed explanation.

3.4 Linear Estimation of Shape

Now we explain how TimToShape estimates shape $\boldsymbol{s}^*$ for arbitrary timbre $\boldsymbol{t}^*$. As described in Section 3.1, this problem is an irregular-grid multivariate interpolation (or extrapolation) problem. Some methods that can solve this problem include nearest neighbor interpolation [4], linear interpolation (with a triangulated irregular network) [4, 41], natural neighbor interpolation [42], inverse distance weighting (IDW) [4], and Kriging (Gaussian process regression) [30].

Linear interpolation and extrapolation were used in this paper. We used this linear estimation method because the calculation is relatively simple, and the linear extrapolation method extrapolates observations based on the gradient on the boundary of the convex hull of the observation points and is considered suitable for this system.

The basic idea of linear interpolation/extrapolation can be expressed as follows where $\hat{\boldsymbol{t}}$ is an appropriate point in the $\boldsymbol{t}^*$ neighborhood and $\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\hat{\boldsymbol{t}}}$ is the estimated Jacobian of $\boldsymbol{f}_{user}$ at $\hat{\boldsymbol{t}}$. For a detailed explanation, see Appendix A.2.

$$\boldsymbol{s}^* = \boldsymbol{f}_{user}(\hat{\boldsymbol{t}}) + \boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\hat{\boldsymbol{t}}}(\boldsymbol{t}^* - \hat{\boldsymbol{t}}) \tag{1}$$

3.5 Method for Observation of User's Mapping Function

Next, we will explain the method for observation of $\boldsymbol{f}_{user}$, which includes how to select the observation points (representative timbres) $\{\boldsymbol{t}_i\}$ and how to observe $\boldsymbol{s}_i$ for each $\boldsymbol{t}_i$.

[1] https://keras.io

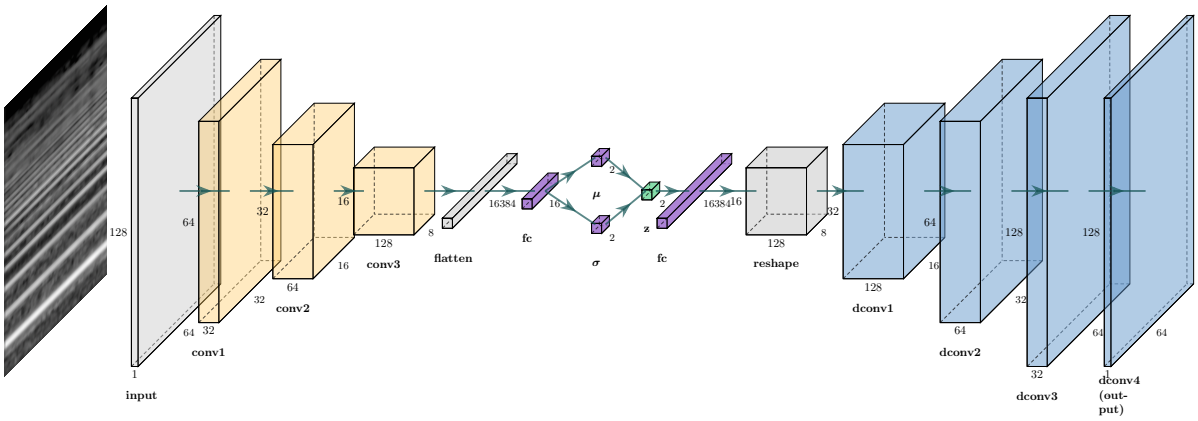

Figure 2: VAE architecture used in TimToShape. The input (128×64) is encoded into a 2-dimensional latent space via three convolution layers and output to the same size as the input via four deconvolution layers.

Before describing the detail of the method, there is an important point to note. Although we have discussed f_{user} as a deterministic function, we suppose that the *actual* human correspondence between timbres and shapes is not deterministic and has the following two properties:

Property 1. The relationship between the timbres and shapes is affected by the relative differences in the timbres. This means that the image of a shape for each timbre is expressed more clearly when the user compares timbres than when the user listens to them separately.

Property 2. Instead of one specific shape corresponding to one timbre, a group of shapes with similar impressions will correspond.

Therefore, f_{user} shall be considered here as a "favorable" one to use in the system among the entire set of actual correspondences of the user between timbres and shapes. We consider a correspondence to be "favorable" if it satisfies the following two conditions.

Condition 1. The variation in the shape mapped to the timbres is large.

Condition 2. The mapped shape changes smoothly for the entire T.

To observe f_{user} such that these conditions are satisfied, we devised the following steps.

(1) Prepare $\mathrm{T_{dataset}}$ as a dataset of timbre frames (raw waveform data of timbre frames) to be targeted by the system and prepare an empty set T_{rep}. For this dataset, we can use the dataset of timbre frames used to acquire the feature space of timbres (described in Section 3.2).

(2) Calculate the image $T_{dataset} \coloneqq \{ f_{timbre}(\mathrm{t}) \in T \mid \mathrm{t} \in \mathrm{T_{dataset}} \}$.

(3) From $T_{dataset}$, select a set of $N+1$ points (call this set $T_{farthest}$) such that all the points are affinely independent and the distance of the closest 2 points in $T_{farthest}$ is maximal. Then calculate the inverse image $\mathrm{T_{farthest}} \coloneqq \{ \mathrm{t} \in \mathrm{T_{dataset}} \mid f_{timbre}(\mathrm{t}) \in T_{farthest} \}$.
The timbres in $\mathrm{T_{farthest}}$ are the first $N+1$ "representative" timbres. These points are selected because the timbres at these points are more likely to be a characteristic combination of timbres in the dataset. According to *Property1.*, users are more likely to associate shapes that greatly differ, which may help f_{user} satisfy *Condition* 1.

(4) Ask the user to listen and compare all the timbres in $\mathrm{T_{farthest}}$, and for each timbre $\mathrm{t_i} \in \mathrm{T_{farthest}}$, ask the user to answer the shape $\mathrm{s_i}$ that they feel corresponds to $\mathrm{t_i}$. Then $N+1$ observations $\{ (t_i, s_i) \coloneqq (f_{timbre}(\mathrm{t_i}),\ f_{shape}(\mathrm{s_i})) \}$ are recorded. After that, add all the observation points $\{t_i\}$ to T_{rep}.

(5) Repeat the following steps: Theoretically, the following should be repeated indefinitely until the termination condition shown in step (5) (b) is met; however, in this paper, 30 iterations were set as the upper limit of iterations, considering the burden on the user.

(a) Select one $t_{farthest}\ s.t.\ t_{farthest} \notin T_{rep} \wedge t_{farthest} \in \underset{t \in T_{dataset}}{\arg\max} \{ \min_{t' \in T_{rep}} \|t - t'\| \}$.
Then find $\mathrm{t_{farthest}} \in \mathrm{T_{dataset}}\ s.t.\ f_{timbre}(\mathrm{t_{farthest}}) = t_{farthest}$.
This $\mathrm{t_{farthest}}$ is the next "representative" timbre. Therefore, the user is asked to listen to $\mathrm{t_{farthest}}$.
This point is selected with the assumption that the least information about f_{user} is obtained at this point.

(b) Display a shape $\mathrm{s^*_{farthest}}$, which is generated from $s^*_{farthest}$ estimated for $t_{farthest}$ based on the observations so far, and ask the user if they feel that the displayed shape $\mathrm{s^*_{farthest}}$ already corresponds to the timbre $\mathrm{t_{farthest}}$.
If the user says yes, $(t_{farthest}, s^*_{farthest})$ is recorded as another observation, $t_{farthest}$ is added to T_{rep}, and the remaining steps of this loop are skipped. If this occurs three times in a row, the entire observation process is completed because the observation can be considered sufficient.
If the user says no, go to step (5)(c).

