ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea
A07: MUSICAL ACOUSTICS
Numerical investigation of effects of mouthpiece geometry on flow and sound generation in a single-reed instrument

Tsukasa YOSHINAGA¹; Hiroshi YOKOYAMA¹; Tetsuro SHOJI²; Akira MIKI²; Akiyoshi IIDA¹

¹ Department of Mechanical Engineering, Toyohashi University of Technology, Japan
² Research and Development Division, Yamaha Corporation, Japan

ABSTRACT

Single-reed instruments produce sounds with fluid-structure-acoustic interactions in a mouthpiece, and it has been known that a small geometrical change may affect sound quality. In this study, we investigate the effects of mouthpiece geometry on the flow and sound generation in a single-reed instrument by conducting numerical flow simulations coupled with a structural analysis of reed vibration. The effects of mouthpiece geometry were explored by changing the mouthpiece outlet diameter. The flow and sound characteristics were predicted by solving the compressible Navier-Stokes equations with high-accuracy finite difference schemes, while the interaction with the reed vibration was simulated by solving one-dimensional dynamic beam equations. The results showed that the waveform of the reed tip opening was changed by the mouthpiece diameter, and the amplitudes of even number harmonics were reduced by the inwards tapers and matched mouthpiece diameter to the resonator inlet. The flow visualization revealed that the vortex rings are generated at the mouthpiece outlet, and the expansion rate of air in the mouthpiece increased earlier with the inwards tapers, suggesting that the timing of air expansion changed the reed tip waveforms and the sound quality of the single-reed instrument.

Keywords: Single-reed instrument, CFD, Aeroacoustics, Fluid-structure interaction

1. INTRODUCTION

The single-reed instruments, like a clarinet and saxophone, produce the sound with the interactions among the airflow from the mouth, reed oscillation, and acoustic resonance in the resonator. To investigate the nature of sound generation, many numerical simulations have been conducted for many years from one-dimensional modeling (e.g., 1-3) to three-dimensional flow simulations (e.g., 4-6). These numerical studies mainly focused on the comparison with the experiments and investigation of the player’s characteristics like blowing pressure and lip forces (5-6). Even portato and staccato articulations were reproduced in the simulation (3). However, a few studies examined the effects of instrument geometry on sound quality with the numerical simulations. Although da Silva et al., (4) performed the flow simulations on the different tip shapes of the mouthpiece geometry, the resulting sound characteristics are still not clear.

It has been known that the instrument geometry, especially mouthpiece geometry, has a great impact on sound generation (7) and its effects were widely discussed (8-11). Pillinger (11) investigated the effects of internal and external dimensions of mouthpieces on the quality of sound generated by an artificial blower. In addition, Ozdemir et al., (12-13) parametrically changed the dimensions of the saxophone mouthpiece with a 3D printer and investigated the sound quality as well as the blowing resistance and flexibility. However, detailed mechanisms of how the sound quality was changed are still unclear without the numerical simulation.

Therefore, in this study, we constructed five mouthpieces of Saxonett, one of the simplest single-reed instruments, with different bore diameters and conducted the flow simulation for the entire instrument geometry to clarify the effects of the mouthpiece geometry on the flow and sound generation. By calculating the compressible flow equations coupled with a one-dimensional dynamic beam equation, the physical phenomena occurring around the instrument are simulated in the computer.

* yoshinaga@me.tut.ac.jp
with different mouthpiece dimensions. The information about a tendency with the parametrical change of mouthpiece dimensions will contribute to the better design of instrument geometry.

2. METHODS

2.1 Single-reed Instrument

We use a Saxonett (JRS700, Jupiter) to simulate the sound generation of a single-reed instrument. The Saxonett consists of a clarinet mouthpiece and recorder-like resonator, and the mouthpiece is set in the artificial blower to mimic the player’s mouth. The Saxonett with the artificial blower is depicted in Fig. 1 (a). The pressure chamber of the artificial blower has dimensions of $70 \times 40 \times 43$ mm ($V = 120$ cm$^3$) which reproduces the average volume of the human vocal tract and trachea. On the far side from the mouthpiece, the inflow region is set with the constant flow supply. The tone holes were covered by the cylinder and the instrument produces the lowest note of C4. The resonator’s inner diameter is 13.3 mm, and the total length from the mouthpiece tip to the resonator outlet yields 307 mm. The axis $x_1$ is set in the longitudinal direction of the resonator, and the origin is set at the tip of the mouthpiece. The reed is fixed to the mouthpiece at the point of ligature ($x_1 = 38.5$ mm), and the lip force is set at $x_1 = 10.5$ mm with the lip width of 4.1 mm.

Details of the mouthpiece geometry are shown in Fig. 1 (b). From the inlet of the tip opening, the side walls of entry gradually expanded to the throat, and the throat connects to a cylindrical bore. The total length of the mouthpiece is 89.6 mm, and the bore has a length of 48.7 mm. In this study, the diameter of the bore outlet is changed from $d = 11.3$ to 15.3 mm. As a result, the taper angle of the bore increased from $-1.176^\circ$ to $1.176^\circ$ (minus value indicates the inwards taper). The dimensions are summarized in Table 1. The original diameter of the Saxonett mouthpiece bore is approximately 15 mm. However, the bore outlet smoothly connects with the resonator when $d = 13.3$ mm, because the inner diameter of the resonator is 13.3 mm. This smooth connection more accurately mimics the connection of the mouthpiece and barrel for the standard clarinet. There is a cavity with a length of 1.8 mm between the mouthpiece and the resonator for all cases.

![Figure 1 – Schematic of single reed instrument. (a) Resonator and mouthpiece of Saxonett and pressure chamber of artificial blower. (b) Mouthpiece geometry with different diameters. The dotted line shows the cross-section of $d = 11.3$ and 15.3 mm.](image)

<table>
<thead>
<tr>
<th>Total length of mouthpiece (mm)</th>
<th>Bore length (mm)</th>
<th>Outlet Diameter $d$ (mm)</th>
<th>Cross-sectional area $A$ ($mm^2$)</th>
<th>Taper angle (deg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.3</td>
<td>100.3</td>
<td>$-1.176$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.3</td>
<td>118.8</td>
<td>$-0.588$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>89.6</td>
<td>48.7</td>
<td>13.3</td>
<td>138.9</td>
<td>0</td>
</tr>
<tr>
<td>14.3</td>
<td>160.6</td>
<td>0.588</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.3</td>
<td>183.9</td>
<td>1.176</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.2 Numerical Simulation

To consider the fluid-structure-acoustic interactions in the single-reed instrument, we numerically solve the three-dimensional compressible Navier-Stokes equations. The compressible flow simulation enables to calculate the sound generation from the flow passing through the oscillating reed. In addition, a one-dimensional beam equation was solved to consider the fluid-structure interaction of the reed oscillation. The deforming reed with the surrounding airflow is expressed by the volume penalization (VP) method, one of the immersed boundary methods (14). The VP method expresses the wall with porous material in the structured grids, and we fix the parameter of porosity to realize the acoustic reflectivity of 99%. The spatial derivatives are solved by a sixth-order-accuracy compact finite difference scheme, while the time integration is performed using the third-order-accuracy Runge-Kutta method. To simulate the turbulent flow in the mouthpiece, the large eddy simulation (LES) is applied with tenth-order-accuracy spatial filter. These methodologies enable to simulate the interactions between jet flapping and acoustic resonance in a recorder (15).

The reed oscillation is simulated by solving a dynamic Euler-Bernoulli beam equation with the external force of airflow, lip, and contact forces. The lip force is calculated as a linear stiffness which is proportional to the distance between the lip position and the reed surface. The spatial derivatives are solved by a second-order accuracy finite-difference scheme, whereas the time integration is performed using the implicit θ-scheme (16). The material constants like Young’s modulus and reed density are chosen from the experimental values, while the damping and internal loss coefficients are manually tuned by comparing the reed tip waveforms of the experimental measurements (17). The details are presented in our previous paper (6). In addition, the lip stiffness is adjusted to obtain the self-sustained oscillation with the typical tip opening.

The computational grids are shown in Fig. 2. The minimum grid size is 0.025 mm at the tip of the mouthpiece, and the grid size is gradually expanded towards the outlet to reduce computational costs. To resolve vortices generated near the bore outlet of the mouthpiece, the grid sizes around the connecting part between the mouthpiece and resonator are kept smaller than 0.2 mm. The total grid number is approximately 1.72 × 10^8.

As a boundary condition, a uniform flow of flow rate of 138 cm^3/s and pressure of 5.5 kPa were set at the inflow region of the artificial blower. We confirmed that the mean pressure inside the chamber was kept at 5.5 kPa for the entire simulation time for all mouthpiece geometry. At the outlet of the computational domain, a non-reflecting boundary with a buffer region is set to prevent the reflection from the outlet. The far-field sound was evaluated by sampling the pressure at 100 mm from the outlet of the resonator. The time step for the time integration was set to 2.68 × 10^{-8} s. As the initial condition, the pressure values in the artificial blower, mouthpiece, and resonator are set to 5.5 kPa, and the initial tip opening was set to 1.1 mm which is the rest position of the reed for the original Saxonett mouthpiece.

![Figure 2](image.png)

Figure 2 – Computational grids around the mouthpiece. The brown and yellow parts show the regions of the rigid wall and oscillating reed, respectively.

3. RESULTS AND DISCUSSION

The waveforms of the tip opening \( h \) are plotted from the initial position to \( t = 52 \) ms for mouthpieces with \( d = 11.3, 13.3, \) and 15.3 mm in Fig. 3. At the initial state, no lip force was applied for three cycles (up to \( t = 14 \) ms) and added the lip force with the stiffness of \( K_{lip} = 2.3 \times 10^5 \) N/m^2 at \( t = 14 \) ms. Then, the lip stiffness was slightly decreased to \( K_{lip} = 2.2 \times 10^5 \) N/m^2 at \( t = 28 \) ms to keep
the amplitudes of $h$ around 0.5 mm for the rest of the time steps. The stiffness of $K_{lip} = 2.3 \times 10^5$ and $2.2 \times 10^5$ N/m$^2$ can be converted to approximately 2.24 and 2.14 N, respectively, based on the maximum of $h$. By keeping the lip stiffness, the self-sustained oscillation was obtained with almost the same displacement amplitudes. The oscillation frequency was the same with different diameters until the lip force was applied, and then, the frequency increased with the lip force for the smaller diameter ($d = 11.3$ mm) whereas the frequency decreased with the larger diameter ($d = 15.3$ mm), compared to $d = 13.3$ mm.

The last three cycles of each mouthpiece diameter are plotted in Fig. 4 with normalized time by its period. The time $t = 0$ is defined at the beginning of the closed phase. The fundamental frequencies of the mouthpiece with $d = 11.3$, 12.3, 13.3, 14.3 and 15.3 mm were $f_0 = 271.8$, 269.0, 268.3, 265.7, and 264.5 Hz, respectively. With the increase of $d$ from 11.3 to 13.3 mm, the maximum of $h$ decreased from 0.55 to 0.47 mm. Then, with the increase of $d$ from 13.3 to 15.3 mm, the maximum of $h$ increased from 0.47 to 0.51 mm. The maximum of $h$ appeared in the earlier phase when $d < 13.3$ mm, whereas the maximum of $h$ appeared in the later phase when $d > 13.3$ mm. With the smooth connection from the mouthpiece to the resonator ($d = 13.3$ mm), the waveform became a rectangular shape in the open phase, and this waveform was similar to those observed in the previous experiments of the clarinet with relatively low tones (18-19).

![Figure 3 – Waveform of tip opening $h$ with different diameters $d = 11.3$, 13.3, 15.3 mm.](image)

![Figure 4 – Waveform of tip opening $h$ with normalized time for the last three cycles.](image)

The spectra of sound pressure sampled at 100 mm from the resonator outlet are plotted in Fig. 5 (a) for $d = 11.3$, 13.3, and 15.3 mm. The frequency is normalized by the fundamental frequency $f_0$ of the reed oscillation. The maximum amplitude of approximately 99 dB appeared at the fundamental tone for all cases, and the amplitude decreased with the increase of frequency. The even number harmonics (i.e. second and fourth harmonics) were 10 to 30 dB smaller than the other odd number harmonics (i.e. first, third, and fifth harmonics), and this result indicates that the Saxonett in the current setup produced the clarinet-like sounds (7) in the same way as the experiment (17).

The discrepancies in sound pressure levels (SPLs) of each harmonic from that of the fundamental tone (SPL$_{f0}$) are plotted in Fig. 5 (b). The second harmonic decreased from SPL − SPL$_{f0} = −30$ to $−39$ dB with the increase of $d$ from 11.3 to 13.3 mm and increased from SPL − SPL$_{f0} = −39$ to $−34$ dB with the increase of $d$ from 13.3 to 11.3 mm. These values are correlated with the shape of the tip waveform $h$, indicating that the rectangular shape of the waveform in $h$ with the smooth connection of $d = 13.3$ mm reduces the amplitude of the second harmonic. While the third and fifth harmonics were almost
the same values with different $d$, the fourth and sixth harmonics increased with the increment of $d$. These results suggest that the second harmonic is correlated with how the mouthpiece smoothly connects with the resonator, whereas higher even harmonics (fourth and sixth) are correlated with the angle of taper of the mouthpiece. If we assume that the typical clarinet tone is formed by the smaller amplitudes of even harmonics (7), the taper angle should be as small as possible with the smooth connection to the resonator in the current setup. Meanwhile, the higher amplitudes in higher harmonics of $d = 15.3$ mm indicate that clearer and brighter tones are generated with the larger bore diameters, and this result is consistent with the previous experiment (11).

![Image](image_url)

Figure 5 – Sound pressure levels sampled at 100 mm from the resonator outlet. Spectra of sound pressures are plotted in (a) with different diameters $d = 11.3$, 13.3, and 15.3 mm. The frequency is normalized by the fundamental frequency $f_0$. The differences of amplitudes of harmonics from the fundamental tone are plotted in (b).

The instantaneous flow pressures and vortex structures in the instrument are visualized in Fig. 6 for $d = 11.3$, 13.3, and 15.3 mm at $t/T = 0.9$. To visualize the vortex structures, the second invariant of velocity gradient tensor $Q$ is calculated, and iso-surfaces of $Q = 10^5$ $s^2$ are depicted with a contour plot of the pressure distribution on the medial plane of the instrument. At $t/T = 0.9$, the reed tip started closing and the velocity at the mouthpiece outlet was the maximum. The turbulent jet entered through the tip opening, and the vortices are dissipated at the throat of the mouthpiece for all cases. When $d = 11.3$ mm, the inwards taper increased the velocity in the bore, and vortex rings were generated from the outlet edge of the bore. On the contrary, the vortex rings were generated at the resonator edge with $d = 15.3$ mm, because the flow directly impinging on the resonator edge due to the larger diameter of the mouthpiece. The vortex tubes were also generated at the cavity of tone holes. The pressure contours showed that the pressure decreased earlier with the smaller $d$, whereas the higher-pressure values remained in the resonator with the larger $d$.

The waveforms of pressure and velocity in the longitudinal direction $u_1$ at the center of the mouthpiece outlet ($x_1 = 89.6$ mm) are shown in Fig. 7 for one cycle. The pressure of $d = 11.3$ mm increased earlier than that of $d = 13.3$ mm, whereas the pressure of $d = 15.3$ mm increased later than that of $d = 13.3$ mm in the reed’s opening phase. The maximum and minimum values of pressure waveforms were almost the same for different mouthpiece diameters. In contrast, the velocity at the mouthpiece outlet started increasing at almost the same time for different diameters. However, the slope of the increase was different, and the velocity reached the maximum of 11.1 m/s for $d = 11.3$ mm, while the maximum was 9.1 m/s for $d = 15.3$ mm.

To elucidate the relationship between the flow and reed tip waveform, the expansion rate $\rho U_2 - \rho U_1$ in the mouthpiece was calculated with mass flow rate $\rho U_1$ at the tip opening and mass flow rate $\rho U_2$ passing through the bore outlet at $x_1 = 89.6$ mm. The waveforms of the expansion rate with three diameters are plotted in Fig. 8. The gray part in Fig. 8 indicates the positive pressure period in the mouthpiece. When the pressure became positive at around $t/T = 0.4$, the expansion rate started decreasing, and the air in the mouthpiece compressed in all cases. The compressed air pushes the reed from the inside, and the reed tip started opening. Then, at $t/T = 0.55$, the expansion rate started increasing, pushing the reed further to open. This expansion rate increased earlier with the smaller diameter $d = 11.3$ mm, whereas the expansion rate increased later with the larger diameter $d = 15.3$ mm. These timings of increase in the expansion rate are correlated with those of the reed tip waveforms, indicating that the different bore diameters changed the expansion rate of air in the mouthpiece and changed the reed tip waveforms. This tip opening waveform changed the frequency.
characteristics of sound sources near the reed tip and resulted in different far-field sound quality. The three-dimensional fluid-structure interaction simulation enabled to show how the flow and sound generations were changed in the mouthpiece with different dimensions, and this helps makers to improve the design of the single-reed instrument.

Figure 6 – Instantaneous pressure distributions and vortex structure visualized by the iso-surfaces of second invariant of velocity gradient tensor $Q = 10^7 \text{s}^{-2}$. The pressure and vortices are depicted with $d = 11.3$ mm (a), 13.3 mm (b) and 15.3 mm (c).

Figure 7 – Waveforms of pressure (a) and velocity $u_1$ (b) at the center of the mouthpiece outlet ($x_1 = 89.6$ mm). The time is normalized by the period of one cycle.
Figure 8 – Expansion rate $\rho U_2 - \rho U_1$ in the mouthpiece. The mass flow rate $\rho U_1$ is measured at tip opening whereas $\rho U_2$ is measured at the mouthpiece outlet ($x_1 = 89.6$ mm). The time is normalized by a period of one cycle. The gray part shows the positive pressure period, while the other parts show negative pressure periods.

4. CONCLUSIONS

In this study, the effects of mouthpiece geometry on the flow and sound generation in a single-reed instrument were investigated by conducting compressible flow simulations. By changing the diameter of the bore outlet, the phase of the maximum tip opening was shifted, and the amplitudes of even number harmonics were changed. The results suggest that the smooth connection between the mouthpiece and resonator reduces the amplitude of second harmonics, while the increase of the inwards taper angle reduces the amplitudes of the other even harmonics and produces more clarinet-like sounds. The flow visualization revealed that the vortex rings were generated at the edge of the mouthpiece connection, and the expansion rate increased earlier with inwards tapers, whereas the expansion rate increased later with the outwards tapers. These results suggest that the expansion rate of air in the tapered mouthpiece changes the reed tip waveforms and resulting sound quality. In future work, we need to evaluate the blowing resistance and flexibility from the current analysis for the design of better instruments. In addition, further analysis of the relationship between the harmonic amplitudes and the expansion rate in different playing conditions is expected.

ACKNOWLEDGEMENTS

This work was partly supported by JSPS Grant-in-Aid for Early-Career Scientists (KAKENHI, Grant number: JP20K14648). We acknowledge H. Onitsuka, K. Arimoto, and R. Baba of Yamaha Corporation for helpful comments and discussion.

REFERENCES

Spectral composition of the attack portion of a recorder tone: How the player can affect the tone color

Nicholas GIORDANO1; Jared W. THACKER
Auburn University, USA

ABSTRACT
We have used Navier-Stokes-based simulations to study the spectral content of tones produced by a soprano recorder. Our focus is on the attack portion of a recorder tone and how it can be influenced and controlled by the player. We show that both the blowing speed and the time taken to initiate blowing can be used to control the amplitudes and time variations of the upper partials during the attack. Our simulations are consistent with previous experimental results reported in the literature for flue instruments and with a new analysis of tones produced by a soprano recorder. The results illustrate how a player can control the expressiveness of a recorder tone.

Keywords: Musical instruments, recorder, Navier-Stokes

1. INTRODUCTION
Physical modeling has yielded important insights into tone production by a wide variety of musical instruments. Wind instruments have been especially challenging for such studies, since the fundamental equations of fluid mechanics, the Navier-Stokes (NS) equations, are a complicated set of nonlinear equations that require a computational approach when applied to a realistic instrument. Even so, available computer resources now make it possible to study a variety instruments using the NS equations. During the course of such simulations we have observed interesting that interesting spectral content is predicted during the attack phase of tones produced by flue instruments. This behavior was first noted some time ago in NS simulations in (1); in this paper we report further studies which give some qualitative insights into the behavior and also show that similar spectra features are found in simple tones produced by a recorder played by a beginner (i.e., one of the authors). It is also worth noting that unknown to the authors at the time of reference (1), the same basic behavior has been reported in a number of previous studies of flue instrument tones.

2. Modeling Details
We use an explicit finite-difference-time-domain algorithm to numerically solve the compressible 3D Navier-Stokes equations for a soprano recorder. Simulation details are described elsewhere (1). The numerical model had dimensions similar to the Yamaha YRS23 soprano recorder, with a chamfer on the lower edge of the windway (Fig. 1) and an inner diameter that tapered as one moves from the labium to the open end. The simulations positioned the instrument inside a closed “box” (which could be visualized as a very small room) with walls that reflect and damp sound produced by the instrument. The calculations yielded the air density and velocity inside and outside the recorder. Variations in the density are proportional to the variations of the pressure, and the results below shown below show the sound pressure at a location above the labium. The sound pressure beyond and off the axis of the open end exhibit similar behavior. In previous work we have verified [1] that the absolute magnitude of the sound pressure inside the instrument was approximately 130 dB, which is comparable to that found in real instruments. The calculated value of the sound pressure outside the recorder and about 10 cm distant and off axis from the open end was smaller by typically two orders of magnitude.

All of the results in this paper were obtained with the tone holes in Fig. 1 all closed so that the

1 njg0003@auburn.edu
instrument produced its lowest note. For the soprano recorder this is C5, which has a fundamental frequency of approximately 525 Hz. These tones also contained components (higher partials) at frequencies that are approximately integer multiples of the fundamental frequency.

Figure 1. Model of the recorder studied in the simulations. The distance from the exit of the channel the labium was 4.0 mm, and from the end of the channel to the end of the resonator was 28 cm. The model contained tone holes that could be open or covered as if by a player. For all the results in this paper the tone holes were all closed.

3. Simulated Tones

Figure 2a shows the sound pressure outside the instrument for the note C5 when the recorder plays at a relatively low volume. Specifically blowing was initiated at t = 0 with the blowing speed ramped up uniformly to its final value of 10 m/s (in the windway) at t = 5 ms. This choice of ramp up time is at the low end of times reported in experiments in the literature but seems consistent with what can be achieved by human players. This blowing speed is then maintained until 250 ms at which time the blowing speed is ramped to zero and the tone dies away. From Fig. 2b it is seen that a steady state oscillation with a frequency of about 525 Hz (corresponding to the pitch of C5) is reached by about t = 60 ms and maintained thereafter. The time it takes to reach steady state is much greater than the ramp up time due to the time it takes for the sound to reach steady state inside the recorder tube and in the somewhat reverberant enclosure than surrounds the instrument.