(c) Ask the user to search a shape that they feel corresponds to $\mathrm{t_{farthest}}$ somewhere around $s^*_{farthest}$ according to the following Eq. (2). The J in Eq. (2) is the Jacobian estimated in Eq. (1), and the user moves the $t_{neighbor}$ around

$\boldsymbol{t}_{farthest}$. In actual use, N sliders are presented, and the user manipulates $\boldsymbol{t}_{neighbor}$ by manipulating each slider. If the user finds the shape, let $\boldsymbol{s}^{**}_{farthest}$ be the feature vector of that shape, then $(\boldsymbol{t}_{farthest}, \boldsymbol{s}^{**}_{farthest})$ is recorded as another observation, $\boldsymbol{t}_{farthest}$ is added to $\boldsymbol{T}_{rep}$, and the remaining steps of this loop are skipped. Otherwise, go to step (5)(d).

$$\boldsymbol{s}^{**} = \boldsymbol{s}^{*} + \boldsymbol{J}(\boldsymbol{t}_{neighbor} - \boldsymbol{t}_{farthest}) \tag{2}$$

The reason for preparing this step was to use *Property* 2. to observe $\boldsymbol{f}_{user}$ without increasing the dimension of the image of $T_{dataset}$ mapped to S by $\boldsymbol{f}_{user}$ as long as possible, which may help $\boldsymbol{f}_{user}$ satisfy *Condition* 2.

(d) Similar to step (4), ask the user to answer the arbitrary shape $\mathrm{s}'_{\mathrm{farthest}}$ that they feel corresponds to timbre. Then $(\boldsymbol{t}_{farthest}, \boldsymbol{s}'_{farthest})$ is recorded as another observation, and add $\boldsymbol{t}_{farthest}$ to $\boldsymbol{T}_{rep}$.

Note that this method of observing the user's mapping function is applicable regardless of the timbre type of the dataset. Since the method of acquiring the timbre feature space can also be applied to any dataset (as explained in Section 3.2), the entire process of constructing TimToShape is independent of the timbre type, which is one of the major advantages of this method.

The next section describes how the user answers the shapes in Steps (4) and (5).

3.6 Process for Answering Shape for Timbre

As explained in Section 3.3, the frequency vector of the shape is adopted as the feature vector of the shape; however, it is almost impossible for users to manipulate the frequency vector to create the desired shape. Therefore, we implemented a method based on that of Wada et al. [45] to generate shapes from semantic shape parameters to enable users to use them when answering shapes to timbres. The following six parameters were used to generate shape: *number of spikes* (0 to 30), *length of spikes* (0 to 0.5), *randomness* (0 to 1), *random seed* (0 to 10), *roundness of the base of spikes* (0 to 1), and *roundness of the tip of spikes* (0 to 1), referring to the study by Oyama et al. [32]. The details of how shapes are generated from these parameters are provided in Appendix A.3.

Using this method of generating shapes from semantic parameters, users can answer the shape they feel corresponds to the timbre in the following process:

(1) The user selects one shape from a set of 25 basic shapes we defined (*Shape Samples*, Fig. 3), which is created by varying the six semantic shape parameters described above, so that the user feels best to correspond to the timbre. We prepared this step because it is easier for users to compare several shapes and choose one from them than think of a shape from scratch.
(2) The user modifies the shape selected in the previous step with the six semantic shape parameters so that the user feels that the shape corresponds more to the timbre.

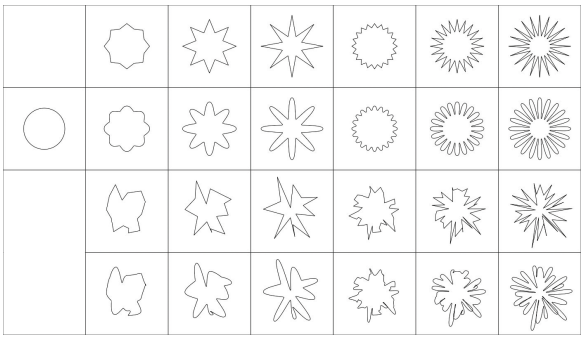

Figure 3: Shape Samples: A set of 25 fundamental shapes that can be created from the semantic shape parameters. The display order of the shapes is randomly rearranged each time.

4 INTERFACE FOR ANNOTATION AND VISUAL FEEDBACK

4.1 Interface for Timbre–shape Annotation

In Sections 3.5 and 3.6, we described the methodology to observe $\boldsymbol{f}_{user}$ and briefly explained the implementation. In this section, we explain how they were implemented as an interface and how to use them from the user's perspective. Henceforth, we will refer to the step where the user answers a shape to each representative timbres as "timbre–shape annotation."

From the user's perspective, the timbre–shape annotation can be divided into two main steps: one is to answer shapes for the first three timbres, and the other is to answer shapes for the fourth and subsequent timbres.

The first step, in which the user answers the shapes for the first three timbres, corresponds to step (4) in Section 3.5. The reason for the number of timbres being 3 is that, as explained in Section 3.2, the number of dimensions of the latent space of timbres (N) is 2 in our implementation. In this step, the user answers a shape to each of the three timbres, following the process described in Section 3.6. The interface used here is shown in Fig. 4 (a), where the three timbres are shown simultaneously, and the *Shape Samples* are shown on the right side. After the user selects a shape from *Shape Samples*, the sliders appear, allowing the user to edit shapes using the six semantic shape parameters.

The second step, in which the user answers the shapes for the fourth and subsequent timbres, corresponds to step (5) in Section 3.5. This step consists of the following three phases.

Phase 1. Verify whether the estimated shape is acceptable (Fig. 4 (b)); this corresponds to Step (5) (b) in Section 3.5. Select whether a modification is necessary by clicking on a button. If the shape requires modification, proceed to *Phase* 2. Otherwise, proceed to the next timbre. The timbre–shape annotation is complete if this occurs three times in a row.
Phase 2. Verify the shapes around the estimated shape (Fig. 4 (c)); this corresponds to Step (5) (c) in Section 3.5. Two sliders were shown to allow searching for shapes in the vicinity

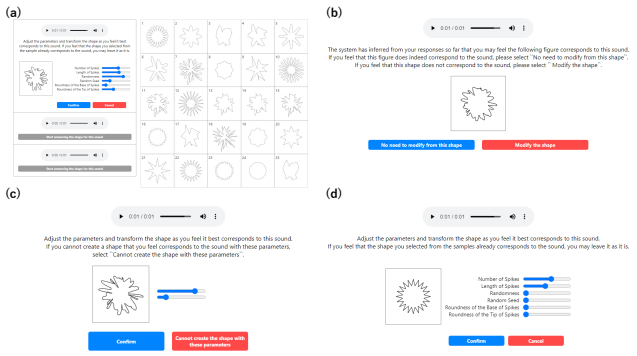

Figure 4: Interface implemented for timbre–shape annotation. (a) Interface to answer shapes to the first three timbres. (b) Interface to determine whether the estimated shape needs modification. (c) Interface to search the shape in the neighbor of the original estimation. (d) Interface to answer shape to fourth and subsequent timbres (*Shape Samples* is omitted). See Section 4.1 for detail.

of the estimated shape. If the user cannot find a shape that they feel corresponds to the timbre, proceed to *Phase* 3.

Phase 3. Answer the shape that the user feels corresponds to the timbre (Fig. 4 (d)), following the process described in Section 3.6, which corresponds to Step (5) (d) in Section 3.5.

4.2 Real-time Shape Feedback Corresponding to Timbre

TimToShape estimates and displays shapes in real-time, corresponding to the timbre input to the microphone. To achieve this, the following calculations must be performed with low latency:

Calculation 1. Computation of the log-mel-spectrogram of the microphone input

Calculation 2. Encoding into the latent space by the VAE

Calculation 3. Estimation of frequency vector of shape using linear interpolation (or extrapolation)

Calculation 4. Inverse Fourier transform of the frequency vector and drawing of the shape

We implemented these using JavaScript to run on a browser. Several implementation efforts have meen made to speed up the above calculations (see Appendix A.4 for details). The final computation time was: 5 ~ 6 ms for calculation 1; 20 ~ 40 ms for calculation 2; 1 ~ 2 ms for calculation 3; and 1 ~ 2 ms for calculation 4. Overall, it took less than 60 ms from the time the last sample point of a frame was input to the microphone until it was reflected in the drawing. Furthermore, using asynchronous processing, we achieved a frame rate of approximately 40 fps (PC: Macbook Pro, M1, 2020 with version 12.4 of macOS Monterey; Browser: Google Chrome of version 106.0.5249.119).

5 STUDY 1: CONGRUENCY OF TIMBRE AND GENERATED SHAPES

This study aims to verify that TimToShape can generate a shape that users feel corresponds to any arbitrary timbre based on each user's input of timbre–shape correspondences. We conducted an online user study to determine whether the shapes generated by TimToShape were more likely to be perceived as congruent by the participants than randomly generated shapes.

This study did not use TimToShape as a real-time feedback system (the shape changes dynamically) but as a system that generates one static shape for one frame ($\fallingdotseq 1.5s$) of timbre. This is because if the shape changes dynamically, it is impossible to measure the congruency of the timbre and shape in each frame.

5.1 Dataset

In this paper, the timbres we focused on were of the "open A-string" of violin, which corresponds to the simplest and most basic playing action of the violin. The timbre dataset used in this study was a smartphone recording (from iPhone XR) of the open A-string of a violin by one of the authors. He is an advanced violinist who has been playing the violin for more than 19 years. The dataset contains timbres of the violin's "open A-string" played with various force levels, bow positions, and bow angles, including timbres imitating those produced by a beginner as well as an advanced player.

The recordings were made in a quiet room using a standard smartphone recording application (Voice Memos[2]). A smartphone was used for the recording because we aimed to make our system operate without requiring special equipment or applications. The recording time was approximately 20 min. The silent time was trimmed, frames (one frame containing 66,560 sample points) were cut out with shifting every 500 sample points (frames overlapped) and then a dataset containing 90,159 frames was created.

5.2 Training

Using this timbre dataset, unsupervised learning of the VAE (described in Section 3.2) was performed. The VAE training was performed on a Windows 10 PC with two GeForce GTX1080Ti GPUs. We used 80% of the dataset for training and 20% for validation. The optimizer was Adam [19] with a learning rate of 0.001, the batch size was 512, and there were at most 200 training epochs. The training was automatically interrupted by early stopping based on the validation loss.

The resulting latent space ($\boldsymbol{\mu}$ of $\boldsymbol{z}$) of timbres is shown in Fig. 5 (a). Although no labeling was performed on the timbre dataset, the latent space mapping for the timbres exhibited a rough trend, as depicted in the figure (note that the labeling of the areas in Fig. 5 (a) represents our impression of the timbres).

Based on the method described in Section 3.5, the observation points in this latent space are shown in Fig. 5 (b). The red dots represent the observation points adopted in Steps (3) and (4) in Section 3.5, while the yellow dots represent the observation points (representative timbres) adopted in the 30 loops in Step (5) in Section 3.5. The numeric labels next to the yellow dots indicate the

[2]https://apps.apple.com/us/app/voice-memos/id1069512134

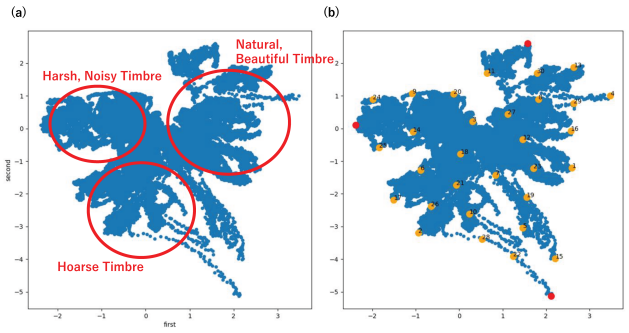

Figure 5: Latent space of timbres used in Study 1. All timbres in the dataset are mapped on the latent space as blue dots. (a) Overall tendency of timbres (the labeling for each area simply represents our impression of the timbres), (b) Observation points (representative timbres) selected in this latent space (see section 5.1)

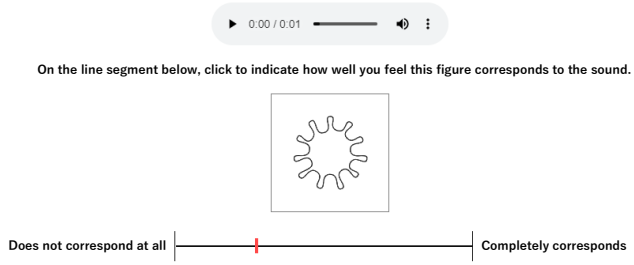

Figure 6: VAS used in Study 1. The length of the horizontal bar was made to be equal to 100mm according to [9]. The participants were asked to initialize the window scale to make the bar size with 100mm at the beginning of the study.

order of selection (all participants annotated the shapes for timbres in this order).

5.3 Participants

This study included 101 participants through a crowdsourcing platform[3]. Sixty-four participants were male, and 37 were female, with a mean age of 42.5 years and standard deviation of 7.43. All participants were required to participate in the study using a PC and wear earphones or headphones. Participants were paid 550 JPY for their participation.

5.4 Procedure

This study was divided into two sessions.

In the first session, the participants performed the timbre–shape annotation, following the flow described in Section 4.1. The interface is identical to that shown in Fig. 4. The timbres presented in this step correspond to the red and yellow dots in Fig. 5 (b) (as mentioned in Section 5.1).

In the second session, 30 timbres were randomly selected from the dataset, and there were two trials for each, resulting in a total of 60 trials. The trials were randomized. In each trial, either a shape generated by TimToShape or a randomly generated shape was presented along with the timbre (both shape generation methods were used once per timbre). Here, "randomly generated shapes" were generated by randomly setting the shape semantic parameters described in Section 3.6. For each presented timbre–shape pair, participants were asked to answer how well they felt the shape corresponded to the timbres by visual analog scale (VAS) (the left side was "Does not correspond at all" while the right side was "Completely corresponds") [9] (Fig. 6).

5.5 Results

Based on the time taken for the entire study, data from those who took an excessively short (less than 10 min) or long (more than 36 min) time were eliminated as inappropriate effort. These time criteria were based on the $\pm 1\sigma$ interval from the mean of the total time required for all participants. As a result, data from 75 of 101 participants were used in the analysis.

For each participant, the "congruence score" is calculated for each of the "TimToShape" and "Random" methods by the following process:

(1) For each trial, congruency was scored from 0 (does not correspond at all) to 100 (completely corresponds), depending on the position of the participant selected in the VAS.
(2) The average score of 30 trials for each method was used as the congruence score for that participant.

The distribution of the congruency scores for all 75 participants is shown in Fig. 7. The median congruence score of "TimToShape" was 66.1 out of 100 while the median score of "Random" was 43.1. A normality test (Shapiro-Wilk test) conducted on the distribution of scores for each method showed that the distribution of "Random" can be regarded as normal ($p = .15 > .05$) but that of "TimToShape" was not normal ($p = .03 < .05$). Therefore, as a paired nonparametric data, a Wilcoxon signed rank test was performed. The results indicated that "TimToShape" had a significantly higher congruence score than "Random" ($p = .000 < .001,\ r = 0.86$).

These results indicate that TimToShape can generate shapes that are felt to be correspond more to any timbre (in the dataset) than randomly generated shapes.

6 STUDY2: USE AS A FEEDBACK SYSTEM FOR VIOLIN PRACTICE

This study aims to determine the advantages, disadvantages, and other possible impacts of using TimToShape as a real-time feedback system for practicing musical instruments. Because the quantitative analysis and evaluation of timbre are difficult, we mainly analyzed the comments provided by the participants who used our proposed system during violin practice. We asked six violin players from various backgrounds to practice the violin using two systems: TimToShape and a system that visualizes the position of the timbre in the latent space as a 2D map. Without revealing the violin players

[3]https://www.lancers.jp

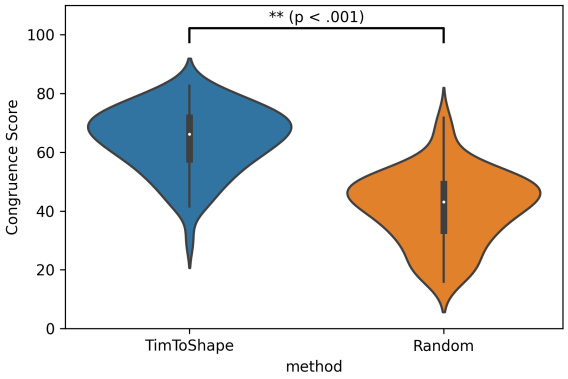

Figure 7: Comparison between the congruence scores obtained by the two methods: TimtoShape and Random (n=75).