Figure 3 shows spectra analysis of the waveform in Fig. 2. These different spectra were calculated using Fast Fourier transforms of data segments extending over about 20 ms centered at times given in the caption. The blue spectrum, calculated at 80 ms, shows the steady state behavior. This spectrum is dominated by the first partial at frequency f1 = 525 Hz, with only a very small peaks at the second partial (f2 approximately 1050 Hz) and the third partial (f3 approx. 1600 Hz). The position of the second partial is extremely weak so we have assumed its frequency is twice f1. Interestingly, the position of the third partial at the earliest time appears to be a bit higher than found at longer times; a similar shifting of the partial frequencies will be seen at other blowing pressures below.

Figure 2. Simulated sound pressure versus time for the note C5 played by a soprano recorder at a blowing speed of 10 m/s and a ramp up time of 5 ms. (a) Full note. (b) Early portion of the tone.
Figure 3. Spectra for the sound pressure waveform in Fig. 2 showing the approximate positions of the first, second, and third partials of the tone at different times. Black: Fourier spectrum calculated using data centered at t=20 ms using data extending approximately 10 ms before and after that time in Fig. 2b. Red: Spectrum at t=40 ms. Blue: Spectrum at t=80 ms. Note that the vertical scale is logarithmic.

Another way to view how the spectral composition of the tone in Fig. 2 varies during the course of the tone is to plot the peak powers of the different partials as functions of time; see Fig. 4. The first partial dominates during the entire course of the tone, with the power at the second and third partials being typically three orders of magnitude smaller at all times. In other words, the amplitudes of the partials are typically smaller by a factor of at least 30. This tone, produced at a low blowing pressure, is thus very close to being a pure tone.

Figure 4. Power at the peaks of the first, second, and third partials in Fig. 3 as a function of time during the full course of the tone. Black: First partial. Red: Second partial. Blue: Third partial.

The spectral properties of the simulated tones change as the recorder is blown harder. Figure 5 shows results for the early part of a tone produced with a steady state blowing speed of 12 m/s. The only difference between this tone and the one in Fig. 2 is that the steady state blowing speed was increased from 10 to 12 m/s. There is now a region between about 30 and 60 ms in which there seems to be a new spectral component (in addition to the first partial near 525 Hz). Figure 3 shows spectra analysis of the waveform in Fig. 2. These different spectra were calculated using Fast Fourier transforms of data segments extending over about 20 ms centered at times given in the caption. The orange spectrum, calculated at 100 ms, shows the steady state behavior. This spectrum is dominated by the first partial at frequency $f_1 = 525$ Hz, with only a very small peaks at the second partial ($f_2$ approximately 1050 Hz) and the third partial ($f_3$ approx. 1600 Hz). At early times, especially at 30, 40 and 60 ms, the third partial is quite significant. Also apparent at early times are substantial components around 800 and 1300 Hz; these frequencies do not appear to be related to the base ($f_1$) component of the tone and we refer to them as inharmonic partials. We will consider their origin below.
Figure 5. Simulated sound pressure versus time during the initial (attack) portion of the tone for the note C5 played by a soprano recorder with a blowing speed of 12 m/s and a ramp up time of 5 ms.

Figure 6. Spectra for the sound pressure waveform in Fig. 5 showing the approximate positions of the first, second, and third partials of the tone at different times. Also indicated are spectral components referred to as inharmonic partials in the text. Black: Fourier spectrum calculated using data centered at t=20 ms using data extending approximately 10 ms before and after that time in Fig. 5b. Red: Spectrum at t=30 ms. Blue: Spectrum at t=40 ms. Green: Spectrum at t=60 ms; Orange: Spectrum at 100 ms.

Figure 7 shows how the power at the different partials vary with time for the tone produced by blowing at 12 m/s. While the first partial again dominates throughout the tone, at early times (before about 30 ms) the third and inharmonic partials are about an order of magnitude weaker than the first partial, corresponding to an amplitude smaller by about a factor of 3-5. This is consistent with what is seen by eye in the waveform in Fig. 5.

We next consider a simulated tone produced by blowing at the even larger blowing speed of 15 m/s. Again, all the other conditions and parameters are the same as for the two tones considered earlier. The resulting waveform at the first part of this new tone is shown in Fig. 8. While the behavior at long times is again nicely periodic with the expected frequency (525 Hz) a range of times during which other spectral components are important is again seen. In addition, the time required to reach steady state is significantly longer than found at a blowing speed of 12 m/s, which was itself longer than found at 10 m/s. The length and spectral nature of the attack portion of the tone thus vary systematically as the blowing speed is increased.
Figure 7. Power at the peaks of the first, second, and third partials (labeled as f1, f2, and f3), and the inharmonic partial near 800 Hz (fin) in Fig. 6 as functions of time.

Figure 8. Simulated sound pressure versus time during the initial (attack) portion of the tone for the note C5 played by a soprano recorder with a blowing speed of 15 m/s and a ramp up time of 5 ms.

Spectra at several times during the course of this tone are shown in Fig. 9. The first partial is seen to increase with time, while the second partial initially increases and then decreases as the steady state is reached. The third partial behaves in an opposite manner, being small initially and then increasing at long times. In addition, the second partial seems to be somewhat smaller than 2xf1 at early times and then shifts to a slightly higher frequency at longer times. The uncertainties for the third partial are greater, but it too seems to shift some as time progresses.

The behavior of the different partials is shown in Fig. 10 and we see that the second partial now dominates until about 80 ms before becoming smaller than the third partial after about 160 ms. The spectral composition of the tone is thus complex and varies substantially during the course of this tone, which a listener would presumably still recognize as C5.
Figure 9. Spectra for the sound pressure waveform in Fig. 5 showing the approximate positions of the first, second, and third partials of the tone with blowing speed 15 m/s at different times. Black: Fourier spectrum calculated using data centered at $t=30$ ms using data extending approximately 10 ms before and after that time in Fig. 5b. Red: Spectrum at $t=80$ ms. Blue: Spectrum at $t=100$ ms. Green: Spectrum at $t=200$ ms.

Figure 10. Power at the peaks of the first, second, and third partials in Fig. 9 as a function of time during the full course of the tone. Black: First partial. Red: Second partial. Blue: Third partial.

The results in Figs. 2-10 show that our simulations predict a systematic variation of the tone color as the recorder is blown harder, with an essentially pure tone at low blowing speeds to a tone in which the attack is dominated by the second partial. We should note that these tones are all in the blowing range in which C5 is produced; the regime change transition to C6 occurs at a higher blowing speed near 18 m/s. That transition and a fuller discussion of these tones will be discussed in a future paper.
Figure 10. Simulated sound pressure versus time during the initial (attack) portion of the tone for the note C5 with a blowing speed of 12 m/s and a ramp up time of 10 ms.

Figure 12. Spectra for the sound pressure waveform in Fig. 11 showing the approximate positions of the first (f1), third (f3), and inharmonic (fin) partials of the tone with blowing speed 12 m/s and ramp up time 10 ms, at different times during the tone (Fig. 11). Black: Fourier spectrum calculated using data centered at t=50 ms using data extending approximately 10 ms before and after that time in Fig. 11. Red: Spectrum at t=70 ms. Blue: Spectrum at t=100 ms.

4. EXPERIMENTAL OBSERVATIONS

Tones produced by flue instruments – organ pipes, recorders, and flutes – have been studied for many years by many researchers. The phenomena of interest to us in this paper have been observed previously (see for example Refs. [2-7], which consider both organ pipes and recorders), but to the best of our knowledge have not been analyzed in quite the ways or in the kind of detail for a particular instrument as required for comparison with the modeling work presented earlier in this paper. We therefore next present an analysis of a single tone produced by a real soprano recorder and analyze that tone in the same way as we have analyzed the simulated tones discussed earlier. The recorder is a Yamaha model YRS25 (a student quality instrument composed of plastic), and we consider the lowest note produced by this instrument with all tone holes closed, C5, the same tone as considered in our simulations.

Figure 11 show the waveform for a typical C5 tone produced by the YRS25 blown by one of the
authors; (a) shows the entire tone while (b) shows the initial, attack portion in more detail. An attempt was made to ramp the blowing speed smoothly to a final value that was then maintained until the end of the tone. The periodicity achieved around 0.08 ms continued until the tone finished.

Figure 11. Sound pressure versus time waveform for the note C5 produced by a soprano recorder as played by one of the authors. (a) Full note. (b) Attack portion of the tone. Note that t=0 in (b) has been shifted to be nearer to the start of the tone.

An interesting feature seen in Fig. 12 is the way the frequencies of the second and third partials vary small but significant amounts during the course of the tone. We believe that these frequency shifts are real, and note that they resemble similar shifts seen in the simulations.

Figure 13 shows how the power in each partial varies with time. It is notable that the second partial is as large or nearly as large as the first partial until about 80 ms, while the first partial dominates by about an order of magnitude or more at longer times. This again is similar to the behavior of the simulated tones at the larger blowing speeds.

Figure 12. Spectra showing the first, second, and third partials of the tone during different portions of the attack. Black: Fourier spectrum calculated using data centered at t=10 ms in Fig. 1b. Red: Spectrum calculated at t=30 ms. Blue: Spectrum calculated at t=100 ms. The approximate positions of the first (f1), second (f2), and third (f3) partials are indicated.
Figure 13. Power of the first, second, and third partials of the tone as functions of time during the full course of the tone. Black: First partial. Red: Second partial. Blue: Third partial.

It is seen that the second partial is essentially equal in power to the first partial for the first 50 ms of the tone with the first partial then dominating at longer times. It is well known (see for example [8-10]) that a human listener is able to perceive pitch and timbre during portions of a tone as short as 20 ms, so the change in timbre evidenced in Fig. 13 is certainly perceptible.

One other feature of Fig. 12 is also worth noting. The frequencies of the different partials appear to also vary some with time. This is most noticeable with the second partial which moves to higher frequency at long times, while the frequency of the third partial seems to fall and then increase during the course of the tone.

5. CONCLUSIONS

In a broad sense, the simulation results are quite similar to that found in the real tone considered in Figs. 11-13. During the attack the higher partials are as large or larger than the first partial, thus contributing very significantly to the timbre. It is well known (see for example [8-10]) that a human listener is able to perceive pitch and timbre during portions of a tone as short as 20 ms, so the changes in timbre we have observed are certainly perceptible. For the real tone it is the second partial that competes for early dominance while for the simulated tone it is the third partial. This difference is certainly significant, but here we have considered simulation results for only one note at one blowing speed and with one ramp up time. Other simulations, which will be reported in the future, show that the behavior during the attack phase depends on those parameters. Our simulations have also found that the behavior can depend on the specific instrument geometry, so while the simulation model in Fig. 1 is intended to resemble the Yamaha soprano recorder, the effect of remaining differences between the geometry of the actual instrument and the model still need to be explored and understood.

Our results also clearly show the effect of the player. The recorder is often the first instrument taught to a young student, presumably because it is easy for a beginner to produce a nice tone, in contrast to many other instruments, such as the violin. However, as with other instruments, it is important for the player to have some ability to affect the timbre. For the case of the recorder, simulations like those presented in this paper show that the blowing speed and the ramp up time can both profoundly affect the timbre, especially during the attack portion of the tone.

As a general result, our results suggest that Navier-Stokes-based simulations can be used study rather detailed aspects of flue instruments, and address subtle but important questions concerning the tones produced by real instruments. We plan to report more extensive results of this kind in the future.
ACKNOWLEDGEMENTS
This work was supported in part by the U.S. National Science Foundation through grant PHY1806231.

REFERENCES
Numerical study of a French horn mouthpiece accompanied by vibrating lips and an oral cavity with compressible direct numerical simulation

Rei SUMITA(1), Ryoya TABATA(1), Sho IWAGAMI(1), Yuki NAKAHARA(1), Takeshi NANRI(2), Taizo KOBAYASHI(2), Yuji HATTORI(3), Kin’ya Takahashi(1)

(1)Kyushu Institute of Technology, Japan, takahasi@phys.kyutech.ac.jp
(2)Kyushu University, Japan, tkoba@cc.kyushu-u.ac.jp
(3)Institute of Fluid Science, Tohoku University, JAPAN, yuji.hattori@tohoku.ac.jp

ABSTRACT
A two-dimensional model of a French Horn mouthpiece is numerically studied with a 2D direct numerical simulation (DNS) of the compressible Navier-Stokes equations to investigate the sound generation mechanism from the viewpoint of aeroacoustics. That is, we consider the sounding mechanism of buzzing, when the mouthpiece without a bore is played. Our numerical tool is highly accurate due to the minimum mesh size of the order of the micro-meter, and details of fluid motion and acoustic oscillation inside and near the mouthpiece are successfully reproduced. In particular, we focus on the roles of vibrating lips and an oral cavity in the sound generation mechanism. When the mouthpiece without lips and an oral cavity is driven by a periodic flow with a certain frequency, a single tone without overtones is observed. On the other hand, when the mouthpiece is driven by vibrating lips with an oral cavity, a generating sound includes rich overtones and its waveform is similar to that observed experimentally. Since the bore is a linear element and cannot generates overtones from a single tone by itself, the sound of a horn including rich overtones is generated by a mouthpiece with the vibrating lips and oral cavity.

Keywords: French Horn Mouthpiece, Buzzing, Aerodynamic sound, compressible DNS

1 INTRODUCTION
The sounding mechanism of brass instruments is a long-standing important problem in the field of musical acoustics and many authors have contributed to this problem [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. The dynamics of brass instruments is well modeled by delay differential equations under the one-dimensional approximation, for example, the Adachi-Sato models[1, 7, 8], and such models well reproduce oscillating motions in the mouthpiece and even in the bore. The vibrating lips of a player, namely lip reeds, are the key for understanding the sound generation mechanism. The motion of lips is roughly categorized into two types of reed valves, the so-called “outward-swinging door” and “sliding door”, which affect characteristics of generating sound, e.g. behavior of the sound frequency with respect to the lip resonance frequency [1, 5, 6, 7, 8]. In reality, the lips of a player behave as a mixture of these models. For the models of wind instruments with an oscillating reed(s), we merely assume that a volume flow injected through the oscillating reed slit \( U \) is changed to acoustic pressure \( p \) in the mouthpiece as \( p = Z_0 U \), where \( Z_0 \) is the characteristic acoustic impedance, and this assumption apparently works well[1]. However, we do not have a clear answer to the problem of how and where an acoustic motion is generated from a fluid flow injected through the oscillating reed slit. This is a problem of aeroacoustics and the generating sound is regarded as an aerodynamic sound[15, 16]. Thus, we need to take an approach based on aeroacoustics to reproduce and understand the sound generation process[14].

In this study, we focus on the sound generation process observed in a French horn mouthpiece detached from a resonance tube from the viewpoint of aeroacoustics. Actually, brass players use the practice of playing a mouthpiece alone, called mouthpiece buzzing, and can play notes on a scale as possible. Note that the mouthpiece on brass instruments works as a Helmholtz resonator and enhances resonance peaks of the air column, i.e., notes, in a middle range of its register[1]. Actually, we numerically reproduce the sound generation...
process as accurate as possible by using a compressible fluid solver, which simultaneously reproduces fluid and acoustic motions. To do this, we adopt a compressible Direct Numerical Simulation (DNS), which is considered the most accurate and reliable scheme, although other less accurate schemes, e.g. Large Eddy Simulation (LES) and Lattice Boltzmann Method (LBM), are often used and provide some successful results [14, 17, 18, 19, 20]. Due to the ability of supercomputers at present, we adopt a two-dimensional (2D) model and consider the effects of vibrating lips together with an oral cavity on the characteristics of fluid motion and generating sound [14, 17, 18]. In the previous work[14], we briefly reported the result of the model with vibrating lips without an oral cavity. In this paper, we discuss the effect of the vibrating lips in detail and also discuss the effect of the oral cavity combined with the vibrating lips, because the existence of the oral cavity should affect the sound generation process to a greater and less extent.

2 MODEL AND NUMERICAL METHOD

Figure 1 shows the cross-section of a French horn mouthpiece that we adopt and the dimensions of its 2D model to which an oral cavity is connected through a lip reed slit. We prepare three models which are different in the geometry of the lip reed slit and oral cavity: the model without the lip reed slit and oral cavity called “Mouthpiece”, the model with the lip reed slit called “Mouthpiece+lip” and the model with the lip reed slit and oral cavity called “Mouthpiece+lip+cavity”.

We consider the following four cases. In the “Case 1”, the “Mouthpiece” model is driven by a Poiseuille flow with the maximum velocity of $V = 10\text{ m/s}$, and a frequency of damped oscillation of pressure in the mouthpiece is observed to detect the resonance frequency of the mouthpiece, the Helmholtz resonance frequency. In “Case 2”, to mimic a flow injected through the vibrating lip reeds, the “Mouthpiece” model is driven by a periodic flow like an alternating current with velocity given by

$$V = V_0 \left(1 - \cos \omega_{d} t \right)/2,$$

where $V_0 = 10\text{ m/s}$ and $\omega_{d}$ is the resonance angular frequency obtained in the “Case 1”. In “Case 3”, the effect of the vibrating lips is considered with the “Mouthpiece+lip” model. The lip reed slit is changed alternatively opening and closing with period $T_d = 2\pi/\omega_{d}$ while a constant flow with a velocity of $V = 10\text{ m/s}$ is injected from the left end of the slit. In “Case 4”, the effect of the oral cavity together with the vibrating lips is considered with the “Mouthpiece+lip+cavity” model. The lip reed slit is changed as “Case 3”, while a constant flow with a velocity of $V = 1\text{ m/s}$ is injected from the left end of the oral cavity so that the flow velocity approximately reaches 10m/s at lip reed slit when it is opened.

To perform numerical calculation, we adopt a compressible Direct Numerical Simulation (DNS) developed by Komatsu et al.[17]. By using the higher-order finite difference schemes with the volume penalization (VP) method, the two-dimensional (2D) compressible Navier-Stokes equations are solved[17, 18]. Figure 2 (a) and (b) show the numerical grids around the mouthpiece for the “Mouthpiece” model and “Mouthpiece+lip+cavity” model, respectively. The points $P_i$ and $P_o$ indicated by the filled yellow circles are observation points inside and outside the mouthpiece, respectively. For the “Mouthpiece” model and “Mouthpiece+lip” model, the mouthpiece is embedded in a solid box. However, for the “Mouthpiece+lip+cavity” model, the outer shape of the mouthpiece is reproduced to explore a more realistic performance. The minimum grid size near the mouthpiece is
of $0.737 \times 10^{-2}$ mm, which is enough to reproduce flows in boundary layers near the solid walls inside the mouthpiece, lips and oral cavity. The entire computational domain is of $2000 \times 2000 \text{mm}^2$ for all cases. The calculation is done up to $0.01 \text{sec}$ for each case. Table 1 shows mesh parameters and time step for the “Mouthpiece” model and “Mouthpiece+lip+cavity” model. Table 2 shows physical parameters of compressible fluid. The simulations were performed on the high-performance computing cluster ITO (subsystem A) with 1 node of 36 cores (36 parallel threads) in Kyushu University taking dozens of days for each calculation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of grids</th>
<th>Minimum grid size</th>
<th>Time step $\Delta t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouthpiece</td>
<td>$2806 \times 2480$</td>
<td>$0.737 \times 10^{-2} \text{mm}$</td>
<td>$1.8 \times 10^{-9} \text{s}$</td>
</tr>
<tr>
<td>Mouthpiece+lip+cavity</td>
<td>$2807 \times 3096$</td>
<td>$0.737 \times 10^{-2} \text{mm}$</td>
<td>$1.8 \times 10^{-9} \text{s}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>parameters</th>
<th>Density at rest $\rho_0$</th>
<th>Viscosity $\mu$</th>
<th>Sound speed $c_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>values</td>
<td>$1.162 \text{kg/m}^3$</td>
<td>$1.846 \times 10^{-5} \text{Pa} \cdot \text{s}$</td>
<td>$347.127 \text{m/s}$</td>
</tr>
</tbody>
</table>

### 3 Numerical results

#### 3.1 Case 1: Mouthpiece model driven by a constant flow

The horn mouthpiece is a Helmholtz resonator, but it is not easy to obtain its resonance frequency theoretically, because the neck cannot be clearly distinguished from the body. Thus, we estimate the resonance frequency from a damping oscillation of pressure in the mouthpiece, when a constant flow is injected into it.

Figure 3 (a) shows the velocity distribution at $t = 0.01 \text{s}$ and figure 3 (b) shows the change in pressure at the observation point $P_i$ in the mouthpiece. As shown in Figure 3 (a), an air-jet goes straight for a short time and is rolled up into an eddy near the throat entrance. Then, the rolled-up eddy is divided into two parts: the greater part goes along the upper or lower wall of the cup, and the remaining part enters the throat, first forming a wavy jet and then seemingly changing to an acoustic wave.

However, as shown in Figure 3 (b), the pressure oscillation, which is regarded as an acoustic oscillation, decays in a short time. This is because a constant flow cannot sustain permanent oscillation with a definite pitch. Next, we estimate the resonance frequency without using the Fourier transform as follows. If $T'$ denotes
3.2 Stationary oscillations for Case 2, Case 3 and Case 4

For Case 2, Case 3 and Case 4, stationary oscillations of acoustic waves are observed on the inside and outside of the mouthpiece. Let us see fluid motions in the mouthpiece at first. Figure 4 shows the velocity distributions at \( t = 0.01 \) s near the mouthpiece for the three cases. In Case 2 shown by Figure 4 (a), the periodically injected flow changes to vortices in the cup and the major parts travel into the throat giving rise to an acoustic oscillation. Compared with Case 1, fewer vortices are observed in the cup and the wavy jet almost disappears in the throat. In the Case 3 shown by Figure 4 (b), the air jet injected through the vibrating lips is broken into small pieces of vortices distributing irregularly in the cup and some parts go into the throat causing an acoustic oscillation. In Case 4 shown by Figure 4 (c), the air jet injected through the vibrating lips is broken into small pieces of vortices, which distribute more densely in particular on the left side than those of Case 3. Small pieces of vortices are also observed on the right side of the oral cavity. However, an acoustic oscillation together with a few vortices is observed in the throat.

Figure 5 shows the pressure distributions at \( t = 0.01 \) s in the whole calculation region for the Case 2 and Case 4. For Case 2, a stationary spherical wave is radiated from the open end of the mouthpiece. For Case 4, a stationary wave is also observed, but the radiated wave is not a sinusoidal-type wave and is rather modulated. Furthermore, it is smaller in amplitude than that in Case 2.

Figure 6 shows pressure oscillations at the observation points Pi and Po for the three cases. For Case 2, almost sinusoidal oscillations are observed in the stationary states at both points. However, the oscillation at the point Pi is much larger in amplitude than that at the point Po, because a small piece of the resonance wave leaks as a radiation wave from the open end. For Case 3, the pressure oscillation at the point Pi is almost sinusoidal, while the oscillation at the point Po is apparently modulated. The individual oscillations at the points Pi and Po are slightly smaller in amplitude than the corresponding ones for Case 2. For Case 4, the pressure oscillation at the point Pi is apparently modulated due to the effect of the oral cavity. Indeed, a pressure oscillation is also observed in the oral cavity and affects the motion in the mouthpiece through the oscillating reed slit. Similar pressure oscillations were observed in the mouthpieces of Brass instruments in experiments[1, 4]. The oscillation at the point Po is much modulated including a large amount of the 2nd and 3rd harmonic components. The individual oscillations at the points Pi and Po are much smaller in amplitude than the corresponding ones for Case 2.

Figure 7 shows the power spectra of the pressure oscillation at the observation point Po for the three cases. For Case 2, ignoring a tiny 2nd harmonic component, only the fundamental component dominates the oscillation so that an almost sinusoidal wave is observed. For Case 3, the fundamental component is still dominant, but the 3rd harmonic component is secondly dominant and the 2nd and 4th harmonic components appear with smaller
Figure 4. Velocity distributions at $t = 0.01$ s. (a) Case 2. (b) Case 3. (c) Case 4.

Figure 5. Pressure distributions at $t = 0.01$ s. (a) Case 2. (b) Case 4.
Figure 6. Pressure oscillations at the observation points Pi and Po. (a) Observation point Pi of Case 2. (b) Observation point Po of Case 2. (c) Observation point Pi of Case 3. (d) Observation point Po of Case 3. (e) Observation point Pi of Case 4. (f) Observation point Po of Case 4.
amplitudes; thus, the wave is apparently deformed as shown Figure 6 (d). Although the result is not shown here, the power spectrum of the oscillation at the point Pi includes some amount of the harmonics, in particular lower-order harmonics. Thus, we think that these 2nd to 4th harmonic components are relatively enhanced in the oscillation at the point Po due to the characteristic of the radiation impedance of the mouthpiece. For Case 4, the fundamental and the 3rd harmonic are in the same order in the spectrum and the 2nd harmonic also take a large value in amplitude; thus, a wave is periodic but includes 2nd and 3rd harmonic components as shown Figure 6 (f). Even for the power spectrum of the oscillation at the point Pi, lower-order harmonic components take large values in amplitude and the wave is apparently deformed as shown Figure 6 (e). Thus, such harmonic components are relatively enhanced in the oscillation at the point Po due to the characteristic of the radiation impedance of the mouthpiece. Therefore, the nonlinear interaction between the mouthpiece and the oral cavity through the vibrating lips plays an important role for creating an acoustic wave with rich harmonics.

4 CONCLUSIONS
In this paper, we numerically studied the sound generation process of the French horn mouthpiece detached from a resonance tube, namely mouthpiece buzzing, with the 2D compressible DNS. In particular, we focused on the influence of the oral cavity combined with the vibrating lips on the fluid and acoustic motions in the mouthpiece as well as on the radiated acoustic wave. To do this, we treated the three models, which are individually different in geometry and driven by different ways, and found the following facts concerning the construction of harmonic components in the acoustic oscillation.

When the mouthpiece is driven by a periodic flow with the resonance frequency of the mouthpiece, the acoustic oscillation involves negligible harmonic components except the fundamental. Even only the model including the vibration of lip reeds however induces some amount of lower-order harmonic components on the acoustic oscillation in the mouthpiece. Since the radiation impedance of the mouthpiece relatively enhances these harmonic components, the radiated wave observed in a far field is considerably deformed. The existence of the oral cavity combined with the vibrating lips significantly enhances harmonic components in the acoustic oscillation in the mouthpiece, which is similar to those observed in experiments on brass instruments[1, 4]. Since the radiation impedance of the mouthpiece relatively enhances the harmonic components of the radiated wave, an acoustic wave with rich harmonics is observed in a far field. Therefore, the nonlinear interaction
between the mouthpiece and the oral cavity through the vibrating lips plays an important role in the sound generation process.

Unfortunately, we did not deal with the problem of how and where an acoustic oscillation is generated from a fluid flow injected through vibrating lips. To attack this problem, we need to take an approach based on the aeroacoustic theory, e.g. Lighthill’s acoustic analogy [15, 16]. For another, to explore the interaction between the mouthpiece and oral cavity through the vibrating lips in detail, we also need an analysis of the interaction between compressible fluid and vibrating lips. Further developments in these directions are left for future work.

ACKNOWLEDGEMENTS
The present work was supported by Grant-in-Aid for Scientific Research (C) Nos. 19K03655 from the Japan Society for the Promotion of Science (JSPS) and “Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures” and “High Performance Computing Infrastructure” in Japan (Project ID: jh220001). Part of the work was carried out under the Collaborative Research Project of the Institute of Fluid Science, Tohoku University.

REFERENCES


Developing an interactive drum model: Using the discrete cosine transform

Tim KIRBY\(^1\), Mark SANDLER\(^2\)

\(^1\)QMUL, UK, t.c.kirby@qmul.ac.uk
\(^2\)QMUL, UK. mark.sandler@qmul.ac.uk

ABSTRACT

This paper expands upon the analysis and synthesis technique proposed by Kirby and Sandler in 2021 (1), which investigates how the evolution of drum modes change with strike intensity. This was observed in the discrete cosine transform (DCT) of unwindowed drum sounds and enabled the synthesis of drum sounds whose intensity is intermediate between those analysed. In this paper, the 67 modal features of a tom-tom drum are modelled analytically, to move towards a complete synthesis. The phase function of each modal feature is modelled with a linear rational function. The amplitude function of the modal feature is first decomposed into odd and even components, and each component fitted using functions such as an inverse Gaussian. The evolution of the parameters of these analytical functions can be mapped onto a single controllable parameter that represents strike intensity. This leads to an analytical model of a modal feature that can be used to build a physical model of a tom-tom, which will be illustrated by an interactive demonstrator.

Keywords: Analysis, Synthesis, Drum, Modelling

1. INTRODUCTION

Music production settings are becoming increasingly reliant on virtual instruments to augment or replace traditional acoustic setups. Of particular interest here are drum synthesis techniques that incorporate physical modeling. These instruments are not limited to prerecorded samples; model parameters can be continuously varied through user automation or randomisation to improve realism or create experimental sounds. Physical drum models can provide the user with complete control over the specifications of the instrument and its surroundings, and these instruments can be extremely responsive in terms of strike position, velocity, gesture, and tool.

At the same time, synthesising realistic drum sounds is still an open problem. Finite difference methods, such as those generated in the NESS project (2), have produced some of the most realistic results, but these models are computationally intensive and can’t be run in real time. Hybrid methods have been used to bridge this gap. IK Multimedia’s Modo Drums (3) is one example, which manipulates existing samples using modal synthesis in real-time. This is an interesting approach, though it is much more limited than a full physical model and the manipulated samples can still sound synthetic.

This research aims to develop a real-time interactive drum model that is informed by reference samples, but is not reliant on them. Here we dynamically model modal features to move towards a full physical drum model. The aim is to show that it is possible to analytically model modal features accurately enough that expected time domain behaviours can be generated.

1.1 Framework Recap

In (1), it was shown that the discrete cosine transform of an unwindowed tom-tom sound reliably contains modal features, \(X_{m,n}(\omega)\), which encode the full time domain activity of a given mode. A representative modal feature is shown in Figure 1. These modal features have a chirp like form which can be decomposed into instantaneous amplitude and phase functions, also shown in Figure 1.
Figure 1. (Top) DCT representation of the fundamental frequency of a 9x10" Yamaha Beech Custom tom-tom, struck at the moderately high velocity of 102 out of 127. This is $X_{0,1}(\omega)$, a representative example of a chirp-like modal feature, with an unmodulated fundamental frequency of 131 Hz. (Middle) Instantaneous amplitude (envelope) of the above modal feature, $A_{0,1}(\omega)$, extracted using the Hilbert transform. (Bottom) Unwrapped instantaneous phase of the modal feature, $\phi_{0,1}(\omega)$, also extracted via the Hilbert transform.

An ideal circular membrane has Bessel-function solutions in the radial direction, and cosine function solutions in the azimuthal direction. These solutions correspond to modes, which are classified by their number of nodal diameters, $m$, and their number of nodal circles (4). Modes can be written in the form $(m, n)$, with the fundamental frequency being $(0, 1)$. The instantaneous amplitude, $A_{m,n}(\omega)$, and the instantaneous phase, $\Phi_{m,n}(\omega)$, evolve with strike velocity in a clear fashion. This evolution was demonstrated by interpolating the amplitude and phase functions extracted from real drum samples and using these to synthesise modes of intermediate velocity.

Here, these functions are modelled analytically, to work towards a complete modal model that doesn’t rely on reference samples. This type of model would be much more flexible, as it can be fully parameterised to model drums of arbitrary specification. This research represents the first attempt at modelling these functions analytically. The aim is to investigate whether the evolution of $A_{m,n}(\omega)$ and $\Phi_{m,n}(\omega)$ can be modeled with simple functions, to a degree that is sufficient to return the expected behaviour in the time domain. The analytical functions used here are not intended to capture every single nuance found in the reference data, but rather to model the general evolution of these reference functions, as a first approximation. The same goes for the time domain activity, the aim is not to synthesise an exact copy of the original time domain signal, but to synthesise a version that has a very similar pitch trajectory and envelope, and is perceptually convincing.

If the expected time domain activity is recovered, it will serve two purposes. Firstly, it would provide a proof of concept for a novel drum synthesis technique. Secondly, it would help us to better understand how the IDCT representation is linked to the time domain activity, by understanding the relative perceptual importance of different aspects of the IDCT representation. There has been little prior work in this type of synthesis. The closest method is arguably IFFT synthesis, which was introduced in 1980 (5). This has been used to generate large ensembles of sinusoids with fixed frequencies (6), but it hasn’t been used to model the more complex non-linear modal activity discussed here.

2. METHOD

2.1 Data Set

Tom-tom samples were extracted from Superior Drummer 3 by Toontrack (www.toontrack.com) by triggering every sample from 1 to 127 midi velocity with all velocity layers loaded and all hit variation features turned off. Each sample is 8s long to ensure a full decay. The following analysis is based on the 9x10" Yamaha Beech Custom tom-tom but is applicable to any tom-tom and is expected to transfer to any drum that exhibits pitch glide.

Central strikes were used as this is the best strike position to excite the fundamental mode. Out of the 68 unique samples obtained, only 1 was deemed to be anomalous as the modal feature did not fit
the trend and it is audibly off-centre. The integrated loudness, as defined by (7), was also calculated for each drum sound using the “integratedLoudness” command in MATLAB (The MathWorks, Natick, MA).

2.2 Modeling the Phase Function

The phase function is the most interesting component of a modal feature, as it encodes the time domain pitch glide of the drum mode. The unwrapped phase function of the fundamental, \( \Phi_{0,1}(\omega) \), was extracted as before, as shown in Figure 2. The active regions of the phase functions were estimated to be within \((130 < f < 180\text{Hz})\). This only includes the varying parts of the function. The plateaus found at \((f \lesssim 130\text{Hz})\) merely indicate the lack of sinusoidal oscillation that is found within the active region of a modal feature.

![Figure 2. Extracted phase functions for the fundamental of a 9x10" Yamaha Beech Custom tom-tom (top), for comparison with the linear rational function model (bottom). The phase functions of all unique sample in the data set are overlaid. The maximum magnitude of phase in the curved section increases with strike velocity, as does the frequency width. The active regions were estimated to be within 130 < f < 180Hz.](image)

The active region was modelled with a linear rational function:

\[
\Phi_{m,n}(\omega, v) = \frac{a_{m,n}(v)\omega + b_{m,n}(v)}{\omega + c_{m,n}(v)}
\]

Where \(a_{m,n}, b_{m,n},\) and \(c_{m,n}\) are real-valued constants, for a given strike velocity, \(v\).

The parameters were fitted using the ‘rat11’ option in MATLAB. The parameters were first fitted using random start points, which is the default option. This provided the expected results for the vast majority of strike velocities. The fitting was then repeated using a starting point of the median value of each parameter from the original fits (\(a_{m,n} = 0.5095, b_{m,n} = -483.9, c_{m,n} = -854\)). This method provided repeatability and provided the expected results for all 67 samples.

This is illustrated in Figure 2, where the model is compared to reference data. The active region of the model appears to be virtually indistinguishable by eye, with the exception of the transition from the plateaus into the active region of the model feature \(f \approx 130\text{Hz}\). The was deemed negligible in this model, but this could be smoothed out or modeled analytically if necessary.

2.3 Evaluation of Phase Function Model

Figure 3 displays the coefficient of determination, \(R^2\), values obtained when the model phase functions are compared to their respective reference function. The mean value is 0.9811 and the standard deviation is 0.0240, which supports the idea that the model is a good fit of the reference function. The model performs particularly well for the range of strike velocities where there is
significant pitch glide. This is reassuring as this is arguably the most important region. The $R^2$ values decrease somewhat for the lowest strike velocities, though this at least partly due to the change in shape of the reference phase function.

The phase function encodes the time domain pitch glide of the drum mode. It is therefore worth comparing the pitch trajectories of the reference data with that obtained by replacing $\Phi_{m,n}(\omega)$ with the linear rational function model. This pitch trajectories are shown in Figure 3. As expected, the pitch trajectories of the model are not identical to the reference pitch trajectories, however, they do follow a similar evolution, with both the magnitude of the pitch glide and it’s duration, increasing with strike velocity, to the same approximate values. This is further demonstrated in Figure 4, where the magnitude of the pitch glide for both reference and model data, is plotted against strike velocity.

![Figure 3](image1.png)

**Figure 3.** (Top) Goodness of fit analysis using the coefficient of determination, $R^2$. Comparison of each model phase function with the respective reference. (Middle) Pitch trajectories from reference modes obtained using the pYIN method (8) in Sonic Annotator (9). (Bottom) Pitch trajectories obtained with linear rational function model. All strike velocities are overlaid. The magnitude of the pitch glide increases with strike velocity for both the reference and model.

![Figure 4](image2.png)

**Figure 4.** The estimated change in pitch of the fundamental, as a function of strike velocity. Integrated loudness is used as a proxy for strike velocity. Pitch estimates were obtained using the pYIN method (8) in Sonic Annotator (9). The blue circles are for the reference data, the red circles are for the model data.
The linear rational function model has 3 parameters \((a,b,c)\). The evolution of these parameters with strike velocity is shown in Figure 5. Each parameter was clearly correlated with the integrated loudness of the drum sample, and in each case, this correlation was modelled with a dual exponential. This means that these three parameters can be controlled by a single parameter which represents strike velocity.

As the model phase function was able to recover the expected pitch glide behaviour in the time domain, and the model parameters evolved in a smooth and continuous fashion, the linear rational function model is deemed suitable as a first approximation.

Figure 5. Evolution of phase function parameters with strike velocity. Integrated loudness is used as a proxy for strike velocity.

### 2.4 Modeling the Amplitude Function

The amplitude function contributes to the envelope of the time domain signal. At low strike intensities, the amplitude function is a bell-shaped curve. As strike intensity increases, the area under the curve increases, and the peak becomes increasingly asymmetrical. A characteristic dip to the left of the peak becomes present, as shown in Figure 6.

Figure 6. Evolution of the decomposed amplitude function with strike velocity, for the fundamental of a 9x10” Yamaha Beech Custom tom-tom. The amplitude function components of all unique samples between midi velocity 1-66 are overlaid, as these illustrate the general trend in a clear manner. (Top) Evolution of the amplitude function, \(A_{0,1}(\omega)\). (Middle) Evolution of the even component of the amplitude function, \(A_{0,1}^{e}(\omega)\). The area under \(A_{0,1}^{e}(\omega)\) increases with strike velocity. (Bottom) Evolution of the odd component of the amplitude function \(A_{0,1}^{o}(\omega)\). The area under \(A_{0,1}^{o}(\omega)\) increases with strike velocity.
No single function was found that could adequately model the evolution of this function, so the amplitude function was decomposed into its odd and even components. This was done by taking the discrete cosine transform of the amplitude function. This yields a vector that has two distinct trends, one found in the odd numbered samples, and one in the even numbered samples. These trends were separated by duplicating this vector and setting the unwanted (odd or even) samples to zero. These vectors can then be transformed back to the original domain using the inverse discrete cosine transform, to yield the odd and even components of the amplitude function. The odd and even symmetry is about the line \( f=131\text{Hz} \), which is the unmodulated fundamental frequency. These components are shown in Figure 6.

The addition of these components returns the full amplitude function, \( A_{0,1}(\omega) \). The even component, \( A_{0,1}^e(\omega) \), is the dominant component at all strike velocities. At the lowest strike velocities the odd component, \( A_{0,1}^o(\omega) \), is negligible as \( A_{0,1}(\omega) \) is itself is essentially an even function. As strike velocity increases, \( A_{0,1}^o(\omega) \) becomes increasingly significant, increasing the asymmetry of \( A_{0,1}(\omega) \). The odd and even components are simple enough to model, and modeling them separately provides the user with a high degree of control over the shape of the amplitude function. Both components are centred on the unmodulated fundamental frequency which is intuitive and allows the unmodulated fundamental frequency to easily be parameterised, to model a membrane of arbitrary tension or radius. Each component exhibits (odd or even) symmetry about this central frequency, which means that there is redundancy, so only half of each function needs to be modelled.

There is some distortion that occurs to the right of the peak of the amplitude function at the highest amplitudes. This was previously shown to have little perceptual effect on the time domain signal (1). Nonetheless, it complicates the decomposition of the amplitude function, so these samples are not included in this first approximation modeling of the amplitude function. This distortion could be included in later models.

### 2.5 Modeling the Odd Component

The right half of \( A_{0,1}^o(\omega) \) resembles a probability distribution. This was originally modeled as an inverse Gaussian, but this has now been superseded by a linear rational function:

\[
A_{m,n}^o(\omega,v) = \frac{d_{m,n}(v)\omega + e_{m,n}(v)}{\omega^2 + f_{m,n}(v)\omega + g_{m,n}(v)}
\]  

(2)

Where \( d_{m,n}, e_{m,n}, f_{m,n}, \text{ and } g_{m,n} \) are real-valued constants, for a given strike velocity, \( v \).

The parameters were fitted using the 'rat12' option in MATLAB. The parameters were first fitted using random start points, which is the default option. The fitting was then repeated using a starting point of the median value of each parameter from the original fits \( (d_{m,n} = 1.2219, e_{m,n} = 0.0338, f_{m,n} = -6.5728, \text{and } g_{m,n} = 69.7205) \).

### 2.6 Modeling the Even Component

It is easier to model \( A_{0,1}^e(\omega) \) in the transformed domain, before the inverse discrete cosine transform step described in Section 2.4. The transformed vector alternated between positive and negative values. A dual exponential was used to model the magnitude of this vector:

\[
A_{m,n}^e(\omega,v) = h_{m,n}(v) \cdot e^{i_{m,n}(v)\Omega} + j_{m,n}(v) \cdot e^{k_{m,n}(v)\Omega}
\]

(3)

Where \( h_{m,n}, i_{m,n}, j_{m,n}, \text{ and } k_{m,n} \) are real-valued constants, for a given strike velocity, \( v \). \( \Omega \) is the transformed version of \( \omega \) after the DCT has been performed on the amplitude function.

### 2.7 Evaluation of Amplitude Model

Figure 7 displays the modeled versions of the amplitude components, along with their superposition which yields the full amplitude model.
Figure 7. Modeled evolution of the amplitude function (top) produced from the superposition of the even component model (middle) and the odd component model (bottom).

Figure 8 displays the coefficient of determination, $R^2$, values obtained when the amplitude function model is compared to the reference function. The top graph shows the values for the full amplitude model. This comparison is made in the time domain, to assess whether the model recovers the expected time domain behaviour. The mean $R^2$ value is 0.9970 with a standard deviation of 0.0048, which suggests a very good fit across all tested strike velocities (The majority of values are even larger, but they begin to decrease at higher strike velocities). This supports the conclusion that the amplitude model is sufficient to capture the expected time domain behaviour.

![Figure 8](image)

Figure 8. Goodness of fit analysis using the coefficient of determination, $R^2$. (Top) Comparison in the time domain of a mode synthesised using the full amplitude model, against the reference mode. (Bottom) Comparison of the modeled odd and even components against their respective reference component. This comparison was made in the domain in which each component is modeled.

The bottom graph compares the separate odd and even component models, in the domain in which they were modelled. The dominant even component values are all above 0.9963, which again suggests a good fit across all strike velocities. The odd component has much lower values at the lowest strike velocities. This is because the odd component is almost negligible at these strike velocities, as evidenced by the fact that the values for the full amplitude model appear unaffected.
3. CONCLUSION

This model was intended as a proof of concept. Both the phase function and the amplitude function have been suitably modeled to return modal behaviour in the time domain, that is consistent with that expected for a tom-tom, while not quite identical to the reference data. It is unclear whether any differences are perceptible; a listening test could be used to investigate this. This model can be parameterised to model modal behaviour in a dynamic fashion. Each parameter can be mapped to a single master parameter which represents strike velocity. This facilitates the creation of an interactive demonstrator that could be triggered by a control surface such as an electronic drum kit.

Transformations can also be used to apply this model to drums of different specification. It may also be possible to link the model functions and their parameters to the underlying physics of a drum.

The phase model could be improved by modelling the transition from the plateau into the active region. This could be done analytically, or a machine learning model could be trained on the differences between this analytical model and the reference data.

The amplitude model appears to perform extremely well over the tested data. Some samples were excluded from the amplitude model as these over complicated the decomposition and analysis of this initial proof of concept modeling. Now that this proof of concept has been achieved, these samples could be modeled in detail.

ACKNOWLEDGEMENTS

This research was funded by Queen Mary University of London. The author would also like to thank Toontrack, and all those involved in creating the meticulously sampled library of drums sounds found in Superior Drummer 3, which has served as an invaluable dataset. Thanks are also given to all those in the Centre for Digital Music for their valued support.