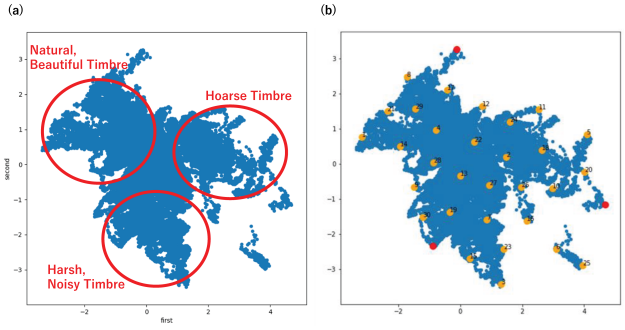

Figure 8: Latent space of timbres used in the Study 2. All timbres in the dataset are mapped on the latent space as blue dots. (a) Overall tendency of timbres (the labeling for each area simply represents our impression of the timbres), (b) Observation points (representative timbres) selected in this latent space (see Section 6.1)

which of the two we proposed, we asked them to practice the violin using each and conducted questionnaires and interviews. This study was approved by the ethics review board of our institution.

6.1 Dataset & Training

This study used a dataset different from that used in Study 1. This was because this study was originally planned to be conducted using a smartphone; however, because of the large computational load, TimToShape could not run in real-time on a smartphone. Therefore, it was modified to be conducted using a computer (Macbook Pro, M1, 2020 with version 12.4 of macOS Monterey). The recording method was almost the same as that described in Section 5.1, with the "open A-string" played with various techniques, but recorded with the standard recording application (Voice Memos[4]) of the computer. The total recording time was approximately 22 min, and frames were cut out as described in Section 5.1, creating a dataset containing 109,447 frames.

The VAE training was conducted in exactly the same setting that described in Section 5.2. The latent space resulting from the training is shown in Fig. 8 (a) (note that the labeling of the areas depicted in Fig. 8 (a) simply represents our impression of the timbres). Because the VAE was trained on a different dataset, the distribution of timbres was arranged differently than that shown in Fig. 5 (a). The observation points (representative timbres) selected from this latent space according to the method described in Section 3.5 are shown in Fig. 8 (b) (the red dots represent the observation points adopted in steps (3) and (4) in Section 3.5, while the yellow dots represent the observation points adopted in the 30 loops in step (5) in Section 3.5).

6.2 Practice Goal

In a user study of TimToShape, setting a goal for practice is an issue to consider. One possible approach would be to randomly select several points in the timbre latent space and use them as the goal timbres. However, this could lead to the possibility of having the participants practice reproducing "bad" timbres. To prevent this possibility, we asked a professional violinist to record a sample of an "open A-string" performance to be used as the goal timbre. The sample was recorded using the same instrument, microphone, and similar recording environment, as when the timbres included in the dataset were recorded. We requested the professional violinist to produce the best timbre that he felt could play on this violin.

We set the goal timbre to approximately 20 s sound of the 20-min recording, during which the professional stated that he had the best-sounding performance. Here, because the goal timbre was not perfectly constant for 20 s, there was variance in the position of the latent space. However, when using TimToShape in practice, users try to approach one shape that corresponds to a specific point in the latent space; therefore, we set the average position of the goal timbre in the latent space over 20 s as the goal point expediently. This practice goal was identical for every participant throughout the entire process of this user study.

6.3 Feedback Systems

We prepared two types of feedback systems for the user study. Both use the same VAE, but with different visualization methods.

The first is our system, TimToShape. When used in practice, a shape corresponding to the goal timbre (denoted "goal shape") and shape corresponding to the timbre currently being played (denoted "current shape") are displayed side by side, as shown in Fig. 9 (a). To make the differences in shapes easier to understand, the goal shape was shown as a translucent guide in the background of the current shape.

The other is a system that visualizes the location of the timbre in the timbre latent space like a map (we call this "Latent Space Visualization"). As shown in Fig. 9 (b), the point of the goal timbre and the point of the timbre currently being played are displayed on the map. This system is based on SonoSpace by Kimura et al. [18], but there are some differences. While SonoSpace does not update

[4]https://apps.apple.com/us/app/voice-memos/id1069512134

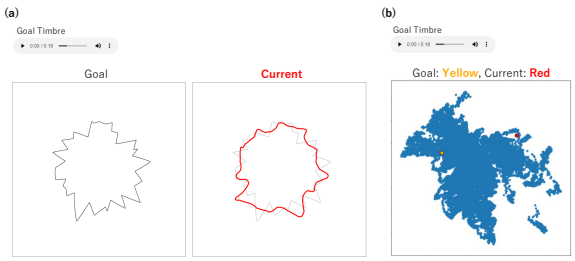

Figure 9: Feedback systems used in Study 2. (a) TimToShape. The shape corresponding to the goal timbre is displayed on the left side, while that corresponding to the played timbre is displayed on the right side. (b) Latent Space Visualization. The yellow dot represents the position of the goal timbre, while the red dot represents the position of the played timbre.

the points corresponding to timbres in the latent space in real-time, our implementation does. SonoSpace employs some additional user interfaces, but our implementation does not include such interfaces.

6.4 Participants

In this user study, six violin players (P-1 to P-6) were recruited. They ranged from beginners to advanced players. They were paid 3,000 JPY as an honorary.

Their experiences in playing their respective instruments were as follows:

P-1: A 55-year-old female and a beginner in the violin. She has some knowledge of playing techniques because her daughter is a violin student, and she has listened to her teacher's instructions with her. Aside from the violin, she had studied the piano for ten years.

P-2: A 57-year-old female and a beginner in the violin. She had previously taken violin lessons at school for one year. Aside from the violin, she had over 50 years of experience playing the piano and had used to be a piano teacher.

P-3: A 24-year-old male with little experience in the violin but an intermediate viola player. He has been learning how to play the viola for approximately nine years and continues to play in an orchestra.

P-4: A 23-year-old male and intermediate-to-advanced violin player. He had taken violin lessons for approximately ten years in the past, and he has been playing violin for approximately 18 years.

P-5: A 20-year-old male and advanced violin player. He has been taking violin lessons for 15 years and continues to play in an orchestra. He also has 16 years of experience playing the piano.

P-6: A 24-year-old female and advanced violin player. She has been taking violin lessons for 20 years and continues to play.

6.5 Procedure

This study was conducted in a quiet environment similar to the one in which the dataset was recorded. The violin and microphone used were identical to those used to create the dataset. The two feedback systems used in the study and the procedure were explained before the user study started. Written informed consent was obtained from all participants.

The procedure of this study consisted of eight steps.

(1) Timbre–Shape annotation: Timbre–Shape annotation was performed according to the flow described in Section 4.1 to estimate the timbre–shape correspondences of each participant.
(2) Pre-Test: The participants listened to the goal timbre and then, conducted a 1-min trial in which they tried to reproduce the goal timbre.
(3) Practice1: Ten minutes of practice using the first feedback system. To balance the effects of the order, "TimToShape" was performed first for P-1, P-3, and P-5, and "Latent Space Visualization" was performed first for P-2, P-4, and P-6.
(4) Acquisition Test1: Immediately after Practice1, a 1-min trial was conducted to reproduce the goal timbre using the feedback system.
(5) Retention Test1: After a 10-min break following Acquisition Test1, a 1-min trial to reproduce the goal timbre was conducted without the feedback system.
(6) Practice2: The same practice as that conducted in Practice1, but with the other feedback system.
(7) Acquisition Test2: The same trial as that conducted in Acquisition Test1, but with the other feedback system.
(8) Retention Test2: After a 10-min break following Acquisition Test2, a 1-min trial was conducted to reproduce the goal timbre without the feedback system.

The entire study took approximately 80 min.

6.6 Measurements

6.6.1 Objective Measures. We set the mean distance to the goal point in the latent space to be an objective measure of the performance in the 1-min trials in Pre-Test, Acquisition Test1, Retention Test1, Acquisition Test2, and Retention Test2. We will refer to this value as the "Distance Error Score" (the lower the score is, the better the performance is).

We adopted this measure mainly for two reasons. The first reason is that the distance on the latent space is an index that can objectively evaluate the physical closeness of timbre to some extent because the latent space can be considered as the dimension reduction of the log-mel-spectrogram of the sound. The second reason is that since the latent space used in "TimToShape" and "Latent Space Visualization" is identical, we can compare the two methods through this measure.

6.6.2 Subjective Measures. In addition to objective measures, semi-structured interviews and questionnaires were conducted to clarify the subjective assessments during practice.