References
Modeling and measurements of organ pipe sound radiation

Samuel D. BELLOWS¹; Timothy W. LEISHMAN²

¹,² Brigham Young University, USA

ABSTRACT

Pipe organs have played an important role in Western musical genres for centuries and are unique in their acoustic radiation as sound is produced from numerous pipes. At low frequencies, the directivity of an individual organ pipe can be roughly modeled using one or two point sources, whose source strengths can be determined, for example, using transmission matrix methods. This work studies the limitations of this approach in predicting the directivities of organ pipes at higher frequencies involving diffraction effects. Modeled directivities are compared to those measured with a spherical array and with a near-field cylindrical array to illustrate the differences.

Keywords: Directivity

1. INTRODUCTION

The transmission matrix method (TMM) represents a complex system as a one-dimensional transmission line, allowing a simple yet robust approach to modeling many musical instruments (1). Because the one-dimensional model is approximately valid as long as the instrument bore remains small compared to the acoustic wavelength, this approach is particularly useful in predicting the frequencies produced using various fingerings of wind instruments even for higher frequencies. While it is common to utilize the TMM to predict played frequencies and input impedance curves, the approach also helps predict the directional characteristics of musical instruments. Nonetheless, the effectiveness of the TMM in predicting directivities have remained inconclusive, partly due to the lack of high spatial resolution directivity measurements for comparisons.

This work contrasts TMM-based directivity models to measured spherical directivities. In particular, this investigation applies the TMM to organ pipes due to their simple natures. Results show that the TMM-based techniques successfully predict directivities at low frequencies. However, BEM simulations and near-field cylindrical scans reveal that the neglecting of diffraction effects limits the effectiveness of the approach.

2. THEORETICAL MODELING

At low frequencies, a simple point-source model reasonably describes the radiation of sound from the organ pipes (2, 3). For open pipes, two sources are necessary: one at the mouth and one at the open end. For a closed pipe, only a single source at the mouth is necessary. For time-harmonic radiation, the pressure at a point \( r \) is

\[
p(r, k) = iz_0 k \sum_{n=1}^{N} U_n G(r, r_n),
\]

where \( N \) is the number of openings, \( U_n \) is the frequency-dependent volume velocity at the nth opening.

¹ samuel.bellows11@gmail.com
²tim_leishman@byu.edu
$i$ is the imaginary unit, $z_0 = \rho_0 c$ is the characteristic specific acoustic impedance of the fluid, $k$ is the wavenumber, and
\[
G(r, r_\text{n}) = \frac{e^{-ik R_n}}{4 \pi R_n}
\]  
(2)

is the free-space Green’s function with
\[
R_n = |r - r_\text{n}|
\]  
(3)

being the distance from the observation position $r$ to the location $r_\text{n}$ of the $n^{th}$ opening.

When considering normalized far-field directivities, the volume velocity of the mouth is arbitrary for a closed pipe with only one opening. However, for an open pipe, the impedance translation theorem relates the volume velocity between the two openings (1, 3). The TMM simplifies the numerical implementation for more complex sources with many openings (3). The acoustic radiation impedance of the open end of the pipe can be modeled as an unflanged tube:
\[
Z_{AR,O} = z_0 \frac{1}{S} \left( 0.25 (ka)^2 + i 0.6 k a \right).
\]  
(4)

The acoustic radiation impedance of the mouth can be roughly modeled as a baffled circular piston:
\[
Z_{AR,M} = z_0 \frac{1}{S} \left( 1 - \frac{I_1(2ka)}{ka} + \frac{i H_1(2ka)}{ka} \right).
\]  
(5)

Translating the impedance of the open end through length $L$ to the mouth opening yields
\[
Z_A = Z_{A0} \frac{Z_{AR,O} + i Z_{A0} \tan (kL)}{Z_{A0} + i Z_{AR,O} \tan (kL)},
\]  
(6)

where $Z_{A0} = z_0 / S$ and $S$ is the pipe’s cross-sectional area. A current divider circuit gives the volume velocity at the mouth as
\[
U_m = U_{\text{in}} \frac{Z_A}{Z_A + Z_{AR,M}},
\]  
(7)

where $U_{\text{in}}$ is an arbitrary input volume velocity. Finally, the volume velocity at the end of the pipe is given by another current divider circuit:
\[
U_O = (U_{\text{in}} - U_m) \frac{-i Z_{A0} \csc(kL)}{i Z_{A0} (\tan(kL/2) - \csc(kL)) + Z_{AR,O}}.
\]  
(8)

Thus, a simple model can estimate the volume velocities at the pipe openings for predicting far-field directivities using Eq. (1).

2.1 Kirchhoff-Helmholtz Integral Equation

A more complete representation of the pressure field follows from the Kirchhoff-Helmholtz Integral Equation (KHIE) (4):
\[
p(r) = iz_0 k \int u_n(r_s) G(r, r_s) dS + \int p(r_s) \frac{\partial}{\partial n} G(r, r_s) dS,
\]  
(9)

where $u_n$ is the normal particle velocity, the integral is taken over the boundary $S$ and $n$ is the normal direction into the domain. In this form, the first integral of the KHIE can be roughly interpreted as the radiation of sound from the source due to surface vibrations, whereas the second integral can be interpreted as the pressure field due to diffraction and scattering effects of the boundary.

If one considers the boundary surface $S$ to fully enclose the exterior of the pipe, it is evident that
Eq. (1) approximates the first integral of the KHIE with

\[ U_n = \int \int u_n(r_s)dS_n, \]

(10)

where \( S_n \) is the surface bounding the \( n \)th pipe opening. With this approach, the openings of the pipe behave as vibrating pistons of effective areas \( S_n \). This approximation should remain valid provided that the acoustic wavelength is larger than the pipe openings. However, Eq. (1) neglects the second integral in the KHIE. Thus, when this term becomes important, deviations between the modeled and measured directivities will arise. Section 5 explores this limitation by comparing measured results with a BEM simulation, which numerically evaluates Eq. (9).

3. MEASUREMENTS

The authors employed spherical and cylindrical arrays to measure the pressure fields produced by two organ pipes: one open metal pipe and closed wood pipe. The spherical array, shown with the closed wooden pipe in Fig. 1(a), consisted of 36 12.7 mm (0.5 in) microphones placed in 5-degree increments in the polar angle at a radius of \( a = 0.97 \) m. The cylindrical array, shown with the open metal pipe in Fig. 1(b), consisted of 36 6.35 mm (0.25 in) microphones placed in 5 cm increments at a circular radius of \( \rho = 0.25 \) m. A turntable rotated each array in 5-degree azimuthal steps. Relative calibrations for each microphone array and a near-field reference microphone ensured normalization between the repeated measurements using frequency-response functions (FRFs) [(5)].

![Figure 1. Spherical (a) and cylindrical (b) microphone arrays used to assess the radiated fields of two organ pipes.](image)

4. RESULTS

4.1 Theoretical Prediction

Figure 2 shows simulated directivity balloons for the first through sixth partials of the open metal pipe using Eq. (1) with \( L = 0.46 \) m. Both radius and color indicate levels on a 40 dB scale. The mouth of the pipe faces the \( \phi = 0^\circ \) azimuthal marker. For the fundamental, the directivity appears as a single disc-like shape. As anticipated for a two point-source model, as frequency increases, the number of lobes also increases. The directivity patterns are nearly axisymmetric about the line connecting the center of the top opening of the pipe with the mouth. Because the wooden pipe has only one radiating
opening, the simulated directivity pattern for all frequencies is omnidirectional.

Figure 2. Simulated directivity balloons of the open metal pipe using Eq. (1) for the first six partials: (a) 331 Hz, (b) 662 Hz, (c) 993 Hz, (e) 1323 Hz, (f) 1654 Hz, and (f) 1985 Hz.

4.2 Directivity Measurements

Figure 3 shows the directivity balloons of the measured metal pipe. The patterns are remarkably similar, especially for the first three partials below 1 kHz. As for the simulated case, the balloons are roughly axisymmetric about the axis connecting the openings. However, at higher frequencies, reduced levels appear behind the pipe (near $\phi = 180^\circ$) that reduce the axial symmetry.
Figure 3. Measured directivity balloons of the open metal pipe for the first six partials: (a) 331 Hz, (b) 662 Hz, (c) 993 Hz, (e) 1323 Hz, (f) 1654 Hz, and (f) 1985 Hz.

Figure 4 shows the measured closed wooden pipe. While for the fundamental, the radiation is roughly omnidirectional, at higher frequencies, particularly above 1 kHz, the radiation is significantly reduced behind the instrument. Small undulations appear, although they are weak and not nearly as pronounced as for the open pipe.

Figure 4. Measured directivity balloons for the measured closed wooden pipe for the first six partials: (a) 335 Hz, (b) 670 Hz, (c) 1005 Hz, (e) 1340 Hz, (f) 1675 Hz, and (f) 2010 Hz.
5. ANALYSIS

As suggested by the measurement results, simple point-source models provide reasonable estimates of source directivities at lower frequencies. However, reduced levels behind the pipes suggest that diffraction around the pipe itself may play an important role in the radiation patterns, particularly at higher frequencies. An enhanced model using a BEM evaluation of the KHIE provides additional insights into this assertion.

To simulate the pressure while maintaining similarity to the modeling approach used in Eq. (1), a single boundary enveloping the exterior of the pipe describes the source geometry. This approach divides the pipe into an exterior and interior portion so that the openings of the pipes become inhomogeneous Neumann (velocity) boundary conditions. Then, as shown in Eq. (10), the net volume velocity used in Eq. (1) is the surface integral of the particle velocity at each opening. The TMM method predicts the particle velocity at each opening by dividing the predicted values from Eqs. (7) and (8) by the area $S_n$ of the opening. While a more accurate model would follow from treating the pipes as thin-walled structures and allowing the computational domain to include both the interior and exterior of the pipe, this method is useful because it maintains a consistent approach between the simple point-source model of Eq. (1) and the first term of the KHIE. The second term of the KHIE, numerically evaluated using BEM, can then be seen as a correction term to the original model.

Because the field produced by the second term decays at a faster rate than the $1/r$ dependence of the first term due to the normal derivative of the Green’s function, this field is more apparent in the near-field of the source. Consequently, measurements produced by the cylindrical array are beneficial for identifying the effects of the second term in the KHIE.

Figure 5 plots results for the fundamental of the closed wooden pipe. Figure 5(a) shows the pressure evaluated on the cylindrical sampling surface as simulated by Eq. (1) with the point source centered at the mouth of the instrument. As anticipated, the pressure is nearly uniform for constant height $z$, although the levels slightly increase in front of the pipe mouth and decrease behind the pipe as the mouth is not directly at the center of array. Figure 5(b) shows the simulated pressure using only the first term of the KHIE. Because the wavelength is much larger than the opening, Eq. (10) remains a reasonable approximation and the pressure remains similar to that predicted by the point-source model. Figure 5(c) shows the simulated pressure when both terms in the KHIE are included. The inclusion of the second term significantly decreases the levels behind the instrument, and the area in front of the mouth is clearly seen as a red circular patch. The BEM simulation agrees well with the measured data shown in Fig 5(d), which also has a similar red circular patch in front of the pipe mouth. Thus, including the second term would improve the accuracy of the TMM-based point-source model.

![Figure 5. Pressure on the cylindrical scanning surface for the fundamental of the closed wooden pipe. (a) TMM-based point-source model simulation. (b) BEM simulation using only the first term of the KHIE. (b) BEM simulation using the full KHIE. (c) Measured pressure.](image-url)
6. CONCLUSIONS

This work explored limitations in modeling the directivities of organ pipes using equivalent point-sources combined with the TMM. The results show that while there is excellent agreement between simulations and measurements at low frequencies, significant deviations occur at higher frequencies. Near-field cylindrical scans and BEM simulations reveal that neglecting diffraction effects around the pipe can cause these deviations. Future work could include applying the method to more complex sources such as woodwind instruments and developing simplified approaches to model diffraction effects.

ACKNOWLEDGEMENTS

The authors express appreciation for funding from the William James and Charlene Fuhriman Strong Family Musical Acoustics Endowed Fellowship.

REFERENCES

Numerical simulation of the acoustic field around a violin

Masao YOKOYAMA¹; Yuya ISHIGAKI¹; Amane TAKEI²; Genki YAGAWA³

¹ Meisei University, Japan
² University of Miyazaki, Japan
³ Toyo University and University of Tokyo, Japan

ABSTRACT
We numerically simulated the sound pressure field around the violin body using the finite-element method. Although there are many studies on the vibration and sound radiation of violins, obtaining results with a neat interpretation from a quantitative and qualitative standpoint remains a challenge. Herein, the highly precise geometry of the violin made by Stradivari was scanned with a micro-computed tomography scanner, where the material of the violin body was wood (spruce and maple) with orthotropic properties. This study aimed to clarify the relationship between the properties of wood and acoustic radiation around a violin vibrated by a force at the bridge.

Keywords: Sound, Insulation, Transmission

1. INTRODUCTION
We conducted numerical simulations of the vibration and sound radiation of violins made by old masters, such as the Antonio Stradivari and Guarneri families. Analyses using numerical simulations have been actively promoted since the beginning of the 21st century.¹ Computed tomography (CT) scanners have also been used to observe the internal structure of violins ² ³ and to investigate the material properties of wood.⁴ ⁵ Furthermore, methods using numerical simulation and/or CT scanners have been investigated as non-invasive and non-destructive methods for the analysis of historical assets such as violins.⁶ Previously, we conducted a coupled numerical simulation of violins using the finite-element method (FEM) and modeled the effects of wood properties on the vibration mode of the violin body.⁷ This paper reports the measurement of the geometric data of a violin made by Stradivari using a micro-CT scanner, the relationship between the mode vibration and properties of wood, and the results of numerical simulations coupling the vibrations by a forced oscillation on a bridge with the acoustic field pressure around the violin.

2. GEOMETRIC DATA AND NUMERICAL SIMULATION

2.1 Three-dimensional geometry using a micro-CT scanner
The geometry of the violin made by Antonio Stradivari (1719) was scanned from the tailpiece to the scroll of the neck using a micro-CT scanner with a precision of 0.1 mm, as illustrated on the left side of Figure 1. The image on the right side of Figure 1 is a snapshot visualizing the interior of the violin body using scanned data.

¹ masao.yokoyama@meisei-u.c.jp
geometric data with computer-assisted design (CAD) software. As the scanned raw-geometric data include many fragments and holes, we must clean the data using CAD software before numerical simulation and visualization.

The import of geometric data, meshing, and FEM calculations were conducted using COMSOL Multiphysics™. The geometric data were divided into various parts, such as top and back plates, ribs, sound posts, and bass bars, to set different physical properties for each of them (Figure 2), and then saved as STEP files (standard for the exchange of product model data). The STEP files were imported into COMSOL Multiphysics as geometric objects. In addition, we set the spherical area of air surrounding the violin (Figure 2, right side). Points A–D represent the positions where the sound pressure was calculated in Section 3.2. In COMSOL Multiphysics, the mesh generator discretizes the domains into tetrahedral second-order mesh elements using the free-mesh method. In total, there are approximately two million elements, including violins and air. The eigenmode frequency, displacement of the body, and sound pressure were calculated using the FEM via the acoustic-structure interaction module of COMSOL Multiphysics.

Figure 1 – View using the micro-computed tomography scanner and visualization of the interior of the violin using computer-aided design software

Figure 2 – Meshes of numerical simulation using the auto-mesh function of COMSOL Multiphysics. Points A–D is the position where the sound pressure is calculated in Section 3.2.
2.2 Parameter setting of the numerical simulation

The physical characteristics of wood can be set in three orthogonal directions in COMSOL Multiphysics: the longitudinal-grain direction (x-axis), radial annual-ring direction (y-axis), and direction tangential to the annual ring (z-axis). The typical values of the physical properties, such as Young’s modulus, rigidity modulus, and Poisson’s ratio, have been measured by Green et al. The representative values of the density for maple and spruce are 0.63 and 0.36, respectively, and Young’s moduli, Ex, are 12.6 GPa and 9.9 GPa, respectively. Ex/Ey, Ez/Ex, and the rigidity modulus (G/Ex) are the ratios of the longitudinal Young’s modulus (Ex) (Table 1).

<table>
<thead>
<tr>
<th>Property</th>
<th>Maple</th>
<th>Spruce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young’s modulus EY/Ex</td>
<td>0.132</td>
<td>0.078</td>
</tr>
<tr>
<td>Ez/Ex</td>
<td>0.065</td>
<td>0.043</td>
</tr>
<tr>
<td>Rigidity modulus Gxy/Ex</td>
<td>0.111</td>
<td>0.064</td>
</tr>
<tr>
<td>Gyz/Ex</td>
<td>0.021</td>
<td>0.003</td>
</tr>
<tr>
<td>Gxz/Ex</td>
<td>0.063</td>
<td>0.061</td>
</tr>
<tr>
<td>Poisson’s ratio μxy</td>
<td>0.424</td>
<td>0.372</td>
</tr>
<tr>
<td>μyz</td>
<td>0.774</td>
<td>0.435</td>
</tr>
<tr>
<td>μxz</td>
<td>0.476</td>
<td>0.467</td>
</tr>
</tbody>
</table>

However, these values differed from the actual scanned values of the violins. We cannot disassemble and measure the properties of a cultural asset such as the Stradivari violin. Thus, we estimated the values for the numerical simulation using the frequency response function (FRF) of the scanned violin, as described in the next section.

3. SIMULATION RESULTS

3.1 Mode vibration and estimation of values of orthotropic properties

In this study, we estimated the orthotropic properties of wood to be set in the numerical simulation by comparing important eigenmode frequencies obtained by numerical simulation with those measured from the real Stradivari violin (1719) by FRF. In other words, this is an inverse problem where we estimate the values of the real properties of the wood by changing parameters such as density and Young’s modulus in the numerical simulation and matching the eigenmode frequencies of the numerical simulation with those of the FRF.

The eigenmodes at low frequencies under 1,000 Hz, called A0, B1−, and B1+, were investigated in the vibration analysis of violins. Figure 3 shows the FRF of the violin proposed by Stradivari (1719). Using a miniature hammer, we tapped the sides of the bridge on both the E and G strings and measured the results using an FFT analyzer. The peak on the left side of the graph (265 Hz) corresponds to the first air cavity mode, called A0. The next peaks were B1− (428 Hz) and B1+ (511 Hz), which were purported to influence the timbre of the sound.

Figure 4 shows the numerical simulation of the changes in the eigenfrequency owing to the changes in density...
and $E_X$. The middle value in each graph represents the frequency calculated using the representative values. This graph depicts eigenfrequencies with varying densities and $E_X$ ranging from $-20\%$ to $+20\%$. The free vibrations of the violin body were calculated without any constraint points.

Figure 4 shows that the eigenfrequency increased as $E_X$ increased and decreased as the density increased. As shown in Table 2, the eigenfrequencies obtained by the numerical simulation using representative values were lower than those obtained by the FRF of the scanned violin. For instance, if we decrease the density of both maple and spruce by $10\%$ and increase $E_X$ by $10\%$ (also $E_Y$, $E_Z$, and $G$), the eigenfrequencies obtained by the numerical simulation approach those obtained by the FRF.

![Graph showing frequency response function.](image)

Figure 3 – Frequency response function of Stradivari (1719) from the hammering test.

![Graphs showing eigenfrequencies.](image)

Figure 4 – Eigenfrequencies in the A0, B1−, and B1+ modes. The eigenfrequency increases with Young’s modulus and decreases with the increasing density.
Table 2 – Comparison of the eigenfrequency (Hz): with the frequency response function (FRF) of Stradivari, numerical simulation with the representative value of the wood property⁹, and with a decrease in density of 10% and increase in Ex of 10%.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Experiment</th>
<th>Numerical simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FRF (Strad. 1719)</td>
<td>Representative⁹</td>
</tr>
<tr>
<td>A0</td>
<td>265</td>
<td>258</td>
</tr>
<tr>
<td>B1−</td>
<td>428</td>
<td>399</td>
</tr>
<tr>
<td>B1+</td>
<td>511</td>
<td>475</td>
</tr>
</tbody>
</table>

Figure 5 shows the displacement in the z-axis direction of the top/back plate in A0, B1 −, and B1+, which was calculated using the modified density and Eₓ, as mentioned above. The red area indicates where the displacement in the z-axis direction is positive, and the blue area indicates where it is negative.

In the A0 mode, the top plate bends widely on the bass bar side (left side of the violin in Figure 6 (a)). In the B1 mode, the top plate bends in the horizontal direction, and the back plate bends in the vertical direction. In contrast, the top plate bent in the vertical direction and the back plate bent in the horizontal direction in the B1+ mode. The displacement of the violin becomes asymmetrical because there is a sound post and bass bar inside the violin body. Other studies have found similar displacement patterns.²

Figure 6 shows the acoustic pressure fields around the violin in the numerical simulations. The radiation above the f-hole of the top plate was significant. The radiation under the back plate was also significant in the B1 mode. The sound radiation from the body appears to depend significantly on the bass bar and bend in the long-side direction (upper- and lower-bout directions).

Figure 5 – Displacement in A0, CBR, B1−, B1+ mode.
3.2 Acoustic pressure field around the violin in forced vibrations on the bridge

This section presents the simulation results when a forced sinusoidal oscillation is applied to the bridge of the violin. Figure 7 depicts the displacement of the body in the z-axis direction by forced vibrations. The right side of Figure 7 shows the oscillating position of the G string. The frequency of the sinusoidal function at 196 Hz (G3, the fundamental frequency of the G string) was inputted in the y-axis direction. On the left-hand side of Figure 7, the red area indicates where the displacement is positive and the blue area indicates where the displacement is negative.

The bridge alternately oscillated from side to side in the y-axis direction, and alternate vibrations of the top plate were induced by the bridge oscillation. In particular, the magnitude of the displacement on the bass bar side (the left side of the violin in Figure 7) was significant. We also speculated that the vibrations of the scroll and fingerboard do not significantly influence the sound volume and timbre; however, these vibrations are not negligible (see the video of detailed simulation results, https://youtu.be/m3cgtTJs9Q).

Figure 7 – Displacement of the violin by the forced vibrations on the bridge where the G string is placed (196 Hz). The bass bar side has significant vibrations. The scroll and fingerboard also vibrate.
The acoustic pressure fields in the y–z plane (including the bridge and sound post) and in the x–y plane (30 mm above the arch of the top plate) are shown in Figures 8 and 9, respectively. We simulated the vibration of the top plate by the sinusoidal oscillation of the bridge and concentric sound radiation from the f-hole and C-bouts. Similar to the experimental results obtained by Wang,\textsuperscript{10} sounds at a low pitch radiated concentrically from the violin body. By contrast, the directivity of the sound was observed at a higher pitch. In the near future, we will numerically analyze the characteristics of sound radiation by varying the frequency and position of forced vibrations.

![Image of sound pressure field](image)

**Figure 8** – Sound pressure field in the y–z plane in the forced vibrations with a sine wave of 196 Hz at the G string position on the bridge.