- At the end of Acquisition Test1 and Acquisition Test2, interviews were conducted based on questions about the advantages and disadvantages they felt about the feedback system, when they felt the feedback got closer to the goal, and how to achieve that. In addition, using a questionnaire, the participants were asked to respond on a 7-point Likert scale, where one indicated high disagreement and seven indicated

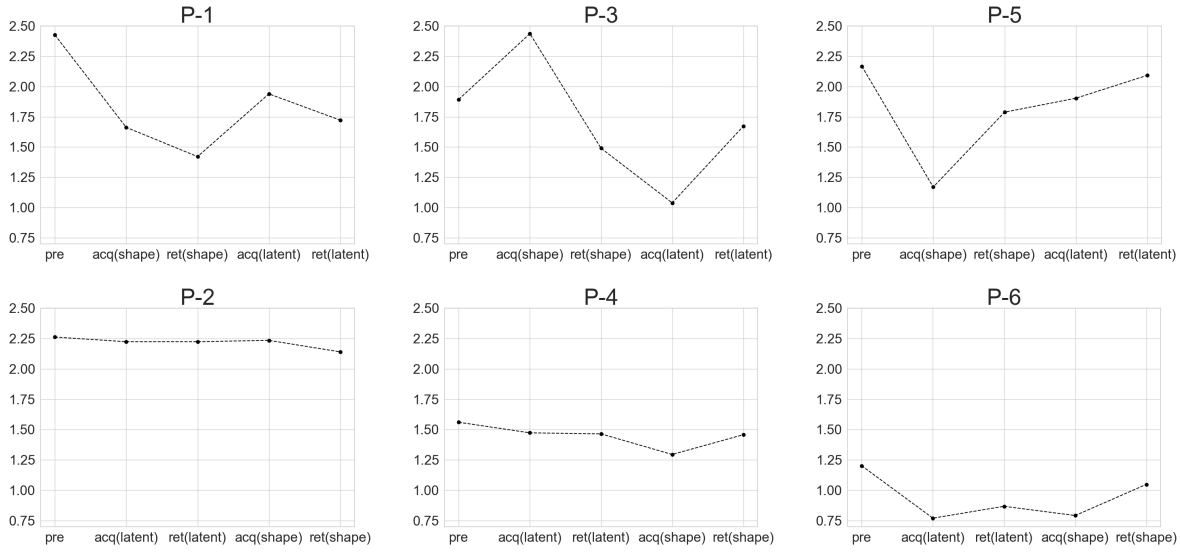

Figure 10: Transitions in Distance Error Scores for each participant (the lower the score is, the better the performance is). The label "pre" means "Pre-Test", "acq" means "Acquisition Test", and "ret" means "Retention Test". Moreover, "shape" means "TimToShape" and "latent" means "Latent Space Visualization".

high agreement, to the question about the interpretability of the feedback, that is, *"Did you understand the relationship between the sound you played and the feedback displayed?"*. Regarding the engagement of the visualization system, they were asked to respond to six questions on VisEngage [17]: "Challenge," "Discovery," and "Interest" (2 questions for each).

- In addition, after the Acquisition Test2, the participants were interviewed about which of the two feedback systems they would prefer to use in their actual practice and why.
- At the end of Retention Test1 and Retention Test2, they were interviewed based on questions about the differences compared to when the feedback was displayed.

6.7 Results and Discussion

6.7.1 Objective Measures. Fig. 10 shows the transitions in the "Distance Error Scores" across the five tests for each participant. The three upper graphs refer to the participants who worked on TimToShape first (P-1, P-3 and P-5), and the three lower graphs refer to the participants who worked on Latent Space Visualization first (P-2, P-4 and P-6).

First, by analyzing the groups that performed TimToShape first, we found that P-1 and P-5 performed better in both the Acquisition Test and the Retention Test with TimToShape, and P-3 performed better in the Acquisition Test with Latent Space Visualization, but performed better with TimToShape in the Retention Test.

Next, by analyzing the group that performed Latent Space Visualization first, we found a slight change in scores over the five tests, but it appeared that the person's ability in the Pre-Test remained dominant until the end.

Looking only at the scores, three (P-1, P-4, and P-5) performed better in the Acquisition Test and five (P-1, P-2, P-3, P-4, and P-5) performed better in the Retention Test with TimToShape than with Latent Space Visualization. Although some score superiority was shown by TimToShape, especially in terms of retention, it is difficult to demonstrate the clear superiority of TimToShape using this objective measure because of the large overall individual differences.

6.7.2 Subjective Measures. The distributions of responses with respect to the interpretability and VisEngage questions are shown in Fig. 11. For the VisEngage questions, the average of the responses to the two questions for each category was considered as one person's score. Although the statistical analysis was not performed due to a small number of participants, TimToShape tends to be higher in terms of "interpretability". Five participants evaluated TimToShape with the same or higher score than Latent Space Visualization. For the question that asked which method TimToShape or Latent Space Visualization they would use in practice, five participants (P-1, P-2, P-3, P-4, and P-6) answered TimToShape. These results suggest that TimToShape has higher interpretability and is preferred by most people over Latent Space Visualization.

From this point on, we will analyze the comments in the interviews to determine why the different methods of visualizing the latent space yielded differing impressions.

First, we introduce some comments on the advantages of TimToShape over Latent Shape Visualization.

"Shape was easy to understand visually and intuitively with a quick glance." (P-2)

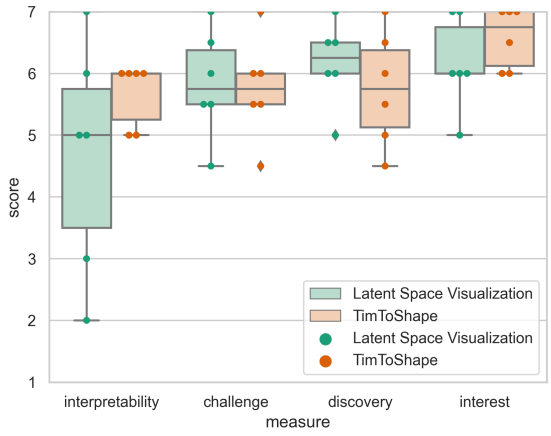

Figure 11: Comparison of scores (7-point Likert Scale) for subjective measures (N=6).

Both beginners (P-1, P-2) mentioned that they were so absorbed in their hands, arms and the movement of the bow that it was difficult for them to keep an eye on the feedback screen. In this context, P-2 stated that she could not follow the moving dot with her eyes (in Latent Space Visualization) but that the changing shape was highly comprehensible at a quick glance (in TimToShape).

"In the map feedback (Latent Space Visualization), it was difficult to understand what the vertical and horizontal axes represented, and it was difficult to move the points as I wanted, but in shape feedback (TimToShape), the transformation of the shape fits my intuition and it was easy to get the hang of getting closer to the goal." (P-6, similar comments from P-3 and P-4)

Several participants indicated that they had difficulty in grasping the meaning of the latent space axes. These comments correspond to the result that "interpretability" was rated higher for TimToShape, as mentioned above. Interestingly, although the same latent space was used in TimToShape, some participants found it easier to understand the relationship between timbre and feedback simply by visualizing it through a sensory-based shape.

Because there was only one goal timbre in this user study, it is not clear whether the results of the objective measures presented in Section 6.7.1 would be similar if the goal timbre were changed. However, the fact that the relationship between timbre and visual feedback was assessed to be easier to understand seems to support the effectiveness of TimToShape in aiming for other timbres.

"When practicing with map (Latent Space Visualization), I did not feel like I was trying to bring the "sound" sounds closer to the goal sound because I was concentrating only on the movement of the dot, and I think I was not listening to my own sound very well during the practice. But practicing with shape (TimToShape), I do not know why, I felt that this problem was lessened and that I could pay more attention to the sound." (P-4 and similar comments from P-5)

We considered this to be related to the previous comment. That is, we consider that the difficulty in interpreting the feedback caused many participants to focus only on the behavior of the feedback (except for beginners). Furthermore, we suspect that the reason this problem was reduced in TimToShape was not only because the relationship between the timbre and feedback was easier to grasp, but also because the timbres were visualized based on crossmodal correspondence, which subconsciously directed the user's attention to the timbre even while looking at the shape.

"When practicing with map (Latent Space Visualization) the dot was usually not close to the goal at all, even when I tried various attempts, which was demotivating. But when practicing with shape (TimToShape), I was able to practice without getting bored because I felt a sense of accomplishment when I got close to the target at some time." (P-1)

This is another advantage of TimToShape. Because VAE does not consider human perception during training, the distance in the latent space does not necessarily correspond to the distance in human perception. Thus, a situation can arise in which two points in the latent space are extremely distant, even though they sound close. This leads to a loss of motivation, particularly for beginners. However, this problem is less likely to occur with TimToShape because the shapes are created based on the user's perception, and even if the two sounds are far apart in the latent space, if the user feels they are similar, the shapes will be displayed close to each other.