![Image of sound pressure fields](image)

**Figure 9** – Sound pressure fields around the body of the violin in the x–y plane 30 mm above the top of the arch.

Figure 10 shows the temporal change in acoustic pressure around the violin. Each line depicts the transition of the acoustic pressure at points A–D, as shown in Figure 2, during approximately two cycles of the sinusoidal function from the start of the forced oscillation ($t = 0.0$–0.01 s). The dotted line in the graph indicates the sinusoidal function of forced vibration on the bridge. The acoustic pressure at each point was similarly varied. Thus, we can also infer from the graph that the sound radiates concentrically when the bridge vibrates at 196 Hz with an open G string.

For instance, we can numerically simulate the directivity of sound radiation for the pitch difference. This
analysis of the acoustic field around the violin can be conducted experimentally in an anechoic chamber using array microphones. However, there are only a limited number of laboratories and facilities with anechoic chambers, and the cost is high. Thus, if we can substitute the numerical simulation for the experiment on sound radiation, coupled simulation can become an extremely useful tool.

4. CONCLUSIONS

A coupled simulation of the mode vibrations and sound radiation of a violin scanned by a micro-CT scanner was conducted using the FEM. We analyzed the relationship between the properties of wood and the eigenfrequency and visualized the vibration of the violin body and sound radiation in mode vibrations. Furthermore, we demonstrated that coupled simulation can analyze sound radiation. In future work, we intend to conduct a numerical simulation of sound radiation caused by string oscillations and analysis of sound radiation in a concert hall using a large-scale parallel computer.

ACKNOWLEDGEMENTS

The authors would like to thank Bunkyo-Gakki Co. Ltd. for providing violins. This study was supported by JSPS KAKENHI (grant number 21K11969).

REFERENCES

Combination of deep-learning-based audio separation and speech enhancement for noise reduction of extracted signal from polyphonic music

Soichiro KOBAYASHI(1), Takashi NOSE(2), Akinori ITO(3)

(1) Faculty of Engineering, Tohoku University, Japan
(2) Graduate School of Engineering, Tohoku University, nose@tohoku.ac.jp
(3) Graduate School of Engineering, Tohoku University, aito.spcom@tohoku.ac.jp

ABSTRACT
The music source separation technique has been dramatically improved with the development of deep learning, but the quality of the separated signal is still to be improved. In this work, we cascaded music source separation and enhancement models to suppress the noise included in the separated signal. We used Demucs, a time-domain music separation model, as the separation model. Demucs is a decoder-encoder model that takes an ensemble music waveform as an input and converts the waveform into the waveform of the desired part using the 1D-CNN and GLU. In addition, we developed a U-net-based denoising autoencoder that worked in the time-frequency domain as the enhancement model. As an experiment, we used mixed music signals in MUSDB18 and separated them into four parts: Drums, Bass, Vocals, and Other. We used the signal-to-distortion ratio (SDR) as an objective metric. The average SDR of the separated four signals using Demucs was 6.08 dB, while connecting the denoising autoencoder gave an SDR of 6.26 dB. Furthermore, we jointly re-trained the connected models, which gave the SDR of 6.28 dB. This result showed that cascading a signal separation model and a signal enhancement model improved the quality of the separated signal.

Keywords: Music signal separation, Speech enhancement, Deep learning, Demucs

1 INTRODUCTION
Sound source separation is a technology to separate a specific sound from a mixed acoustic signal [1]. There are two kinds of source separation techniques: one is the single-channel method [2], and the other is the multi-channel method [3]. In this paper, we focus on the single-channel method and aim to apply the method to music signal separation, where musical sound signals are separated into individual parts such as drums, bass, and vocals.

Speech enhancement is a technique applied to noisy speech signals to emphasize speech sound and suppress noise [4]. The speech enhancement can be viewed as a kind of sound source separation, separating the sound into speech and other sounds.

Until less than a decade ago, source separation and speech enhancement techniques were based on mathematical signal modeling, such as the non-negative matrix factorization (NMF) [2] or the independent low-rank matrix analysis (ILRMA) [5]. Recently, deep learning techniques have been extensively studied [6, 7, 8]. However, it is still impossible to separate each part completely, even with the deep-learning-based sound source separation model. The idea proposed in this paper is to suppress the leakage of the other part using a speech enhancement model as post-processing. According to this idea, we verify that the proposed method improves the quality of separated music sound signals by combining sound source separation and speech enhancement models.
2 RELATED WORKS

2.1 Music source separation

The music source separation is a kind of sound source separation [9, 10]. Music source separation models based on deep learning are roughly classified into time-frequency-domain and time-domain models. The time-frequency domain models, such as Open-Unmix [11] and D3Net [12], receive an amplitude spectrogram obtained by short-time Fourier transformation (STFT) of the original signal and output amplitude spectrograms of individual parts. The time-frequency domain models are good at separating time-varying sounds, such as vocal signals [13].

On the other hand, time-domain separation models such as Wave-U-Net [14] and Demucs [13] input the original signal’s waveform into the model, and the waveform of each part is obtained as the output of the model. Since these models take the waveform as the input and output, it does not need STFT, and the phase information is incorporated into the model automatically. It is said that this kind of model is good at separating sounds with strong attacks, such as drums and bases [13].

Demucs is a model with a U-Net [15] structure for one-dimensional convolution connecting six encoder and decoder pairs. The encoder performs convolution on the input signal’s waveform, and one convolution layer doubles the number of channels. Similarly, the decoder’s transposed convolution layer halves the number of channels, then outputs the specific part’s waveform. In addition to the model structure, the masking structure using the Gated linear units (GLU) is adopted, and a two-layer Bidirectional LSTM (BiLSTM) is used for the bottleneck part. As a result, Demucs gives a high-quality separation in the time domain.

The above-mentioned methods such as Open-Unmix and Demucs only consider a specific part when extracting the signal of that part. On the other hand, KUIELab-MDX-Net [16] introduces the “Mixer” module that exploits information of instruments of other parts when extracting the signal of a specific instrument. It is one of the state-of-the-art methods in the competition for music source separation [17] in 2021.

This model is a combination of the time-domain and time-frequency-domain networks. The time-domain part is the same as Demucs, and the time-frequency-domain part combines U-Net-based signal extractors and the Mixer module. Figure 2 shows the network structure of the time-frequency-domain part of KUIELab-MDX-Net. In the module shown in Figure 2, separation into each part is performed by a separate model, and then the separated signal is enhanced using a network called Mixer. The Mixer module is the signal enhancement mechanism that emphasizes a signal of a specific part, exploiting the information of other parts and the original signal.
2.2 Speech enhancement using deep learning
For speech enhancement, which is a kind of sound source separation, the models that use deep learning are classified into two types: one is a time-frequency domain model, and the other is a time-domain model. An example of speech enhancement in the time-frequency domain is speech enhancement using a mask. First, an amplitude spectrogram mixed with noise is input to the model, and the model estimates which part of the input is the noise to be removed. Then, a mask of the same size as the input is outputted, acting as a filter to hide the noise. The noise can be removed from the amplitude spectrogram by multiplying the input by this mask.

Macartney and Weyde [18] proposed a deep-learning-based speech enhancement model that employed the U-Net structure, similar to the Wave-U-Net. Since then, a number of U-Net-based speech enhancement models have been proposed so far [19, 20].

Recently, it has been proven that the model’s performance can be improved by combining the time-frequency domain and time domain, including speech enhancement, such as Multi-Domain Processing via Hybrid Denoising Networks (MDPhD)[21]. This model connects the sub-networks of both domains, and the network is trained considering the features of both domains. Another example is the Hybrid Demucs [22], which adds waveforms separated in both domains.

3 PROPOSED MODEL
The idea of the proposed method is to combine the audio source separation method (Demucs) and other methods such as speech enhancement or music separation methods to enhance the result of the first separation.

We examined two models. The first is the cascade model, which combines Demucs and the speech enhancement models. We intended to verify whether post-processing could improve the quality of the separated signal. The second one is a combination of Demucs and Mixer part of the KUIE-MDX-Net. Using this model, we examined how using other parts improved the quality of the specific part’s signal.

3.1 The cascade model
In this model, the separation model (Demucs) and the enhancement model are trained independently, and the separated signal is input to the enhancement model to reduce noise. Figure 3 shows the network structure of the entire model. As shown in the figure, this model first separates the input waveform into individual parts using Demucs, and the waveform of each part is enhanced using an enhancement model trained part by part.

Figure 4 shows the model structure of the speech enhancement model. It has a U-Net structure, with an encoder and a decoder with five two-dimensional convolution layers. This network calculates the time-frequency mask, and the estimated mask is multiplied by the original spectrogram to reduce noise [20].

The enhancement models were prepared part by part. Let the amplitude spectrograms separated into four parts (Drums, Bass, Other, and Vocals) by Demucs be \( A_p(k, t) \), where \( 1 \leq p \leq 4 \) is the part number. Then we calculate the log spectrogram as \( \log(1 + A_p(k, t)) \), and this log spectrogram is input to the enhancement model. The enhancement model calculates the binary mask \( M_p(k, t) \). We trained the model so that \( \log(1 + A_p(k, t))M_p(k, t) \) becomes similar to the reference signal.

The detailed specification of the model is shown in Table 1. We used the Leaky ReLU as the activation function, and we used the sigmoid function as the activation of the output layer of the decoder so that the output values of the network (the time-frequency mask) were between 0 and 1. When we transformed the amplitude spectrogram into a waveform, we used the phase of the original signal since the amplitude spectrogram lacked the phase information.

3.2 The mixer-combined model
The second model combines Demucs and Mixer in KUIELab-MDX-Net. Using this model, we investigate how the use of other parts improves a specific part’s speech enhancement model. Figure 5 shows the structure of the second model. Its basic structure is a combination of Demucs and Mixer; the difference is that the original Mixer (as shown in Figure 2) receives the time-domain signal of each part and the original waveform, whereas the proposed model does that in the time-frequency domain. We employed this structure aiming to combine the time-domain-based and time-frequency-domain-based models to compensate for the weak points of those models.
Figure 3. Structure of the cascade model

Figure 4. Structure of the speech enhancement model

[21]. Figure 6 shows the structure of the mixer network, which is the same as that of the enhancement network (Fig. 4), except that it has five input channels (four parts and the original signal) and four output channels (four parts).

When training this model, we first trained Demucs using the training data, and then we trained the entire model. We used the loss function combining two kinds of losses, the loss after separation ($L_{\text{wave1}}$) and after emphasis ($L_{\text{wave2}}$), and the total loss was the sum of these loss functions.

$$L = L_{\text{wave1}} + L_{\text{wave2}}$$

We used the weighted-SDR (wSDR) loss [23] as a loss function.

4 EXPERIMENT

4.1 Materials

We used MUSDB18 corpus [24] for training and testing the models. We used 86 songs for training, 14 for development, and 50 for testing. Five signals, Mixture (the original signal), Drums, Bass, Other, and Vocals, are prepared for one song. The sampling frequency of the files is 44100 Hz, and the quantization is 16 bit/sample. When training a model, we extracted 11-second segments by shifting 1 second and used them as individual data. In addition, we augmented the training data with the following augmentation methods:

- Randomly mixing instruments from different songs [25]
- Randomly swap left and right channels of instruments [25]
- Randomly scale the signal from 0.25 to 1.25 [25]
- Randomly change the phase by multiplying $-1$ [26]
- Random pitch shift from $-2$ to 2 halftones [27]
- Random tempo change from 0.88 to 1.12 [27]

Signals in MUSDB18 have two channels. We input the stereo signal to Demucs part of the model, then we split two channels and input them individually to the enhancement model.
Table 1. Network conditions of the cascade model

<table>
<thead>
<tr>
<th></th>
<th>Cascade model</th>
<th>Mixer-combined model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kernel size and stride</strong></td>
<td>(21, 5), (2, 2)</td>
<td></td>
</tr>
<tr>
<td><strong>Channel size of the encoder</strong></td>
<td>1 – 45 – 90 – 90 – 90 – 90</td>
<td>5 – 45 – 90 – 90 – 90 – 90</td>
</tr>
<tr>
<td><strong>Channel size of the decoder</strong></td>
<td>90 – 90 – 90 – 45 – 1</td>
<td>90 – 90 – 90 – 90 – 45 – 4</td>
</tr>
<tr>
<td><strong>STFT Frame size and shift</strong></td>
<td></td>
<td>4096, 1024</td>
</tr>
<tr>
<td><strong>Window function</strong></td>
<td>Hann window</td>
<td></td>
</tr>
<tr>
<td><strong>Batch size</strong></td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td><strong>Maximum number of epochs</strong></td>
<td>50</td>
<td></td>
</tr>
<tr>
<td><strong>Loss function</strong></td>
<td>MSE wSDR</td>
<td></td>
</tr>
<tr>
<td><strong>Activation</strong></td>
<td>Leaky ReLU / Sigmoid (the last layer)</td>
<td></td>
</tr>
<tr>
<td><strong>Optimizer and learning rate</strong></td>
<td>Adam, $2 \times 10^{-4}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The separation results evaluated by SDR [dB]

<table>
<thead>
<tr>
<th></th>
<th>Drums</th>
<th>Bass</th>
<th>Other</th>
<th>Vocals</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRM</td>
<td>10.04</td>
<td>9.33</td>
<td>7.78</td>
<td>10.36</td>
<td>9.38</td>
</tr>
<tr>
<td>Demucs</td>
<td>6.59</td>
<td>7.04</td>
<td>4.19</td>
<td>6.48</td>
<td>6.08</td>
</tr>
<tr>
<td>Cascade model</td>
<td><strong>6.63</strong></td>
<td>7.11</td>
<td>4.40</td>
<td>6.88</td>
<td>6.26</td>
</tr>
<tr>
<td>Mixer-combined model</td>
<td>6.49</td>
<td><strong>7.22</strong></td>
<td><strong>4.48</strong></td>
<td>6.92</td>
<td><strong>6.28</strong></td>
</tr>
</tbody>
</table>

4.2 Result
Table 2 shows the summary of the objective evaluation for 50 pieces. We used the signal-to-distortion ratio (SDR) [28], representing the ratio of other sounds. The IRM (Ideal Ratio Mask) in the table represents the result when we use the ratio of the clean and noisy spectrograms as the time-frequency mask. This result is an upper limit of the separation quality when using the mask-based speech enhancement.

When we compare Demucs and the cascade model, we can confirm that the enhancement model improves the SDR for all parts. Especially, “Other” and “Vocals” showed larger improvement than “Drums” and “Bass.”

The use of the mixer-combined model gave more improvement for three parts except for Drums. The SDR of the Drums was slightly lower than that of the Demucs separation.

4.3 Examples
Figure 7 shows examples of spectrograms for vocal signal separation. Figure 7 (a) is the reference signal. Figure 7 (b) is the output from Demucs. The circled part in (b) has a straight striped pattern derived from the sound of Other part (guitar). Comparing the figures (a) and (b), the separation result by Demucs is insufficient to suppress the instrumental sound. The instrumental sound is suppressed in the results (c) and is almost removed in the result (d).

Figure 8 shows the examples for drum signals. We can see the horizontal line-like patterns in the parts surrounded by the ellipses derived from the other instrumental sound. This noise is suppressed in (c) and (d). On the other hand, the part surrounded by the dashed circle is the noise from vocal signal. We can see that this
noise is suppressed in (c). However, in (d), the harmonic pattern of the vocal signal is over-suppressed, which causes additional distortion of the separated signal. This distortion is audible as a kind of noise.

5 CONCLUSION
In this paper, we propose a method to improve music source separation performance using speech enhancement. We used the time-domain audio separation method Demucs as a baseline model. We introduced two models; one is the cascade model, where a simple U-Net-based model is used that suppresses noise part by part. Another is the mixer-combined model, in which all parts are considered simultaneously for enhancement.

We conducted an experiment to compare Demucs and the proposed methods, and the result showed the proposed methods improved the SDR over the baseline method. In addition, the mixer network was better than the part-by-part enhancement on average, but it over-suppressed the vocal signal, introducing additional distortion.

In future work, we need to compare the proposed methods to the existing methods that combine the time-domain and time-frequency domain signals, such as Hybrid Demucs. In addition, we plan to incorporate the adversarial generative network technology [29, 30] for further improvement.

ACKNOWLEDGMENT
Part of this work was supported by JSPS KAKENHI Grant Number JP21H00895.

References


Figure 7. An example of enhancement result (Vocals)


Figure 8. An example of enhancement result (Drums)


Automatic classification of blowing properness in flute sounds*

Sai OSHITA(1), Tetsuro KITAHARA(1)

(1)Nihon University, Japan, {oshita,kitahara}@kthlab.jp

ABSTRACT
Our goal is to develop a system that evaluates the technique of novice flute players when playing their instrument. This paper addresses automatic evaluation of flute sounds. We collected flute sounds played with a proper and improper blowing styles. Using the sounds, a subjective evaluation with a five-graded scale were shown to participants. Then, automatic evaluation of those sounds (that is, the prediction of the degree of the appropriateness and the classification of appropriate and inappropriate sounds) was conducted with linear regression and decision trees using results of the subjective evaluation. The results showed that (1) the root mean squared error of the regression-based prediction was 0.65, and (2) the accuracy of the decision-tree-based classification was 93%. On the other hand, the accuracy of the classification of blowing styles using the decision tree was lower (about 70% for two-class (normal/non-normal) classification, 20–40% for five-class classification).

Keywords: flute, wind instrument, acoustic analysis, automatic evaluation

1 INTRODUCTION
The flute is an instrument whose sound changes greatly when the direction of the breath and the size of the mouth are changed. Therefore, it is necessary to carefully control the size of the mouth and the direction and strength of the breath in order to play the flute with appropriate tone quality. However, although there are many books on the market instructing how to play the flute, not many of them clearly describe these points. Even if one reads a detailed book, it is not easy for a beginner to listen to his/her own sound and judge how to improve it by themselves.

Various studies have been conducted on the acoustic analysis of flutes. Ando[1, 2, 3, 4] analyzed the differences in the range of sound and the number of overtones in different registers when the blowing angle and breath strength are changed, as well as the differences among multiple flutes. Onogi et al. [5] analyzed how to make the machine performance more like human performance. Nishimura et al. [7] analyzed the differences in tone preferences between professional and amateur flutists. Yorita et al. [9] and Fletcher [10] analyzed how air pressure, lip shape, and overtones were related to years of experience. These were mainly acoustic analyses and did not map the relationship with the blowing style.

On the other hand, Wilcocks [6] presented practice routines including methods of controlling the tone of the flute. Without taking into account whether or not good sounds were being produced, the paper focused on methods for students to find the proper way to blow by trial-and-error.

Yoonchang Han[8] had created a system to judge whether the player was playing appropriate sounds by judging the head tube relationship, air pressure, and fingering from the sound, but not whether the sound was good or not.

Therefore, our goal is to develop a system that informs novice flute players whether they are producing appropriate sounds. For this aim, we first collected audio samples of flute tones with proper and improper blowing styles. Then, we extracted acoustics features from those, and collected participant responses to the tone qualities. Using those data, we attempted automatic classification of proper and improper tones using linear regression and decision trees.

*This work was supported by JSPS Kakenhi Nos. 22H03711 and 21H03572.
2 METHOD

2.1 Used Data
In order to collect the sounds of the improperly played music, we collected the performance sounds on a crowdsourcing site. They were asked to play the score shown in Figure 1. and we instructed the performers not to play vibrato. In order to reduce the burden on individual performers, we divided the collection into the following two pattern.

1 [normal] [larger mouth] [smaller mouth] [breath upward]
2 [normal] [breath downward] [breath stronger] [breath weaker]

The number of performers and the number of performance sounds collected are shown in Table 1. However, since it is difficult to distinguish [breath stronger] and [breath weaker] from forte and piano as a musical expression, we decided not to use them in the following analyses. In order to compensate for the fact that the sound volume varies depending on the recording conditions, we corrected the amplitudes so that the temporal mean values of the amplitudes are equal.

![Figure 1. Note performed by participants](image)

Table 1. Participants (performers) for flute sound collection

<table>
<thead>
<tr>
<th>Pattern</th>
<th># of performers</th>
<th>flute experience (years)</th>
<th># of sounds per performer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern 1</td>
<td>12</td>
<td>0.1~10</td>
<td>about 76</td>
</tr>
<tr>
<td>Pattern 2</td>
<td>12</td>
<td>0.1~10</td>
<td>about 88</td>
</tr>
</tbody>
</table>

2.2 Subjective Evaluation
Using a Web-based crowdsourcing service, subjective evaluations of the collected sounds were conducted. Participants were limited to current or former students of flute majors in music colleges or high school music departments, and those who have played the flute for at least 12 months. As a result, three participants from Table 2 participated.

When a participant opened the designated web page, 20 randomly selected performance sounds were displayed. They listened to them one by one and entered their answers to the questions in Table 3. The choices for choice-type questions are Table 4 and Table 5.

2.3 Acoustic Analysis Methods
2.3.1 Silence interval removal
In order to remove the sections before and after blowing from the collected acoustic signals, the following process is performed. First, the root mean square (RMS) is calculated for each of the 512 points for a given acoustic signal. Let $a_1a_2\cdots a_T$ be the sequence of the RMS. The maximum value of RMS is $a_{\text{max}}$. Then, we find the smallest value from $n_S$ satisfying $\forall n \in [n_S, n_S + \alpha]: a_n > a_{\text{max}}/4 \ (n = 1, 2, \cdots T)$, which is regarded as the onset time of the sound. Similarly, we find the largest value from $n_E$ satisfying $\forall n \in [n_E - \alpha, n_E]: a_n > a_{\text{max}}/4 \ (n = 1, 2, \cdots T)$, which is regarded as the offset time of the sound.
Table 2. Participants (evaluators) for subjective sound evaluation

<table>
<thead>
<tr>
<th>Participant</th>
<th># of participation</th>
<th>Flute experience (years)</th>
<th>Other instrument experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
<td>Sax, piano</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>5</td>
<td>Piano, harp</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>6</td>
<td>Electric organ, percussion, etc.</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>22</td>
<td>Piano, percussion, piccolo, etc.</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>28</td>
<td>Piano</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>28</td>
<td>Piano</td>
</tr>
</tbody>
</table>

Table 3. Questions used in the subjective evaluation

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
<th>Response Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>How good did it sound?</td>
<td>option 1</td>
</tr>
<tr>
<td>2</td>
<td>how clear was it</td>
<td>option 1</td>
</tr>
<tr>
<td>3</td>
<td>How small was the volume swing?</td>
<td>Response type 1</td>
</tr>
<tr>
<td>4</td>
<td>How little was the pitch swing?</td>
<td>Response type 1</td>
</tr>
<tr>
<td>5</td>
<td>How much breath did you hear?</td>
<td>Response type 1</td>
</tr>
<tr>
<td>6</td>
<td>Were there applicable problems with the blowing method?</td>
<td>Response type 2</td>
</tr>
<tr>
<td>7</td>
<td>Things I noticed other than the above</td>
<td>description</td>
</tr>
</tbody>
</table>

Then, the acoustic signal from $n_S + \beta$ to $n_E - \beta$ is extracted and used in the following analysis. The acoustic signal extracted from the interval from $n_S + \beta$ to $n_E - \beta$ is used for the subsequent analysis. The reason for narrowing the interval to be extracted from $[n_S,n_E]$ by only beta is to avoid the situation where the fundamental frequencies in the vicinity cannot be extracted due to calculation errors at the start and end points. We set $\alpha = T/19$ and $\beta = (n_E - n_S)/50$.