"In the trial without feedback (Retention Test), although the shape was not displayed, I played while imagining the changing shape in real-time. However, (in Latent Space Visualization,) it was difficult to imagine the position on the map in real-time." (P-5 and similar comments from P-2)

We consider that this may also be a strength of the crossmodal correspondence. By visualizing the shapes that may be recalled by each user for timbres,it becomes easier to recall feedback even after it is lost.

On the other hand, the following comments were received regarding the advantages of Latent Space Visualization over TimToShape.

"Map visualization (Latent Space Visualization) seemed to have better resolution regarding goal and current errors." (P-4)

"Map feedback (Latent Space Visualization) seemed to analyze more elements of timbre." (P-5 and similar comments from P-3)

Both comments suggest that by visualizing the timbre as a shape, the amount of information is reduced compared with when the latent space of the timbre is visualized as it is. This is not necessarily a bad thing when used for feedback, in that it reduces the cognitive load of the user; however, further discussion is needed regarding the amount of information to be presented.

In addition, the following comments were received as common drawbacks to both.

"Technical advice is not presented, so sometimes it is not clear how to improve." (P-2 and similar comments from P-1)

"Both systems require me to do a variety of trial and error to find the moment when the feedback approaches the goal by myself. If I could not find it, I would be at a loss." (P-3)

This is a major challenge, particularly for beginners who use this system. TimToShape only visualizes timbres by shape; although this makes it easier to understand the relationship between timbres and feedback, it does not provide technical advice. Because technical advice depends on the type of instrument, attempting to solve this problem makes it impossible to build a method applicable to any instrument, which is the basic idea of TimToShape. Therefore, for beginners, we recommend that TimToShape be used along with advice from an adept professional or a system that provides feedback on technical information (e.g., that developed in [36]).

Overall, our system was well-received by violin players in terms of visual clarity, interpretability (ease of understanding the relationship between the timbre and feedback), and ease of recall (even after the feedback was lost).

7 LIMITATION & FUTURE WORK

7.1 Number of Dimensions of Latent Space of Timbre

In this paper, the dimensionality of the timbre latent space was set to 2 with reference on Kimura et al. [18]. However, further discussion of the dimensionality of the latent space is required. If the dimensionality of the latent space is increased, more timbre features can be reflected, leading to a more accurate estimation of the correspondence between the timbres and shapes. However, the user must answer the shapes for timbres more times in timbre–shape annotation, which would increase the burden for users. Further user studies should be conducted to determine the advantages and disadvantages of increasing the dimensionality and to determine the optimal dimensionality.

7.2 Parameters Making Up Shapes

As discussed in Section 5.5, the median congruence score for TimToShape was 66.1 (out of 100). This leaves scope for further improvement.

There are two ways to address this issue. The first is to revisit the method of creating shapes, as described in Section 3.6. Currently, we provide 25 shape samples, and users select one shape from them and edit six semantic parameters from the selected shape; however, this process may be inadequate for creating arbitrary shapes that the user wants to create. We believe that constructing a process to create shapes more freely would improve the congruency scores.

The second is to consider parameters other than the shape outline. Because timbre–shape correspondences vary from person to person, some may feel a strong correspondence to factors other than shape outline (e.g., size, color, and texture pattern), which has been reported by research on crossmodal correspondences between sound and vision [13, 33, 34]. Our method is so expandable that it can include any element in the output if it can be linearly interpolated as a vector. We leverage this strength to reflect elements other than the outline of the shape in the output and provide shapes with high congruency to the timbre for a larger number of people.

In addition, by increasing the number of parameters that make up the shape, we can provide more information as feedback for the timbre, which will help remedy the problem of the shape's low amount of information, as mentioned in Section 6.7.

7.3 Termination Condition for Timbre–Shape Annotation

There is scope for further studies regarding the termination conditions for timbre–shape annotation. Currently, the termination condition is to state that there is no problem without modifying the estimated shape three consecutive times in step (5) (b) described in Section 3.5, or to loop step (5) thirty times.

However, only 21 out of 75 participants in Study 1 finished with the former termination condition, while the remaining participants annotated the shapes for timbres thirty-three times (including the first three timbres).

We believe that this is because the current system validates the termination only in step (5) (b), which may be too strict. We should devise a method to validate the termination based on the size of the shape modification in step (5) (c). If an appropriate threshold can be determined, the burden on the user of answering with many shapes can be reduced without decreasing the Congruency Score.

7.4 Statistical Analysis of the Learning Effects

Study 2 mainly focused on examining the subjective experience of using TimToShape for feedback in practice. Therefore, violinists from various backgrounds were recruited and asked to use both TimToShape and Latent Space Visualization systems. The questionnaire and interviews revealed the subjective advantages and disadvantages of both systems.

Although the study indicated the visual clarity, interpretability and ease of recall of the proposed system, the objective learning effects of the proposed system have not been statistically clarified. To further clarify this, it would be necessary to conduct an experiment in which a larger number of participants are divided into the following three groups with the same original violin ability and compare the learning effects over a longer period: the group practicing without visual feedback, group practicing with TimToShape, and group practicing with Latent Space Visualization.

7.5 Application for Different Timbres and Environments

In this paper, TimToShape was tested only with the timbre of the "open A-string" of the same violin. However, our method can be easily applied to any timbre of any instrument if a corresponding dataset exists. In the future, we intend to create datasets of timbres from various instruments with various recording environments and build a system using these datasets that can be used with any instrument anywhere.

8 CONCLUSION

In this paper, we proposed TimToShape, a system that can visualize arbitrary timbres using 2D shapes aligned with a person's perception based on the fact that there is a crossmodal correspondence between timbres and shapes. TimToShape aims to provide feedback

in musical instrument practice, making it easy for humans to understand the relationship between timbre and feedback. TimToShape acquires a latent space of timbres through unsupervised learning by the VAE. By asking each user to annotate shapes for some points in the latent space, it can estimate a shape for any point in the latent space with linear interpolation. First, we conducted a user study via crowdsourcing (n=75) to verify the shape estimation method and found that the proposed method could generate shapes with significantly higher congruency to the timbre than randomly generated shapes. Next, we conducted a user study with six violin players to verify the effectiveness of this system when used as feedback for practicing a musical instrument and found that compared to simply visualizing the position of the timbre in the latent space (like a map), visualization by TimToShape was more comprehensible at a glance, was easier to understand the relationship between the timbre and feedback, and was easier to recall the feedback even after the feedback was lost. In the future, we will consider including elements other than shape outlines in the visualization to further improve congruency, and creating datasets of other timbres to improve the generalization performance for various instruments and environments. We hope that this paper will inspire synergy between the HCI and AI domains by providing insights that reflect human perceptual characteristics in an interface and interaction that applies machine learning.

ACKNOWLEDGMENTS

This work was supported by JST, PRESTO Grant Number JPMJPR18JA, Japan.

REFERENCES

[1] Mohammad Adeli, Jean Rouat, and StéAphane Molotchnikoff. 2014. Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in Human Neuroscience* 8, MAY (may 2014), 352. https://doi.org/10.3389/fnhum.2014.00352

[2] Kota Arai, Mone Konno, Yutaro Hirao, Shigeo Yoshida, and Takuji Narumi. 2021. Effect of visual feedback on understanding timbre with shapes based on crossmodal correspondences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. 1–3.

[3] Michael William Buck. 2008. *The efficacy of SmartMusic® assessment as a teaching and learning tool.* The University of Southern Mississippi.

[4] Peter A Burrough, Rachael A McDonnell, and Christopher D Lloyd. 1998. *Principles of geographical information systems.* Oxford university press.

[5] Estefanıa Cano, Christian Dittmar, and Sascha Grollmisch. 2011. Songs2See: learn to play by playing. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011).* Miami, 2231–2240. Retrieved October 3, 2022 from https://www.songs2see.com/

[6] Yi-Chuan Chen, Pi-Chun Huang, Andy Woods, and Charles Spence. 2016. When "Bouba" equals "Kiki": Cultural commonalities and cultural differences in sound-shape correspondences. *Scientific reports* 6, 1 (2016), 1–9.

[7] Ya-Xi Chen and René Klüber. 2010. ThumbnailDJ: Visual Thumbnails of Music Content.. In *ISMIR*. 565–570.

[8] Rocco Chiou and Anina N Rich. 2012. Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception* 41, 3 (2012), 339–353.

[9] Nicola Crichton. 2001. Visual analogue scale (VAS). *J Clin Nurs* 10, 5 (2001), 706–6.

[10] Ophelia Deroy and Charles Spence. 2016. Crossmodal correspondences: four challenges. *Multisensory research* 29, 1-3 (2016), 29–48.

[11] Ivan Galamian and Sally Thomas. 2013. *Principles of violin playing and teaching.* Courier Corporation.

[12] Kostas Giannakis. 2006. A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound* 11, 3 (2006), 297–307. https://doi.org/10.1017/S1355771806001531

[13] Kostas Giannakis and Matt Smith. 2001. Imaging soundscapes: Identifying cognitive associations between auditory and visual dimensions. *Musical Imagery* (2001), 161–179.

[14] Sergio Giraldo, George Waddell, Ignasi Nou, Ariadna Ortega, Oscar Mayor, Alfonso Perez, Aaron Williamon, and Rafael Ramirez. 2019. Automatic assessment of tone quality in violin music performance. *Frontiers in Psychology* 10 (2019), 334.

[15] Thomas Grill and Arthur Flexer. 2012. Visualization of perceptual qualities in textural sounds. In *ICMC*.

[16] Daniel Gurman, Colin R McCormick, and Raymond M Klein. 2021. Crossmodal correspondence between auditory timbre and visual shape. *Multisensory Research* 35, 3 (2021), 221–241.

[17] Ya-Hsin Hung and Paul Parsons. 2017. Assessing user engagement in information visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1708–1717.

[18] Naoki Kimura, Keisuke Shiro, Yota Takakura, Hiromi Nakamura, and Jun Rekimoto. 2020. SonoSpace. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery (ACM), New York, NY, USA, 367–374. https://doi.org/10.1145/3394171.3413542

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*.

[21] Trevor Knight, Finn Upham, and Ichiro Fujinaga. 2011. The potential for automatic assessment of trumpet tone quality.. In *ISMIR*. Citeseer, 573–578.

[22] Wolfgang Köhler. 1947. *Gestalt psychology: An introduction to new concepts in modern psychology.* Liveright Pub. Corp.

[23] Philipp Kolhoff, Jacqueline Preuß, and Jörn Loviscach. 2008. Content-based icons for music files. *Computers & Graphics* 32, 5 (2008), 550–560.

[24] Yin-Jyun Luo, Li Su, Yi-Hsuan Yang, and Tai-Shih Chi. 2015. Detection of Common Mistakes in Novice Violin Playing.. In *ISMIR*. 316–322.

[25] Esteban Maestre, Panagiotis Papiotis, Marco Marchini, Quim Llimona, Oscar Mayor, Alfonso Pérez, and Marcelo M Wanderley. 2017. Enriched multimodal representations of music performances: Online access and visualization. *Ieee Multimedia* 24, 1 (2017), 24–34.

[26] makemusic. 2006. SmartMusic | Music Learning Software for Educators & Students. Retrieved October 3, 2022 from https://www.smartmusic.com/

[27] Lawrence E Marks. 1987. On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of experimental psychology: Human Perception and Performance* 13, 3 (1987), 384.

[28] Dimitri Mavriplis. 2003. Revisiting the least-squares procedure for gradient reconstruction on unstructured meshes. In *16th AIAA computational fluid dynamics conference*. 3986.

[29] Robert D Melara and Thomas P O'Brien. 1987. Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General* 116, 4 (1987), 323.

[30] Margaret A Oliver and Richard Webster. 1990. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System* 4, 3 (1990), 313–332.

[31] Oy. 2020. Yousician | Learn Guitar, Piano, Ukulele With The Songs you Love. Retrieved October 3, 2022 from https://yousician.com/

[32] Tadasu Oyama, Hisao Miyano, and Hiroshi Yamada. 2003. Multidimensional scaling of computer-generated abstract forms. In *New developments in psychometrics*. Springer, 551–558.

[33] Cesare Parise and Charles Spence. 2008. Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters* 442, 3 (2008), 257–261.

[34] Cesare V Parise and Charles Spence. 2012. Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research* 220, 3-4 (aug 2012), 319–333. https://doi.org/10.1007/s00221-012-3140-6

[35] Alfonso Perez-Carrillo. 2016. Statistical models for the indirect acquisition of violin bowing controls from audio analysis. In *Proceedings of Meetings on Acoustics 172ASA*, Vol. 29. Acoustical Society of America, 035003.

[36] Alfonso Perez-Carrillo. 2019. Violin Timbre Navigator: Real-Time Visual Feedback of Violin Bowing Based on Audio Analysis and Machine Learning. In *International Conference on Multimedia Modeling*. Springer, 182–193.

[37] A. Perez-Carrillo and M. M. Wanderley. 2015. Indirect Acquisition of Violin Instrumental Controls from Audio Signal with Hidden Markov Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 5 (2015), 932–940. https://doi.org/10.1109/TASLP.2015.2410140

[38] Vilayanur S Ramachandran and Edward M Hubbard. 2001. Synaesthesia–a window into perception, thought and language. *Journal of consciousness studies* 8, 12 (2001), 3–34.

[39] Oriol Romani Picas, Hector Parra Rodriguez, Dara Dabiri, Hiroshi Tokuda, Wataru Hariya, Koji Oishi, and Xavier Serra. 2015. A real-time system for measuring sound goodness in instrumental sounds. In *Audio Engineering Society Convention 138*. Audio Engineering Society.

[40] Makiko Sadakata, David Hoppe, Alex Brandmeyer, Renee Timmers, and Peter Desain. 2008. Real-Time Visual Feedback for Learning to Perform Short Rhythms with Expressive Variations in Timing and Loudness. *Journal of New Music Research* 37, 3 (2008), 207–220. https://doi.org/10.1080/09298210802322401

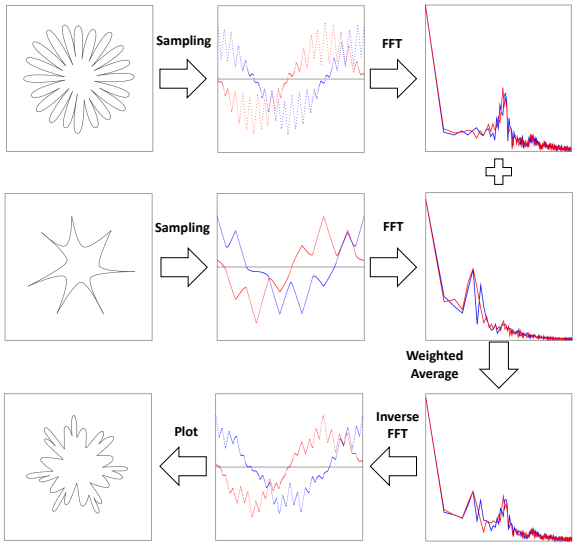

Figure 12: How shapes are Fourier transformed and how a shape is morphed between two shapes using the frequency domain (referring to Wada et al. [45]). In the figures in the center and right rows, the red plots are for the x-coordinate and the blue plots are for the y-coordinate.

[41] WE Schaap and R Van De Weygaert. 2000. Continuous fields and discrete samples: reconstruction through Delaunay tessellations. *arXiv preprint astro-ph/0011007* (2000).

[42] Robin Sibson. 1981. A brief description of natural neighbour interpolation. *Interpreting multivariate data* (1981).

[43] Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics* 73, 4 (2011), 971–995.

[44] Gualtiero Volpe, Ksenia Kolykhalova, Erica Volta, Simone Ghisio, George Waddell, Paolo Alborno, Stefano Piana, Corrado Canepa, and Rafael Ramirez-Melendez. 2017. A multimodal corpus for technology-enhanced learning of violin playing. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*. 1–5.

[45] Yuji Wada, Kazuya Matsubara, Akira Miyamae, and Kazuya Ishibashi. Japanese Patent 6725123 07 2020. Method, program and information processing device for displaying time-varying flavors as time-varying visual elements (in Japanese).

A IMPLEMENTATION DETAILS

A.1 Fourier Transform of Shape

The process of how the shape is converted to a frequency vector is as follows (referring to Wada et al. [45]). Note that this process can also be reversed to generate a shape from the frequency vector.