2.3.2 Features related to temporal variation of amplitude and fundamental frequency

It is important for beginners to be able to produce sounds with little variation in amplitude and fundamental frequency with time, so we focus on acoustic features that represent temporal variation of amplitude and fundamental frequency.

First, the RMS and fundamental frequency (F0) are obtained for each of the 512 points of the acoustic signal after removing the silence interval. Let $a_1a_2\cdots a_T$ and $f_1f_2\cdots f_T$ be the sequence of the RMS and F0, respectively. To extract F0, we used PyWorld, a speech analysis library for Python. From $\{f_n\}$, we excluded the frames whose F0 was not found. Then, the following four features are extracted.

- Average of time variation of amplitude: $d_v = \frac{1}{T} \sum_{n=2}^{T} |a_n - a_{n-1}|$
- Average of time variation of fundamental frequency: $d_f = \frac{1}{T} \sum_{n=2}^{T} |f_n - f_{n-1}|$
- Amplitude range: $r_v = \max_{n} a_n - \min_{n} a_n$
- Fundamental frequency range: the quartile range $f_r$ of $f_1f_2\cdots f_T$. The silent frames were excluded because the F0 calculated from them may be $-\infty$ or NaN in PyWorld.

2.3.3 Features related to the amplitude spectrum

According to Ando’s analysis [1], the better the sound is, the more higher harmonics there are. In addition, the more breath sounds are included, the more they appear as non-harmonic components. Therefore, the following features are extracted.

- Number of harmonic components (including fundamental frequency components) at the beginning of blowing:
From the first 1-sec interval (after removing the silence interval), the amplitude spectrum is obtained, and is denoted by $S[f]$ ($f$: frequency). Let $F_0$ be the fundamental frequency. Then, The amplitude of the $k$-th harmonic, $s_k$, is calculated as $s_k = \max\{S[kF_0-r], \ldots, S[kF_0+r]\}$. The upper bound of $k$ is defined so that $(s_k + s_{k+1})/2 > 2s_{k+1}/2$ is not satisfied or $s_k < 0.01 \cdot \max_f S[f]$. Let $o$ be the number of overtones extracted. The value of $F_0$ is calculated based on the equal temperament ($C_5 \approx 523.25$[Hz]).

- Percentage of overtones (non-fundamental components) in all harmonics at the beginning of blowing: $f_s = \sum_{k=1}^{K} s_k / s_1$, where $s_1, \ldots, s_k$ are the amplitudes of the extracted harmonics.
- Percentage of overtones in the entire spectrum at the beginning of blowing: $n_s = \sum_{k=1}^{K} s_k / \sum_f S[f]$. This feature takes a lower value when more non-harmonic components such as breath sounds are included.
- Number of overtones (calculated from the middle interval),
- Percentage of non-fundamental frequency components in all overtones (calculated from the middle interval),
- Percentage of overtone components in the whole spectrum (calculated from the middle interval):

For extracting these features, the same process as above is applied to the central one second of the acoustic signal. These features are denoted by $o_c$, $f_c$, and $n_c$, respectively.

### 2.4 Automatic Evaluation

From the collected performance sounds, we constructed a linear regression and a decision tree. The average responses to the subjective evaluation were used as the objective variable and the acoustics features mentioned above were used as the explanatory variables. In addition, we conducted a decision tree experiment when the type of blowing styles were used as the objective variable. For all evaluations, half of the data were used for training the models, and the remaining were used for testing them.

#### 2.4.1 Linear regression

With linear regression, the objective variable $y$ is approximated as $y = w_0d_y + w_1r_y + w_2d_f + w_3r_f + w_4o_s + w_5o_c + w_6n_s + w_7n_c + w_8f_s + w_9f_c + \gamma$, where the values of $w_n$ and the constant ($\gamma$) were estimated. The same linear regression was applied to the data without the outliers. Accuracy was evaluated as the root mean square error (RMSE).

#### 2.4.2 Decision tree

To predict the subjective evaluation with a decision tree, we have to discretize evaluators’ responses. We attempted two different ways of discretization: two-class [lower than 2 / 2 or higher], and three-class [lower than
2 / 2 (incl.) to 3 (not incl.) / 3 or higher. For sounds that have more than one evaluation, their average was used. The same experiment was also conducted after removing outliers.

3 RESULT

3.1 Results of acoustic feature extraction and subjective evaluation

The results of acoustic feature extraction (standardized) are shown in Figure 2. This figure shows the sounds of [normal], [large mouth], [small mouth], [upward breath], and [downward breath] from left to right, and the vertical lines indicate the boundaries of those blowing styles. The figure suggests the following features do not depend clearly on the blowing styles:

- the mean value of amplitude fluctuation ($d_v$),
- the number of overtones in the first second ($o_1$),
- the number of overtones in the middle second ($o_{1/2}$),
- the percentage of non fundamental frequency components in the first second ($f_{1/2}$),
- the percentage of non fundamental frequency components in the middle second ($f_{1/2}$),
- the percentage of overtones in the first second ($n_{1/2}$), and
- overtones in the middle 1 second ($n_{1/2}$).

On the other hand, the following features have no outliers for the normal blowing style while there are some outliers for other blowing styles:

- the mean value of fundamental frequency fluctuation ($d_f$),
- the range of amplitude ($r_v$), and
- the range of fundamental frequency ($r_f$),

If such sounds with outliers of acoustic features actually produce inappropriate sounds, these acoustic features may be used to classify the appropriateness of the sounds. We therefore analyzed the sounds with outliers and their subjective evaluation (Table 6). The results show that outliers do not appear in the normal blowing style while outliers appear in all other blowing styles. Also, the sounds with outliers were evaluated low (e.g., 1 or 2 for responses to the subjective evaluation, commented like “almost airy sounds”). Above, we regarded features satisfying the following conditions as outliers: $d_f \geq 1.5$, $r_v \geq 1.0$, and $r_f \geq 1.0$.

Regarding the expected problems of blowing (Table 3, Question 6), the option of too weak breath was selected for all but one of the tones played with the breath facing upward. For all but two of the tones that were blown with the breath upward, the option that the direction of the breath was too downward was selected.

3.2 Results of acoustic feature analysis and subjective evaluation

Figure 3 shows the amplitudes, fundamental frequency, and amplitude spectra (the first 1-sec, the central 1-sec) of the sounds that were subjectively evaluated highly (sound ID: d19, downward breath) and low (sound ID: b17, loud mouth). Looking at the amplitude and fundamental frequency in Figure 3, it can be seen that sounds with higher ratings have less temporal variation. In fact, temporal-variation-related features for Sound d19 ($d_f = -0.193$, $r_v = -0.271$, $r_f = -0.274$) are smaller than those for Sound b17 ($d_f = 0.248$, $r_v = -0.167$, and $r_f = 0.252$). Note that some of these values are negative because they have been normalized so that the average among all sounds is zero. As for the amplitude spectrum, it can be seen that the higher-rated have high amplitudes in higher-order harmonics, which is consistent with the finding by Ando [1].

On the other hand, no difference in the distribution of features was observed between the different blowing styles. That means that it is difficult to identify the cause of the problem in the blowing style, even if it is possible to identify whether the blowing is appropriate or not.
Table 6. Performance sound with outlier feature and subjective evaluation of it (Blank: not rated)

<table>
<thead>
<tr>
<th>Performance method</th>
<th>sound ID</th>
<th>outlier feature value</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>large mouth</td>
<td>b10</td>
<td>$r_v$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2, 4, 6</td>
<td>Unstable and not blowing properly.</td>
</tr>
<tr>
<td></td>
<td>b11</td>
<td>$r_v$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td>The sound is muffled and not properly reproduced.</td>
</tr>
<tr>
<td></td>
<td>b12</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1, 2, 4, 6</td>
<td>Out of the question, Air noise</td>
</tr>
<tr>
<td>small mouth</td>
<td>s9</td>
<td>$r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td>just starting out, and don’t seem to understand anything. can’t play properly at all. haven’t got the knack of making notes.</td>
</tr>
<tr>
<td></td>
<td>s13</td>
<td>$r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td>still haven’t gotten the hang of how to make the sound.</td>
</tr>
<tr>
<td></td>
<td>s14</td>
<td>$r_v$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td>Broken sound</td>
</tr>
<tr>
<td></td>
<td>s15</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td>The place where the breath hits is not right. sound like a person who blew for the first time today.</td>
</tr>
<tr>
<td></td>
<td>s16</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 4, 6</td>
<td>The sound is not right. breath is not in the right place. There is no sound at all.</td>
</tr>
<tr>
<td></td>
<td>s17</td>
<td>$r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 3, 4, 6</td>
<td>Not a problem, just getting started. No sound was coming out. felt like blowing into it, not forming a mouth shape.</td>
</tr>
<tr>
<td>breath upward</td>
<td>u6</td>
<td>$r_v$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1, 2</td>
<td>The breath was too strong and the sound be interrupted in the middle. The sound was choppy and unsteady.</td>
</tr>
<tr>
<td></td>
<td>u8</td>
<td>$r_v$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 4, 6</td>
<td>just starting out and are still getting used to it.</td>
</tr>
<tr>
<td></td>
<td>u9</td>
<td>$r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 4, 6</td>
<td>The sound is not coming out properly. don’t know where to breathe and how much to breathe. You are not blowing.</td>
</tr>
<tr>
<td></td>
<td>u10</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 5, 8</td>
<td>Like someone who blew for the first time today. There is no sound at all.</td>
</tr>
<tr>
<td></td>
<td>u11</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3, 6</td>
<td>Almost air sound.</td>
</tr>
<tr>
<td></td>
<td>u12</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>u13</td>
<td>$d_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td>Sounds terrible. Like someone who blew for the first time today.</td>
</tr>
<tr>
<td>breath downward</td>
<td>d9</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td>can’t get the sound right. not hitting the right spot with breath, not getting the hang of it.</td>
</tr>
<tr>
<td></td>
<td>d13</td>
<td>$d_f, r_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d17</td>
<td>$d_f$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2, 6</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Linear regression

The root mean square error (RMSE) for the prediction of the subjective evaluation using linear regression is 0.670 (rounded to the nearest third decimal place). When the outliers are removed, the RMSE is is 0.642 (rounded to the nearest third decimal place).

The coefficients $w_0, \ldots, w_9$ and the constant $\gamma$ estimated with the linear regression, where the approximation equation is $y = w_0 d_v + w_1 r_v + w_2 d_f + w_3 r_f + w_4 o_s + w_5 o_c + w_6 n_s + w_7 n_c + w_8 f_s + w_9 f_c + \gamma$ are shown in Table 7. Comparing these coefficients with and without outliers, the weight of $r_f$ has changed significantly. This could be because $r_f$ is negative for non outliers but is positive for some outliers. There was little difference in the root mean square error (RMSE) when outliers were removed or not. Figure 4 shows a comparison between predicted and actual values of the subjective prediction. The figure shows that, even though the actual value of the highest subjective evaluation is 4.75, its predicted value is 2.38. When the outliers are removed, the sounds where actual subjective evaluation are greater than 3 have lower predicted values than the actual evaluation.
Table 7. Coefficients of linear regression equation (Rounded to the nearest third decimal place)

<table>
<thead>
<tr>
<th></th>
<th>(w_0)</th>
<th>(w_1)</th>
<th>(w_2)</th>
<th>(w_3)</th>
<th>(w_4)</th>
<th>(w_5)</th>
<th>(w_6)</th>
<th>(w_7)</th>
<th>(w_8)</th>
<th>(\gamma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>−0.202</td>
<td>−0.168</td>
<td>−0.239</td>
<td>−0.104</td>
<td>0.519</td>
<td>−0.202</td>
<td>0.015</td>
<td>0.056</td>
<td>−0.449</td>
<td>0.240</td>
</tr>
<tr>
<td>Outliers excluded</td>
<td>−0.204</td>
<td>−1.525</td>
<td>−0.606</td>
<td>8.312</td>
<td>0.219</td>
<td>0.091</td>
<td>0.009</td>
<td>−0.014</td>
<td>−0.137</td>
<td>−0.066</td>
</tr>
</tbody>
</table>

3.4 Decision tree
3.4.1 Objective variable: subjective evaluation
As described above, the prediction of subjective evaluation with decision trees were conducted as two-class and three-class classification. The results of these classification are shown in Figure 5. The depth of the decision tree is 2. Table 8 shows the classification accuracy when the depth of the decision tree is 2, which has the maximum classification accuracy among different tree depths.

Table 8. Decision trees’ depth and classification accuracy (in parentheses: outliers removed)

<table>
<thead>
<tr>
<th>Classification</th>
<th>Maximum (Depth 2)</th>
<th>Maximum Depth (Maximum time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-class (lower than 2 / 2 or higher)</td>
<td>0.93 (0.94)</td>
<td>0.93 (0.94)</td>
</tr>
<tr>
<td>Three-class (lower than 2 / 2 to 3 / 3 or higher)</td>
<td>0.83 (0.84)</td>
<td>0.86 (0.84)</td>
</tr>
</tbody>
</table>

The classification accuracy was about 80–90%, regardless of whether outliers were included or not. For top-level decision rule, \(d_f\) was often used, while only amplitude- and F0-related features such as \(d_r\), \(r_v\), and \(r_f\) were used for other-level decision rule. This is considered to be consistent with the prediction that a sound with less temporal variation in amplitude and fundamental frequency is a good sound. It also indicates that amplitude and fundamental frequency related features are important in predicting subjective evaluation.

3.4.2 Objective variable: type of blowing
We conducted different classification tasks: two-class [normal / other], three-class [normal / mouth-size-related / breath-direction-related] and five-class [normal / mouth big / mouth small / breath upward / breath downward]. The results of these tasks are shown in Figure 6, where the depth of the decision tree is 2. Table 9 lists the classification accuracy with the depth of the decision tree of 2, which has the maximum classification accuracy. On average, the accuracy was lower than 20% for the five-class classification, about 30% for the three-class classification, and about 40% for the two-class classification (even though the maximum classification accuracy is about 70%). The classification accuracy decreased when the outliers were removed from the prediction of classification.

Table 9. Decision Tree Depth and Discrimination Rate 2 (in parentheses: outlier removal)

<table>
<thead>
<tr>
<th>Classification</th>
<th>Maximum (Depth 2)</th>
<th>Maximum Depth (Maximum time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>two-class (normal, others)</td>
<td>0.7 (0.41)</td>
<td>0.71 (0.75)</td>
</tr>
<tr>
<td>three-class (normal, oral, breath)</td>
<td>0.39 (0.31)</td>
<td>0.49 (0.44)</td>
</tr>
<tr>
<td>five-class (each blowing style)</td>
<td>0.36 (0.19)</td>
<td>0.36 (0.19)</td>
</tr>
</tbody>
</table>

Features related to the fundamental frequency such as \(d_f\), \(r_f\) were often used in the top-level decision rule, while features other than \(o_r\), \(o_c\) were often used in the other-level decision rule. The \(n_s\) and \(n_c\) were used more frequently in the two-class classification than the other classifications. Compared to the classification of the subjective evaluation, features related to the amplitude spectrum were used more frequently. These imply that amplitude, fundamental frequency, and amplitude spectrum are effective for the prediction of the blowing styles.

4 CONCLUSION
The purpose of this study is to investigate whether it is possible to identify blowing styles (especially improper blowing styles) from performed sounds. Our results imply that it is possible to classify sounds with proper and
improper blowing with accuracy of 80∼90% , whereas it is difficult to identify the cause of improper blowing from acoustic features.

However, the number of responses to the subjective evaluation is still small, so the results may be affected by individual differences among the evaluators. We have to increase the number of responses to obtain more reliable analysis. In addition, we have to analyze various performances including long tones, melodies, and arpeggio, which will be important to achieve a practical performance suppoer system.

REFERENCES


[5] Onogi K, Yokoyama H, Iida A, Mechanism analysis of timbre change by blowing angle of flute (in Japanese), Supercomputing Division, 22 (Special Issue 1), 2020


Figure 2. Distribution of the feature values (standardized) for each performance note (from left to right: [Normal], [Large mouth], [Small mouth], [Breath upward], [Breath downward])
Figure 3. Amplitude, fundamental frequency, and amplitude spectrum of the performance sounds that were good/bad rated.

Figure 4. Comparison of actual subjective evaluation and its prediction with linear regression (horizontal: actual evaluation, vertical: predicted)
Two-class (lower than 2, 2 or higher), outliers removed

Three-class (lower than 2, 2–3, 3 or higher), outlier removed

Figure 5. Decision trees obtained with predicted subjective evaluations
Figure 6. Decision trees obtained for predicting blowing styles.
Effects of the temperature on the vibration duration of tuning forks

Kyota NOMIZU¹; Sho OTSUKA²,³,⁴; Seiji NAKAGAWA²,³,⁴

¹ Dept. of Medical Eng., Grad. School of Sci. & Eng., Chiba Univ., Japan
² Ctr. for Frontier Medical Eng., Chiba Univ., Japan
³ Grad. School of Eng., Chiba Univ., Japan
⁴ Med-Tech Link Ctr., Chiba Univ. Hospital, Japan

ABSTRACT

Tuning forks are used in the fields of music, clinical tests of vibrotactile sensation/audition, and healing. For the tuning fork, longer duration of the fundamental tone is required as well as accuracy of the frequency of the fundamental tone. Additionally, it is desirable that the harmonic tone decreased immediately after the stroke. However, in the manufacturing process of the tuning forks, only the fundamental frequency is tuned, and durations of tones have not been evaluated. In addition, although previous studies have revealed the effect of temperature on the frequency, the effect of temperature on durations of tones has not been investigated. In this study, we developed a tuning-fork-striking device and evaluated effects of temperature on the effective durations of the fundamental and harmonic tones of several tuning forks made of different metallic materials. As results, the effective durations of fundamental tone varied with change of the temperature and the variations tended to depend on materials, although the magnitudes of the variation were different among individual tuning forks.

Keywords: Tuning fork, Vibration, Effective duration, Fundamental, Harmonic

1. INTRODUCTION

The tuning fork is a U-shaped metal instrument that emits sounds of a specific frequency when struck. It has been used for a variety of purposes; music, medicine, healing, and others. Each fork is made manually by artisans, and the frequency is adjusted by polishing. Therefore, the production process of tuning forks are still dependent on individual skills of the artisan.

Sounds produced by a tuning fork is divided into a fundamental tone (sound at the resonant frequency) and harmonics. Immediately after a strike of a tuning fork, a harmonic called "clang" tone is produced in abundance. The fundamental vibration mode, in which the tines move symmetrically, produces the fundamental tone, and the second mode produces the clang tone that has approximately six times higher frequency of the fundamental tone (1). Considering the purpose of using a tuning fork, it is desirable that the fundamental tone persists as long as possible and the clang tone decays as quickly as possible.

In the manufacturing process of the tuning forks, the fundamental frequency is tuned by hardening and polishing, however, the vibration duration is not evaluated. Further, while there have been some studies on relationships between the temperature and the frequency (2,3) and the radiated sound field (4) of tuning forks, there have been few studies on the vibration duration. Based on the above, we have been conducting research on the vibration duration of tuning forks and evaluated individual differences and the effects of holding strength (5).

Tuning forks are made of various metal materials such as aluminum, carbon steel, and stainless steel. The frequency of tuning forks depends on the properties of the metallic materials, such as Young's modulus and density, and varies with temperature (6). Therefore, it is highly likely that the variation of the vibration duration with temperature change will also vary depending on the material. In this report, we struck several tuning forks made of different metallic materials at different temperatures and compared the variation of the vibration duration with temperature change.

All experiments were approved by the Institutional Review Board of the Life Science Research of Chiba University. Necessary information about the experiments was provided to the participants and
informed consent was obtained from each participant before the experiments.

2. EXPERIMENTS

2.1 Tuning Fork
In this experiment, we used four aluminum tuning forks with a resonance frequency of 440 Hz (Niti-On), four carbon steel tuning forks with a resonance frequency of 440 Hz (Niti-On), and four stainless steel tuning forks with a resonance frequency of 528 Hz (Niti-On 04-040). All forks are shown in Figure 1, and the length and mass of each fork are shown in Table 1.

2.2 Basic Procedure
First, the tuning forks were heated to 70 °C in a thermostatic bath. After a sufficient heating time, the tuning forks were placed in a tuning-fork-striking device in an anechoic room. Tuning forks placed at room temperature gradually decreased in temperature. A thermocouple temperature sensor (A&D AD1214) was attached to the tuning fork holder of the tuning-fork-striking device to measure the temperature. We struck the tuning fork at 2 °C intervals as the temperature decreased. Aluminum tuning forks were measured at 11 temperature levels ranging from 26 to 46 °C, carbon steel tuning forks at 12 levels ranging from 28 to 50 °C, and stainless steel tuning forks at 8 levels ranging from 26 to 40 °C. The above procedure was repeated four times for each tuning fork.

2.3 Tuning Fork Striking
The tuning fork was struck using a newly-devised tuning-fork-striking device, as shown in Figure 2. The tuning-fork-striking device consists of a tuning fork holder and a mallet driving unit with a spring. The tuning fork holder was made by using a 3D printer to match the shape to the stem of each tuning fork. The mallet strike position was 1/3 to 1/4 of the tuning fork tine length from the free end of the tine, which is considered a desirable strike position because it corresponds to the node of the second mode and thus reduces the clang tone (7,8). In this experiment, the mallet drive intensity was kept constant for all measurements.

<table>
<thead>
<tr>
<th>Material</th>
<th>Aluminum</th>
<th>Carbon steel</th>
<th>Stainless steel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>Length [cm]</td>
<td>18.2 18.3 18.3 18.3</td>
<td>16.8 16.7 16.8 16.8</td>
<td>13.3 13.4 13.4 13.2</td>
</tr>
<tr>
<td>Mass [g]</td>
<td>64 65 63 64</td>
<td>143 141 143 143</td>
<td>55 55 54 52</td>
</tr>
</tbody>
</table>

Figure 1 - Tuning forks used in the experiment.