(1) The shape is sampled at 512 points to divide the perimeter equally.
(2) Fourier transform the x- and y-coordinates of sampled points separately and obtain a 512-dimensional complex vector of frequency dimensions for each.
(3) Separate the real and imaginary parts of the complex vectors and merge the four vectors into a 2048-dimensional real vector, which is the feature vector of the shape.

As shown in Fig. 12, a shape can be morphed smoothly between the two shapes using the weighted average of the frequency vectors of them.

A.2 Linear Interpolation and Extrapolation of Vector

Here we fully explain the linear interpolation/extrapolation method used in TimToShape. The symbols used in this section follow the definitions in Section 3.1.

When $\boldsymbol{t}^*$ is inside the convex hull of observed points $\{\boldsymbol{t}_i\}$ (hereafter $CH(\{\boldsymbol{t}_i\})$), linear interpolation is performed. The linear interpolation method used here is basically an extension of the Delaunay Tessellation Field Estimator (DTFE) [41] to multidimensional output. Specifically, let Del_{in} be the simplex that contains $\boldsymbol{t}^*$ when the interior of $CH(\{\boldsymbol{t}_i\})$ is partitioned by N-dimensional Delaunay tessellation, and let $\boldsymbol{t}_{a_0}, \ldots, \boldsymbol{t}_{a_N}$ be the vertices of Del_{in}, then $\boldsymbol{s}^*$ is estimated according to the following equation.

$$\boldsymbol{s}^* = \boldsymbol{f}_{user}(\boldsymbol{t}_{a_0}) + \boldsymbol{J}_{\boldsymbol{f}_{user}}|_{Del_{in}}(\boldsymbol{t}^* - \boldsymbol{t}_{a_0}) \tag{3}$$

Note that $\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{Del_{in}}$ in Eq. (3) is the estimated Jacobian of $\boldsymbol{f}_{user}$ that is constant in the Del_{in}, and this $\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{Del_{in}}$ can be computed easily by evaluating Eq. (3) for each N points $\boldsymbol{t}_{a_1}, \ldots, \boldsymbol{t}_{a_N}$ as $\boldsymbol{t}^*$.

On the other hand, when $\boldsymbol{t}^*$ is outside of $CH(\{\boldsymbol{t}_i\})$, linear extrapolation is performed. Specifically, let $\boldsymbol{t}_{nearest}$ be the point in $CH(\{\boldsymbol{t}_i\})$ that is closest to $\boldsymbol{t}^*$ (note that $\boldsymbol{t}_{nearest}$ does not have to be the observation point), then $\boldsymbol{s}^*$ is estimated according to the following equation.

$$\boldsymbol{s}^* = \boldsymbol{s}^*_{nearest} + \boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{nearest}}(\boldsymbol{t}^* - \boldsymbol{t}_{nearest}) \tag{4}$$

Note that $\boldsymbol{s}^*_{nearest}$ is the estimated $\boldsymbol{s}$ at $\boldsymbol{t}_{nearest}$, and $\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{nearest}}$ is the estimated Jacobian of $\boldsymbol{f}_{user}$ at $\boldsymbol{t}_{nearest}$. $\boldsymbol{s}^*_{nearest}$ can be obtained by evaluating Eq. (3) for $\boldsymbol{t}_{nearest}$ as $\boldsymbol{t}^*$. $\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{nearest}}$ can be obtained as follows.

(1) Since $\boldsymbol{t}_{nearest}$ lies on the border of $CH(\{\boldsymbol{t}_i\})$, there exists a facet of $CH(\{\boldsymbol{t}_i\})$ which contains $\boldsymbol{t}_{nearest}$. Call this facet $Face_{nearest}$, and let $\boldsymbol{t}_{b_0}, \ldots, \boldsymbol{t}_{b_{N-1}}$ be the vertices of $Face_{nearest}$. Also, $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_{N-1})$ satisfying both $\sum_{i=0}^{N-1} \lambda_i \boldsymbol{t}_{b_i} = \boldsymbol{t}_{nearest}$ and $\sum_{i=0}^{N-1} \lambda_i = 1$ is uniquely determined.
(2) At each of $\boldsymbol{t}_{b_i}$, estimate the Jacobian of $\boldsymbol{f}_{user}$ ($\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{b_i}}$) by the following method.
   (a) Let T'_{b_i} be the set of points connected to $\boldsymbol{t}_{b_i}$ by an edge when the $CH(\{\boldsymbol{t}_i\})$ interior is partitioned by Delaunay tessellation.
   (b) $\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{b_i}}$ is estimated as $\boldsymbol{J} \in \mathbb{R}^{M \times N}$ that minimizes the L below. This method is an extension of [28] to multi-dimension.

$$L = \sum_{\boldsymbol{t}' \in T'_{b_i}} \left\| \frac{1}{\|\boldsymbol{t}' - \boldsymbol{t}_{b_i}\|} [\, \boldsymbol{f}_{user}(\boldsymbol{t}') - \{\, \boldsymbol{f}_{user}(\boldsymbol{t}_{b_i}) + \boldsymbol{J}(\boldsymbol{t}' - \boldsymbol{t}_{b_i}) \,\} \,] \right\|^2 \tag{5}$$

(3) Finally, $\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{nearest}}$ is estimated according to the following equation.

$$\boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{nearest}} = \sum_{i=0}^{N-1} \lambda_i \boldsymbol{J}_{\boldsymbol{f}_{user}}|_{\boldsymbol{t}_{b_i}} \tag{6}$$

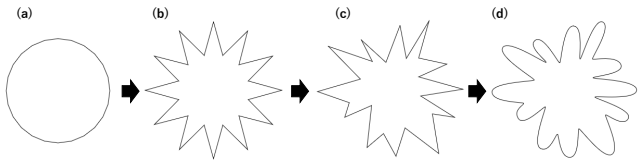

Figure 13: How 2D shape is generated from parameters. The parameters of this shape are *"number of spikes"*: 12, *"length of spikes"*: 0.3, *"randomness"*: 0.6, *"random seed"*: 5, *"roundness of the base of spikes"*: 0.15, *"roundness of the tip of spikes"*: 1.0 .

A.3 Method for Generating Shape from Semantic Parameters

The following shows how shapes are generated from six parameters (referring to Wada et al. [45]): *number of spikes* (0 to 30), *length of spikes* (0 to 0.5), *randomness* (0 to 1), *random seed* (0 to 10), *roundness of the base of spikes* (0 to 1), and *roundness of the tip of spikes* (0 to 1).

(1) Place "$number\ of\ spikes$" $\times$ 2 vertices evenly on the circumference of the circle (Fig. 13 (a)).
(2) Shift each vertex alternately outward and inward in the radial direction according to *"length of spikes"* (Fig. 13 (b)).
(3) Shift each vertex randomly in both radial and angular directions according to *"randomness"* and *"random seed"* (Fig. 13 (c)). The position of the vertices are now determined.
(4) Connect all two adjacent vertices with a quadratic Bézier curve (Fig. 13 (d)). Since two of the four Bézier curve operation points are end points of the curve, they are set at both vertices, and the remaining two points are determined according to *"roundness of the base of spikes"* and *"roundness of the tip of spikes"*.

A.4 Implementation Efforts to Run TimToShape in Real-Time on Browser

The followings are our implementation efforts to use TimToShape in real-time on browser.

- *Calculation* 1 in Section 4.2 is performed in a thread separate from the main thread using the "audio-worklet" functionality recently implemented in the browser. However, the calculation cost is high if the log-mel-spectrogram for 66,560 sample points is calculated simultaneously. The log-mel-spectrogram can be computed faster with $n\_fft = 2,048$ and $hop\_length = 1,024$ (see Section 3.2) by computing as follows.
  - Always keep the last 2,048 sample points as a buffer, and apply the mel-filter bank to the buffer and calculate the logarithm of it (call the result of it a "log-mel-vector" for convenience) every time 1,024 samples are updated. Then, the 64 consecutive log-mel-vectors will be the input of one frame for VAE. In this way, the VAE input can be updated every 1,024 sample points (23ms).
- The VAE calculation (*Calculation* 2 in Section 4.2) was implemented using "TensorFlow.js"[5], an open-source framework. The VAE model implemented by Keras was converted into a form usable for TensorFlow.js. By using this JavaScript framework, we implemented asynchronous inference by VAE To avoid bottlenecks in inference time.

[5] https://www.tensorflow.org/js