Figure 2 – Tuning-fork-striking device.
2.4 Sound Recording

The sound produced by the tuning fork was recorded for 30 s each in an anechoic room using a 1/4-inch free-field microphone (Brüel & Kjær 4939-A-011) and a conditioning amplifier (Brüel & Kjær 2690-0S4). The microphone was placed at the same height as the fork. The distance between the microphone and the fork was kept at 10 cm. The tuning fork and microphone were placed at a sufficient distance from the inner walls of the anechoic room.

2.5 Vibration Evaluation

In sound analysis, we focused only on the fundamental and clang tones of the tuning fork. Fast Fourier Transformation (FFT) and Short-time Fourier Transformation (STFT) were applied to the obtained acoustic signal data to estimate the frequencies of the fundamental and clang tones. Next, we applied a band-pass filter to the time waveform to extract each tone. We extracted the amplitude maximums of both tones from the absolute values of the peaks observed immediately after the tuning fork was struck, and we calculated the time required for the amplitude to reach 1/10 of the maximum amplitude (we defined this as "vibration duration").

Figure 3 shows an example of vibration evaluation using the above method. The acoustic signal data was recorded as shown in Figure 3(a), the Short-time Fourier Transformation (STFT) was applied as shown in Figure 3(b), and the tone was extracted by band-pass filtering as shown in Figure 3(c) based on the frequency analysis results. Figure 3(d) shows the calculation of the vibration duration based on the absolute values of the peaks of the tone.

3. RESULTS

3.1 Frequency

The frequencies of the fundamental and clang tones decreased with increasing temperature for tuning forks of all materials.

For the aluminum and carbon steel tuning forks, the effect of temperature was almost the same among tuning forks of the same type. For the aluminum tuning fork, the fundamental frequency decreased by approximately 2.8 Hz, and the clang frequency decreased by approximately 18 Hz for a temperature increase of 20 °C. For the carbon steel tuning fork, the fundamental frequency decreased

![Figure 3 – Vibration evaluation.](image-url)
by approximately 1.6 Hz, and the clang frequency decreased by approximately 11 Hz for a temperature increase of 22 °C.

On the other hand, the stainless steel tuning forks showed more significant individual variability, with fundamental and clang frequencies decreasing by 1.2 to 2.2 Hz and 8 to 13 Hz, respectively, with a temperature increase of 14 °C. Considering the temperature range, the carbon steel tuning fork clearly showed a smaller decrease in frequency with increasing temperature than the tuning forks made of other metals.

### 3.2 Vibration Duration of the Fundamental Tone

Figure 4 shows the relationship between the temperature and the vibration duration of the fundamental tone of each tuning fork.

For aluminum and stainless steel tuning forks, the vibration duration of the fundamental tone decreased with increasing temperature, except for the stainless steel TF4, which recorded an extremely short fundamental tone. There were individual differences in the magnitude of the effect of increasing temperature on both materials.

On the other hand, for the carbon steel tuning fork, the vibration duration gradually increased with increasing temperature, with the longest vibration durations recorded at around 44°C for all forks. There was a tendency for vibration duration to shorten at higher temperatures. However, the vibration duration at 50°C, the highest temperature measured in this experiment, exceeded the vibration duration at the lowest temperature for all of the tuning forks.

### 3.3 Vibration Duration of the Clang Tone

Figure 5 shows the relationship between the temperature and the vibration duration of the clang tone of each tuning fork.

The clang tone of the aluminum tuning fork tended to prolong as the temperature increased. In contrast, the carbon steel and stainless steel tuning forks did not show any common variation due to the temperature effect.

![Figure 4](image)
(a) Aluminum tuning fork.

![Figure 5](image)
(b) Carbon steel tuning fork.

(c) Stainless steel tuning fork.

Figure 4 – Vibration duration of the fundamental tone.
4. **DISCUSSION**

Tuning forks of all materials showed a change in the vibration duration of the fundamental tone as the temperature increased. However, the variation in vibration duration varied with the tuning fork material. The fundamental tone of the aluminum and stainless steel tuning forks tended to shorten as the temperature increased, and the effect was greater for the aluminum fork. On the other hand, the carbon steel tuning fork tended to prolong the fundamental tone as the temperature increased, up to around 44°C, a result different from that observed for other metals. Material properties such as Young's modulus and coefficient of thermal expansion are considered to cause these differences in the effect of temperature. Also, the significant difference in the size of the tuning forks used in the experiments (Table 1) may have influenced the results.

The effect of temperature on the frequency and duration of the fundamental vibration showed that the carbon steel tuning fork exhibited minor variations with the temperature of any of the tuning forks. In addition, there was a tendency for the fundamental tone to prolong with an increase in temperature. These results suggest that carbon steel tuning forks are relatively more suitable for use at high temperatures than other materials, although a 1°C increase results in a decrease in the fundamental frequency of approximately 0.07 Hz. On the other hand, aluminum tuning forks, which showed a significant decrease in frequency and shortening of the fundamental tone with increasing temperature, are not considered suitable for use at high temperatures in terms of both frequency and vibration duration.

It should be noted that polishing, which is an integral part of the manufacturing process for carbon steel and stainless steel tuning forks, generates a great deal of heat. This result suggests that evaluating the vibration duration during the manufacturing process is not easy, which involves repeated polishing and tuning. In particular, the stainless steel tuning forks recorded almost identical vibration durations at high temperatures. However, at room temperature, differences were observed (Figure 4). This means
that the tuning of tuning forks at high temperatures after polishing does not necessarily result in optimal operation at room temperature, and it is necessary to take countermeasures.

5. CONCLUSIONS

In this research, we compared the effect of temperature on the vibration duration of tuning forks made of aluminum, carbon steel, and stainless steel. As a result, the vibration duration of the fundamental tone of the tuning forks made of aluminum and stainless steel tended to shorten as the temperature increased. On the other hand, the vibration duration of the tuning fork made of carbon steel gradually increased with increasing temperature, with the longest vibration duration recorded at around 44°C. The differences in material properties such as Young's modulus and coefficient of thermal expansion are considered to be the cause of these differences in the effect of temperature among the metallic materials. This temperature effect may be an issue in the manufacturing process to optimize the operation at room temperature.

Based on this relationship with temperature, we are aiming to establish a method for evaluating the vibration duration in the tuning fork manufacturing process. For this purpose, it is necessary to comprehensively examine the effects of shape parameters and other environmental conditions, in addition to holding strength and temperature, which have been investigated so far. By examining each of these parameters, we hope to contribute to change manufacturing technology of tuning forks, which relies on the experience of artisans now, to explicit knowledge.

ACKNOWLEDGEMENTS

We are grateful to NITI-ON Co. Ltd. for their cooperation. A part of this work was supported by JSPS KAKENHI Grant Numbers JP19K22950 and JP20H04497, and a Research Grant for the Next Generation Research Incubator from the Institute for Global Prominent Research, Chiba University for SN.

REFERENCES

Visualizing mode shape of a drumhead under non-uniform tension for estimating tuning condition using Fourier transform profilometry

Ryota HASHIMOTO(1), Kohei YATABE(2), Yasuhiro OIKAWA(3)

(1) Department of Intermedia Art and Science, Waseda University, Japan, hashimo_108@fuji.waseda.jp
(2) Department of Electrical Engineering and Computer Science, Tokyo University of Agriculture and Technology, Japan, yatabe@go.tuat.ac.jp
(3) Department of Intermedia Art and Science, Waseda University, Japan, yoikawa@waseda.jp

ABSTRACT
In this paper, vibration modes of a musical drumhead were visualized using Fourier transform profilometry (FTP). Purpose of the visualization is to estimate tuning condition of a drumhead under non-uniform tension. Since drums are often tuned independently of a musical scale, their tuning heavily relies on each drummer’s taste and experience. If modes of drumhead can be visualized, it will assist beginners in judging the uniformness of tension, which should enable them to tune drums more easily. To realize such visualization, we applied FTP, which is an optical method to measure the displacement of a surface, to a vibrating drumhead. Experimental tuning condition was set adjusting bolts tightness around a drumhead. Then, a video of a vibrating drumhead was shot with a high-speed camera. Time variation of the displacement and a mode shape were obtained applying FTP to each frame of the video. Comparing the visualized mode shape and tuning condition, correspondence between them was shown.

Keywords: Drumhead, Fourier transform profilometry, Tuning, Vibration mode, Visualization

1 INTRODUCTION
In typical music genres such as rock and jazz, drums are necessary instruments to control the playing speed and give thickness to sounds. For this reason, drums are used very frequently in most popular songs. However, since non-professional drummers often do not have their own drums, they use drums installed in a music rehearsal studio for practice. The use of the same drums by some drummers may cause their tuning to shift, or they may change the tuning. Due to this situation, the tuning of the drums varies each time, so it is necessary to tune them many times. A drummer tunes a drum adjusting the tightness of tension bolts around a drumhead. Since they are often tuned independently of a musical scale, their tuning depends largely on the taste and experience of each drummer. Therefore, it is not easy to tune it quickly. Furthermore, beginners do not know the right sounds for a music, so it is more difficult for beginners. To make the tuning easier and quicker for them and experienced drummers alike, estimating the tuning condition of a drumhead is needed. If it is achieved, the methodology can be applied to practical tuning equipment.

Non-uniform drumhead vibration has been analyzed by several numerical methods. For example, Mei calculated nodal patterns of a circular membrane under various tension conditions using the finite element method [1]. As another example, Sathej and Adhikari proposed a numerical method based on Fourier-Chebyshev collocation for the popular Indian drums tabla, which is eccentrically loaded [2]. In addition, a drumhead under non-uniform tension was modeled and analyzed by Samejima and Fukuda using a Fourier-Chebyshev spectral collocation technique, which is one of the spectral method [3]. They showed that under non-uniform tension, the degeneracy could be lifted and thus each mode split into two frequencies. They also mentioned that whether only one or both of the two degenerate modes are excited depends on the position of the striking.

For a real drum, some optical methods have been proposed to visualize the vibration. Worland measured
the effect of tuning conditions on drumhead vibration by electronic speckle pattern interferometry (ESPI), which uses laser interference [4]. It is shown that not only degeneracy can be lifted by the non-uniformities, but also that the angle of the nodal line changes or curves depending on bolts tightness. However, the laser method require delicate equipment such as beam splitters and several mirrors. Although a single sensor head composed of the necessary equipment is sufficient for ESPI, it is hardly simple enough for actual drum tuning in any case. Another example is the use of Fourier transform profilometry (FTP), a technique that uses only a camera and a video projector. By applying it to each picture frame, the 3D shape of the drumhead is reconstructed and a vibration animation is generated [5]. However, the measurement does not take into account non-uniform tension.

In this paper, we apply FTP to a non-uniform tension drumhead to visualize the effect of its tuning conditions on vibration. Experimental tuning conditions are set adjusting the tension bolts of a snare drum. The amplitude distributions for each conditions are obtained and their shapes are compared. Paying attention to the shape change of the (1, 1) vibration mode, we clarify the corresponding relationship with the tuning condition. The results shows regularity in the angle and bending of nodal lines, consistent with the previous experiments. Visualizing the relationship by FTP, which is easier to set up than the laser method, will improve the convenience of tuning estimation and its applicability to tuning tools.

2 FUNDAMENTAL CONCEPTS

2.1 Ideal circular membrane

An ideal membrane is assumed to have uniform areal density and tension, and there is no air loading or stiffness. The general solution of the wave equation for such an ideal circular membrane is given by

\[
\hat{z}_{m,n}(r, \theta, t) = \hat{\mathbf{C}}_{m,n} J_m(k_{m,n}r) \begin{cases} 
\cos (m\theta) \\
\sin (m\theta)
\end{cases} e^{j\omega_{m,n} t},
\]

where \(\hat{\mathbf{C}}_{m,n}\) is displacement amplitude, \(J_m(\cdot)\) is the Bessel functions of the first kind of order \(m\), the \(\theta\) is an angle in the polar coordinates with an antinode position as zero, and the \(k_{m,n}\) is a constant that makes \(J_m(k_{m,n}R)\) equal to the \(n\)th zero at the circumference \((r = R)\), giving the angular frequency \(\omega_{m,n}\) [6,7]. These form a \((m,n)\)
mode which has $m$ nodal diameters and $n$ nodal circles including the circumference.

Figure 1 shows an example of $(m, n)$ mode of an ideal circular membrane. When the number of nodal diameter is greater than zero ($m > 0$), Eq. (1) has two solutions for one $m$ because it contains both $\cos(m\theta)$ and $\sin(m\theta)$. The pairs of modes corresponding to these solutions have the same frequency and are doubly degenerate. Their nodes and antinodes positions are interchanged (rotated by an angle $\pi/2m$) as shown in Fig. 2. The mode degeneracy can be lifted by perturbations such as non-uniform tension and areal density and become separately observed. The spectral method and ESPI visualizes such degenerate modes and shows that they change the angle and bending of their nodes.

2.2 Fourier transform profilometry

FTP is one of the optical methods to acquire 3-D object shapes. A sinusoidal grating of a certain frequency is projected on an object surface and photographed by a camera from directly above. The surface displacement is obtained by analyzing the distortion of the stripes caused by the unevenness of the surface as phase modulation [8]. Using a high-speed camera and processing one frame at a time, the system can be applied to dynamic objects [9–12]. It can be conducted using only a camera and a video projector, which eliminates the need for detailed and delicate equipment arrangement unlike the laser methods.

Figure 3 shows a schematic diagram of FTP. To begin, a sinusoidal grating is projected onto the drumhead. Let $f_0$ be the spatial frequency of the grating image in the reference plane, and given that the light source of the grating is at infinity $E_\infty$. The observed luminance value is expressed as

$$g_T(x, y) = \sum_{n=-\infty}^{\infty} A_n r_0(x, y) e^{2\pi i n f_0 x},$$  

where $A_n$ is a Fourier coefficient and $r_0(x, y)$ is a non-uniform distribution of reflectivity on the reference plane. However, the actual light source is at $Pr$, not at infinity $E_\infty$. Since there is a phase modulation distribution

$$\phi_0(x, y) = 2\pi f_0 BC$$  

due to the projection angle in practice, the fixed reference plane is as follows:

$$g_0(x, y) = \sum_{n=-\infty}^{\infty} A_n r_0(x, y) e^{(2\pi i f_0 x + n\phi_0(x, y))},$$
which represents a non-vibrating drumhead image.

When the drumhead is vibrating, an image observed at time $t$ is

$$ g(x, y, t) = \sum_{n=-\infty}^{\infty} A_n r(x, y, t) e^{i(2\pi f_0 x + n\phi(x, y))}, $$

(5)

where $r(x, y, t)$ is the reflectivity distribution due to the height distribution of the drumhead, and the $\phi(x, y, t)$ is the phase modulation due to both the height and the projection angle at time $t$,

$$ \phi(x, y, t) = 2\pi f_0 BD. $$

(6)

Computing the space-domain 2-D Fourier transform of $g_0(x, y)$ and $g(x, y, t)$, only the $n = 1$ components corresponding to the fringe frequency $f_0$ are extracted. Applying the inverse Fourier transform to the extracted components, we obtain

$$ \hat{g}_0(x, y) = A_1 r_0(x, y) e^{i(2\pi f_0 x + \phi_0(x, y))}, $$

(7)

$$ \hat{g}(x, y, t) = A_1 r(x, y, t) e^{i(2\pi f_0 x + \phi(x, y))}, $$

(8)

Then, calculating the product of $\hat{g}$ and the complex conjugate of $\hat{g}_0$ yields

$$ \hat{g}(x, y, t)\hat{g}_0^*(x, y) = |A_1|^2 r(x, y, t) r_0(x, y) e^{i\Delta \phi(x, y)}, $$

(9)

$$ \Delta \phi(x, y, t) = \phi(x, y, t) - \phi_0(x, y), $$

(10)

and the phase difference at each frame is given by

$$ \Delta \phi(x, y, t) = \text{atan2} \{ \text{Im} [\hat{g}(x, y, t)\hat{g}_0^*(x, y)], \text{Re} [\hat{g}(x, y, t)\hat{g}_0^*(x, y)] \}, $$

(11)

where the function $\text{atan2}(b, a)$ is the unique phase of a complex number $a + bi$. The $\text{Im}[:]$ and $\text{Re}[:]$ represent the imaginary part and real part, respectively. The phase is often wrapped in the range of $[-\pi, \pi]$, therefore, 3-D phase unwrapping is necessary for practical purposes. In this research, the branch cut method which is one of 2-D phase unwrapping algorithm was applied to the final frame. After that, 1-D unwrapping was performed toward the front in the time direction, resulting in phase unwrapping of all frames [13].
Table 1. Experimental equipment and conditions.

<table>
<thead>
<tr>
<th>Equipment</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-speed camera</td>
<td>MEMRECAM HX-3 (nac)</td>
</tr>
<tr>
<td>Frame rate</td>
<td>2000 fps</td>
</tr>
<tr>
<td>Projector</td>
<td>VPL-PHZ10 (SONY)</td>
</tr>
<tr>
<td>Drum tuner</td>
<td>DATK Torque Key (EVANS)</td>
</tr>
<tr>
<td>$L_0$</td>
<td>1145 mm</td>
</tr>
<tr>
<td>$d$</td>
<td>755 mm</td>
</tr>
<tr>
<td>$f_0$</td>
<td>0.1408 line/mm</td>
</tr>
</tbody>
</table>

Figure 5. Example images of a snare drum taken.

(a) A frame where the drumstick was photographed.  
(b) A vibrating drumhead frame.

The height distribution is obtained from the phase difference $\Delta \phi(x,y,t)$. From Eqs. (3) and (6), Eq. (10) gives the following:

$$\Delta \phi(x,y,t) = \phi(x,y,t) - \phi_0(x,y) = 2\pi f_0 (BD - BC) = 2\pi f_0 CD.$$  

(12)

Furthermore, similarity relation between $\triangle \text{PrHCa}$ and $\triangle \text{CHD}$ gives

$$d : \overrightarrow{CD} = \{L_0 - Z(x,y,t)\} : \{-Z(x,y,t)\},$$

(13)

$$\frac{d}{\overrightarrow{CD}} = \frac{-dZ(x,y,t)}{L_0 - Z(x,y,t)}.$$  

(14)

Substituting Eq. (14) into Eq. (12), we obtain time variation of the height distribution,

$$Z(x,y,t) = \frac{L_0 \Delta \phi(x,y,t)}{\Delta \phi(x,y,t) - 2\pi f_0 d}.$$  

(15)
3 EXPERIMENT

3.1 Equipment and tuning condition

The measurement object is the only batter side of a snare drum with six tuning bolts around it. The batter side is the side that is struck when playing the drums. The system was set up as shown in Fig. 4, and Table 1 shows the conditions for FTP. The projected sinusoidal grating consists of sine waves with 100 gradations from the minimum to maximum. Experimental tuning conditions were set adjusting bolts tightness using a type of drum tuners, “Torque Key.” It idles when a certain amount of torque is applied, allowing bolts to be tightened evenly. Besides, the user can decide the tightening strength with a number from 0 to 20 written on the tuner. We assigned numbers 10 and 20 to loosen and tighten bolts respectively, and set several tuning conditions depending on their position. All bolts around the drumhead on the back side were set to number 10.

At the start of the measurement, the center of the drumhead was struck with a wooden drumstick, and at the same time, a high-speed camera began taking video. Three measurements were taken per tuning condition to examine FTP stability. To avoid using frames where the drumstick was shot, we selected 1000 frames from the 100th to the 1100th frame after the striking. The vibrating drum images of $1920 \times 2560$ pixels was then obtained. Moreover, they were reduced to $1800 \times 1800$ pixels by removing pixels outside the plane. Figure 5 shows the examples of the taken images.

(a) Schematic diagrams of tuning condition. The white and black points are the loosened and tightened bolts respectively.

(b) Linear $(1,1)$ mode shapes.

(c) Curved $(1,1)$ mode shapes.

Figure 6. Variation of $(1,1)$ mode shapes with tuning conditions.

Figure 7. $(1,1)$ mode shapes obtained in three trials for one condition.
3.2 The (1,1) mode
The (1,1) mode shape can be obtained from the calculated time variation of the height. For each pixel \((x, y)\) of the height distribution \(Z(x, y, t)\), the amplitude spectrum was obtained by Fourier transform in the time domain. The frequency whose amplitude is the second largest was obtained in each pixel because the first may include the \((0,1)\) mode, which is not the objective. We counted the number of pixels where the same frequency is obtained and listed ten frequencies with large pixel count. Then, we saved ten amplitude distribution images corresponding to them. Finally, the \((1,1)\) mode shapes were obtained in the saved images.

3.3 Comparing mode shapes
Figure 6 shows the \((1,1)\) mode shapes obtained for each tuning condition. The white and black points in Fig. 6(a) refer to the positions of the loosened and tightened bolts, respectively. The amplitude ranges from 0 mm to 1.1 times the maximum of each distribution.

Figure 6(b) shows amplitude distributions of linear nodal lines. The line rotated as the number of tightened bolts increases. Also, the nodal lines tended to appear as if they connect black and white points located at opposite sides of the circumference. This may be due to the geometric symmetry of the tension distribution. Figure 6(c) shows the other \((1,1)\) mode, which was degenerate. These became able to be observed separately because the degeneracy was lifted by the non-uniform tension. Arc-shaped nodal lines appeared around regions of low tension. This result is consistent with the calculation by the spectral method and the ESPI observation. In addition, the curve of the nodes was more gentle in the condition with more black or white points. This means that as the tension becomes more uniform, the arc node approaches a linear shape such as the \((1,1)\) mode in Fig. 1.

Figure 7 shows the mode shapes obtained from three measurements for one condition. Although the struck spots were not exactly the same throughout the three trials, all the mode shapes looked alike. Thus, the changes in mode shape in Fig. 6 can be directly attributable to the tuning conditions.

4 CONCLUSIONS
In this paper, the \((1,1)\) mode of a non-uniform tension drumhead is visualized with only a camera and a projector by FTP. The stability of the method was demonstrated by making three measurements per tuning condition. Moreover, this allowed us to compare the mode shapes under each tuning condition. Particularly, two regular features were visualized: linear nodes follow geometric symmetry and arc nodes enclose regions of low tension. This indicates that the FTP can provide information that may lead to tuning estimation. In the future, we will investigate specific methods for estimating unknown tuning states from mode shapes.

REFERENCES


Difference in timbre of violin in a seat and on stage

Yuya IHIGAKI1, Masao YOKOYAMA1
1Meisei University, Japan

ABSTRACT
Generally, it is speculated that the tone of Stradivarius is superior to those of other violins, and it sounds farther. However, many points regarding the difference between the acoustic features of Stradivarius and that of the others are still unclear. In this study, we recorded the sounds from six violins from old violin to brand new violin on a stage and a seat. We then compared the frequency spectrum and the signal-to-noise ratio (S/N ratio) of those sound sources. Based on the results, the difference in the power spectrum was observed at approximately 3–5 kHz. And the difference of S/N ratio was also appeared in some frequency bands. In the higher harmonic overtone around 16 kHz, we found that there are violins which have the peaks of harmonic overtone and ones which do not have. We investigated whether the difference in the shapes of these frequency spectrums and S/N ratios were related to the tone color of violin and the so-called distant peal phenomenon.

Keywords: Violin, timbre, Frequency spectrum, S/N ratio

1. INTRODUCTION
Studies on the sound of antique violins such as Stradivarius have been conducted extensively.1-5 We analyzed the tone color with focusing on a phenomenon called the “distant peal” in this study. “Distant peal” is a phenomenon that audience can hear the sound well and clearly even if the sound volume of performance is not so loud. In contrast, a phenomenon called “side rumble” which the sound is not sufficiently audible even when violinist performs violin loudly.6 Generally, the sound by distant peal is considered good so that players keep in mind in playing and maintaining the distant peal. In addition, it is said that there are some violins which sound in distant peal and some which sound in side rumble. Therefore, in this study, we compared the power spectrum and S/N ratio on a stage and seat using six violins and examined the existence of the distant peal of the violins.

2. EXPERIMENT
The recording was conducted in a concert hall (Toppan hall, Tokyo) without an audience. The player was a professional violinist. The six violins used in this study were Stradivarius (1705, violin S), Del Gesu (1742, violin D), the beginner’s violin (violin P) made in Japan, Ferdinand (2011, violin F) made in Germany, and two other brand-new violins manufactured by two Japanese violin makers (violin H and N). The microphones used for measurements were two nondirectional microphones (Earthworks M23). As shown in figure 1, we placed the microphone on the stage (over a 20 cm from the bridge of violin) and the seat (15 m from the stage). We instructed the player to play G major scale (three octaves) with each musical instrument. We performed spectrum analysis after recording the sounds and compared the power spectrum and cross-spectrum for each violin. In addition, we calculated the S/N ratio in the 1/1 octave band around A4 (band 1), at approximately 3-5 kHz (band 2), and 10-12 kHz (band 3).

\[
S/N\text{ ratio} = \frac{\text{Average of peaks}}{\text{Average of noises}}
\] (1)
3. RESULTS
3.1 Comparison of the power spectrums

We present the results for the power spectrum (time average) in the A4 sound of each violin in Figures 2-7.

Figure 2. Spectrum in the A4 sound of violin S

Figure 3. Spectrum in the A4 sound of violin D

Figure 4. Spectrum in the A4 sound of violin P

Figure 5. Spectrum in the A4 sound of violin F
We focus on the difference between the power spectrum of the stage and seat. First, for the spectrum on the stage, some strong peaks appeared from a fundamental frequency (f0) to 60 kHz. Second, while the peaks of violins P and F decreased gently over f0, the peaks of violins D, N, and H are significant at approximately 3 kHz. For the spectrum recorded on the seat, some peaks exceeding 40 dB were observed for violin S in the 3-5 kHz compared with the other violins. In addition, the peaks are not seen at approximately 16 kHz for violin N.

Next, we present the results of the cross-spectrum in the A4 sound of each violin in Figures 8-13.
First, the peak of the fundamental frequency of violin S was not significant compared to those of the other violins. Second, some peaks exceeding 2.5 kdB appeared in violin S from the fundamental frequency to 3 kHz. Third, gentle disruptions of the peaks appeared from 10 to 12 kHz for violins S and D. According to these results, the difference of power spectrum is observed in the A4 sound from a fundamental frequency to 6 kHz. It is seemed that this phenomenon is similar to the singers’ formant which we often observe singers voice in opera. It is reported that the sound which the opera singer sings with emphasizing a specific frequency (2.5 kHz – 3.5 Hz) are not buried among the orchestra sounds. In addition, regarding the higher harmonic overtone over 16 kHz, the difference in the peak of harmonic overtone was observed among 6 violins, and it seems to be a characteristic of violin concerning the ease of distant peal.

3.2 Comparison of the S/N ratios
We show the difference in the S/N ratios of the 6 violins in the A4 sound in Tables 1 and 2.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument S</td>
<td>1.55</td>
<td>1.56</td>
<td>1.36</td>
</tr>
<tr>
<td>Instrument D</td>
<td>1.62</td>
<td>1.36</td>
<td>1.27</td>
</tr>
<tr>
<td>Instrument P</td>
<td>1.66</td>
<td>1.57</td>
<td>1.35</td>
</tr>
<tr>
<td>Instrument F</td>
<td>1.53</td>
<td>1.35</td>
<td>1.23</td>
</tr>
<tr>
<td>Instrument N</td>
<td>1.68</td>
<td>1.57</td>
<td>1.37</td>
</tr>
<tr>
<td>Instrument H</td>
<td>1.71</td>
<td>1.60</td>
<td>1.42</td>
</tr>
</tbody>
</table>
In band 1, the values of the S/N ratio on seat increased for all instruments comparing with that of on stage. In particular, the values significantly increased in violins S, N, and H. In band 2, the values of the S/N ratio increased for all violins except instrument N; the increase for violins S, D, and H are significant. In band 3, the S/N ratio of violins S, D, and F increased, and the increase for violin D was significantly high. Based on these results, the S/N ratio of violins S, D, and F increased in all bands. Especially, that of violins S and D increased significantly. Therefore, it may be the reason that audience at a seat hear the sound well and clearly.

4. CONCLUSIONS
In this study, we observed the differences in the frequency spectrum and the S/N ratio on a stage and a seat and analyzed the tone color of violins. We analyzed the property of each violin by comparing the peaks and S/N ratios in some frequency bands. In particular, the phenomenon similar to the singers’ formant was appeared in 3-5 kHz. This might be the acoustic explanation to the distant peal of violin. In a future study, we will compare the sound of violins in terms of various pitch and analyze other acoustic characteristics.

REFERENCES
ABS-0762

Personal learning portfolio for music university students based on acoustic analysis of performance sound

Masanobu MIURA

1 Kunitachi College of Music, JAPAN

ABSTRACT

University students of music performance are usually under severe training, but they cannot receive objective judgment since the evaluation factors are not clarified. The author proposed a personal learning portfolio for the students based on acoustic analysis of performance sound. In this project, a marimba performance is dealt with to analyze. Several styles of grip for two mallets in a single hand are investigated. Several marimba players play several stroking patterns on the mallet. Sound spectrogram and several MIR parameters (such as Spectral Flux and Spectral Centroid) were used to describe their performance features. Subjective judgements are conducted for the recorded sounds, where marimba players were asked to answer subjective impressions for each sound. The author designed a panel on display where the sound’s waveform, spectrogram and MIR parameters are drawn. By observing the panel, the players are possible to understand their own performances visually. Students are also possible to compare their playing sounds with other students, including their own playing done several years ago. At the conference, the author will demonstrate a concrete example of MIR parameters on the students’ personal learning.

Keywords: Musical Performance, Sound Analysis, Learning Portfolio, MIR Parameters

1. INTRODUCTION

The recent development of IT provides many opportunities for university education, such as online resources for university education and SNS (social network services) styled learning have been provided where students can access the educational resources and can learn without person-to-person instructions from teachers. In the case of musical university students, such kind of development has not been seen so much since many studies of music training are based on personal instructions. Moreover, from the student’s perspective, an ad-hoc style of learning is widespread in musical performance training since measuring the extent of music learning is essentially difficult. In order to improve the educational environment of musical university students, a personal learning portfolio is proposed here, where the musical university student’s performance sound will be accumulated and then compared year by year by him/herself. Here, percussion playing is targeted. The marimba playing is dealt with. The performance was previously recorded by other literature [1,2]. Here the recorded sound is employed to develop a methodology of percussion playing’s learning portfolio.

2. ACOUSTIC ANALYSIS

2.1 Recording Experiment [1,2]

Two students (P1 and P2) majoring in percussion performance participated in this test. The recorded playing is from the former study of the author and Mr. Takeshita, a PhD candidate for percussion playing at Kunitachi College of Music. The instrument used was a Yamaha YM-5000, and mallets were made by Lesta Jay MG-04B and MG-05B with modifications. The performance sound was recorded using two RODE NT2A microphones placed 13 cm from the marimba parallelly located to the change of pitch and a Roland Quad Capture mixer connected to a computer. The sheet music played was the most standard interval in the performance, shown in Figure 1, which includes the widths of the major-third interval. The tempo was set to 100 bpm, and the participants were asked to

---

1 miura.masanobu@kunitachi.ac.jp
play a pattern in which they played two notes simultaneously with one hand, then alternatively stroked each of two mallets. All audio signals are recorded by 44.1kHz, 24bit sampling.

![Sheet music of recorded performance](image)

**Figure 1 – Sheet music of recorded performance [1,2]**

2.2 Analysis

For the recorded performance sound, MIR (Music Information Retrieval) parameters, such as Spectral flux $\text{Flux}_n$ (1) is calculated using the following equations.

$$\text{Flux}_n = \sum_{k=1}^{K} |X_n(k) - X_{n-1}(k)|$$

where $n$ represents the index of frame on short time Fourier transform, $k$ represents bin index for windowed waveform, $K$ is the total number of bin index or half of window length, and $X$ represents power spectrum of audio signal $x(t)$. The Fast Fourier Transform is performed with the length of 4096 samples with hanning window. In general, the spectral flux is said to represent the fluctuation of the sound spectrogram in conjunct two windowed signals. At here, the flux is used to represent timbre change when the stroke occurs.

Spectral Centroid $\text{SC}_n$ (2) is calculated as follows

$$\text{SC}_n = \frac{\sum_{k=1}^{K} \log_2 f(k) X_n(k)}{\sum_{k=1}^{K} X_n(k)}$$

where the $f(k)$ represents the centre frequency of the $k$-th bin index. The spectral centroid is the moment centre of gravity on the sound power spectrum, representing the sensation of brightness on current sound.

2.3 Example of calculated results

Figure 2 shows an example of a recorded performance for the sheet music in Figure 1. The results show four different figures, where the first and second are acoustic signals from two separate microphones. The first is derived from the relatively higher pitch region, whereas the second is from the relatively lower pitch region. The third one represents the change of Spectral Flux, where the blue is from the microphone from the higher pitch region and the red curve is from a relatively lower area. The fourth figure shows the spectral centroid, where the abscissa represents time in music, and the
ordinate represents the pitch as the order of power of two. Due to the essentials of the spectral centroid, when no performance sound is on the recorded audio, the spectral centroid provides no practical information about the sound. At this figure, SC reaches $2^{11}$ to $2^{13}$ (Hz), which is not reflected in the real spectrogram. Figure 3 shows the sound spectrogram of Marimaba’s stroke, from which the spectrum centroid will likely be around $2^8$ or $2^9$.

![Figure 2](image2.png)  
Figure 2 – Example of result for sheet music in Fig.1, played by P1.

![Figure 3](image3.png)  
Figure 3 – Sound spectrogram for the first recorded note in Fig.2

### 2.4 Example of calculated results

Figure 4 shows another example for sheet music in Fig.1. Even if the P1 and P2 use the same marimba and mallet, the performance is not the same. Concerning the first six strokes, those on P1 are played in a clear timbre, whereas those on P1 give a more robust, impulsive and rugged impression. The impression is undoubtedly tricky between the two, but the use of only sound spectrogram is not efficient since the difference is difficult to be displayed. Figure 5 shows the sound spectrogram played by P2 for the first note in Fig.2, which can be compared with Figure 3. As can be seen from Fig.3 and Fig.5, no dramatic difference is observed, but the timbre is inconsistent. In practical situations in training and self-learning, musical university students only have to judge it according to their senses. Following the red and blue curve on the first note in Figure 2 and Figure 4, that on Figure 2 (or P1’s playing) gives no such difference between the blue and red curves. However, Figure 4 (or P2’s playing) provides a substantial difference in time between red and blue. The blue curve is delayed from the red curve. Since the blue show a higher-pitch microphone, the change of spectrum is different between P1 and P2. More specifically, the playing in P2, the higher-pitch component, stays longer than the P1.
Figure 4 – Another example of Figure 3, played by P2.

Figure 5 – Sound spectrogram for the first recorded note in Fig.4

3. OUTLOOK OF PORTFOLIO FOR MUSICAL STUDENTS

3.1 Why needed?

Students of musical universities usually only rely on their sense or intuition when evaluating their playing. It is inefficient because it is not easy for students to judge their playing. The recorded acoustic signal is helpful since they can play back it, but it is time-consuming. The proposed personal portfolio does not always require the student to understand their playing but provides some informative features or aspects of their performance. Therefore, for music university students, the proposed portfolio will be effective.

3.2 Possible method to use as portfolio

Following is the possible flow of the portfolio of students. Firstly standard sheet music is set to be played several times over four years. They played it with the same instrument and mallet, with the same microphone and exact place to record. The recorded acoustic signal is represented as a panel like Figure 2. The recorded acoustic signal and MIR parameters are shown graphically, and by observing and listening to their playing, students are possible to recognize their playing's detailed
features. The recorded playing with the analyzed result is accumulated while the students belong to the university, and parametric comparison is conducted to the students' feedback.

4. CONCLUSIONS

In this presentation, a method to evaluate the timbre of a played sound is explained. Moreover, the possible usage of the result of MIR parameter as the descriptor of musical playing. Finally, the use of MIR parameters to realize the student’s portfolio that enables students to think about their own playing more is introduced. In the near future, the proposal of this system to my affiliation will be discussed.

ACKNOWLEDGEMENTS

This study is partly supported by Research Fund on Kunitachi College of Music.

REFERENCES

1. Takeshita K., Basic study for the four mallet performance on Marimba using motion capture and acoustic analysis, Report on Master Course Study in Graduate School of Kunitachi College of Music (2022.2, in Japanese)
Music signal analysis in the Large Time Frequency Analysis Toolbox

Clara HOLLOMEY*, Nicki HOLIGHAUS, Peter BALAZS
Acoustics Research Institute, Academy of Sciences, Austria
*clara.hollomey@oeaw.ac.at

ABSTRACT
The Large Time-Frequency Analysis Toolbox (LTFAT) is a toolbox for Matlab and GNU Octave that is dedicated to the time-frequency analysis of audio signals. Based on mathematical frame theory, the generalization of vector bases to redundant sets of vectors, LTFAT provides a variety of time-frequency representations that are particularly suitable for analyzing music signals. Besides an implementation of the classical short-time Fourier transform, it comprises invertible auditory filterbanks using the mel-, Bark- and ERB-scale, an invertible implementation of the constant-Q transform, and algorithms for calculating the wavelet transform. Those signal representations can be readily used for audio applications such as denoising, audio data compression, and the separation of tonal and transient components, each of which being exemplified with software demonstrations in the toolbox. LTFAT's blockprocessing framework allows for their real-time application and playback, directly from Matlab and Octave. The toolbox can, along with its extensive documentation, be downloaded from ltfat.org.

Keywords: Audio signal analysis, Filter banks, open source

1 INTRODUCTION
The Large Time-Frequency Analysis Toolbox (LTFAT) has been designed as both an educational and a computational tool for applied harmonic analysis and audio signal processing [19, 14]. Besides the well known windowed Fourier transform, LTFAT contains implementations of perceptually motivated [12], constant-Q [7], and wavelet transforms [15]. All those routines operate on nonlinear frequency scales, and implementing them such that one is still able to reconstruct the time-domain signal from the time-frequency representation can be challenging, even in oversampled scenarios. Moreover, particularly in applications where the reduction of (audio) data, or conversely, increased robustness to noise is required, being able to adjust the resulting time-frequency representation’s redundancy is desirable.

In LTFAT, functionality for achieving time-frequency representations that fulfil those requirements is implemented as filter banks. Embedded in an efficient framework for signal analysis, processing, and synthesis, they are versatile tools for a wide range of applications in music signal analysis and processing. In the following, we summarize the mathematical background of LTFAT’s filter bank framework, describe its structure, and introduce its most useful transforms for music signal analysis.

2 MUSIC SIGNAL ANALYSIS VIA FILTER BANKS
2.1 Mathematical Background
Filter banks split an input signal \(x[n]\) into a set of signals \(x_1[n], x_2[n], \ldots\), such that each of them corresponds to a different frequency region of \(x[n]\). Often implemented as an array of modulated lowpass (i.e. bandpass) filters complemented with a downsampling unit for preventing excessive redundancy, they are an established tool for working with music signals, with applications comprising their coding [13], compression [17], and processing [16].
To ensure that the time domain signal can be recovered, one often imposes strong constraints on the filters and their spacing in frequency, by either requiring the filters’ frequency response to be zero at integer multiples of the downsampling rate $\frac{1}{a}$ [21], or, for greater robustness against modifications and noise, to be altogether band limited consistent with $a$ [18].

Those constraints, however, can impede the analysis of music signals. Music often covers a large frequency range, and consequently, comprises frequency components with a variety of wavelengths. Moreover, it is non-stationary in nature, and changes in the high frequencies may occur on much shorter time scales than those in the low frequency range. Yet, at the same time, a great deal of information, e.g. on the musical instruments used, their relative intensity, or even the playing style can be conveyed by fine grained changes in the time-frequency properties. Hence for music signals, flexibility in assigning the frequency resolution to time-frequency representations is of particular interest. However, the design of filter banks that allow for such flexibility requires the relaxation of the above mentioned constraints. To still obtain invertible time-frequency coefficient matrices, strategies are needed that take variations over the whole time-frequency plane covered by the filter bank into account.

Such strategies are provided by frame theory. A finite set of vectors $\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$ is called a frame [6] when there exist positive real numbers $\lambda_{\min}$ and $\lambda_{\max}$ such that

$$\lambda_{\min} \|x\|^2 \leq \sum_{k \geq 1} |\langle x, \phi_k \rangle|^2 \leq \lambda_{\max} \|x\|^2, \text{ for every } x. \quad (1)$$

Frame theory provides a way to mathematically describe redundant time-frequency representations where the downsampling in time does not necessarily equal that in frequency, such as filter banks [2]. Frames generalize the concept of bases in that they may comprise more vectors than are strictly needed to span the vector space, thus providing a higher degree of flexibility for the design of time-frequency representations while simultaneously ensuring their invertibility. Thus, frame theory provides the mathematical foundation for LTFAT’s filter bank framework. In the following, we detail the framework’s structure and outline the functionality that is particularly suitable for analyzing music signals.

2.2 Filter banks in LTFAT

LTFAT provides the filterbank function, implemented in Matlab/GNU Octave with a backend in C for greater numerical efficiency, for calculating filter bank coefficients according to

$$c_m[n] = \sum_{l=0}^{L-1} x[l] g_m[a_m n - l], \quad (2)$$

with $x$ the input signal, $g_m$ the $m^{th}$ filter, $a_m$, its associated downsampling factor, and $L$ the length of the filter bank. filterbank accepts the input signal along with the filter type and downsampling factors as input arguments. The filter bank coefficients, the frequency responses of its filters, and the filter bank’s overall frequency response can be visualized using the functions plotfilterbank, filterbankfreqz, and filterbankresponse. Moreover, there is functionality for calculating the instantaneous time and frequency (filterbankphasegrad), conducting phase-based spectrogram reassignment (filterbankreassign), and reconstructing the phase from the filter bank’s magnitude coefficients (filterbankconstphase). The frame bounds can be retrieved for complex and real signals via filterbankbounds and filterbankrealbounds respectively. Their ratio indicates whether direct inversion using the ifilterbank function, and possibly a dual window generated via filterbankwin, or iterative inversion via ifilterbankiter is more appropriate for a given time-frequency coefficient matrix.

Moreover, LTFAT provides functionality for generating filter functions that are specifically matched to the requirements of certain types of time-frequency representations. In their minimal configuration, most filter generators accept the sampling frequency, the desired filter bank length, and the frequency range to be covered.
as input arguments and output the appropriately scaled filter functions, their downsampling factors and center frequencies. Thus, they can be readily used in conjunction with \texttt{filterbank}. There are filter generators for designing filter banks in the style of the STFT, the constant-Q transform, the wavelet transform, for perceptually motivated filter banks, and a warped filter bank, the latter accepting nearly arbitrary center frequency to bandwidth ratios. The generic signal flow for signal analysis via filter banks in LTFAT is, along with the most relevant parameters, depicted in Figure 1.

Figure 1. LTFAT’s signal flow for filter bank processing. The filter generators are named after the type of filter they yield (i.e. either \texttt{gabfilters}, \texttt{audfilters}, \texttt{cqtfilters}, \texttt{waveletfilters}, \texttt{warpedfilters}) and output the desired filters and downsampling factors such that they can be readily processed, along with the input signal, via the \texttt{filterbank} function. Depending on the framebounds of the resulting coefficients, either direct or iterative inversion can be applied.

\subsection{2.2.1 Gabor filter banks}

The short time Fourier transform (STFT), where the signal of interest is segment-wise expanded in terms of its Fourier coefficients

\[ c[n] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi \frac{n}{N}} \]  

provides a uniform and fixed resolution for each time-frequency point and is probably the most commonly used time-frequency representation. To reduce redundancy, downsampling in both, the time and the frequency domain is often applied. However, Fourier theory lacks concise criteria for the reconstruction of such downsampled STFTs. In line with the concept of frame theory, Gabor theory views time-frequency representations as expansions of the time-domain signal in terms of shifted and modulated windows

\[ g_{m,n}[t] = e^{-2j\pi mbt} g[t - na], \]  

where \( a \) and \( b \) are the downsampling factor in time and frequency, respectively. The resulting coefficients have the structure

\[ c[m,n] = \sum_{k=0}^{K-1} x[k] g_{m,n}[k]. \]
Thus, Gabor theory provides a unified mathematical framework that encompasses the STFT and provides clear definitions for downsampling in the time-frequency domain. For this reason, in LTFAT, the filter generator targeting time-frequency analysis on equidistantly spaced time-frequency grids is termed *gabfilters*.

### 2.2.2 Perceptually motivated filter banks

The *audfilters* generator provides a linear representation with resolution evolving across frequency. It matches the human auditory time-frequency resolution, both in terms of frequency spacing and bandwidth.

Supported scales comprise two variants of the mel scale [20], the critical band rate in bark [22], and two variants of the "equivalent rectangular bandwidths" (ERB) scale [11]. An overview of all auditory scales available in LTFAT is depicted in Figure 2. To ensure the filter bank’s invertibility, low- and highpass filters, centered around 0 Hz and the Nyquist frequency, respectively, are part of the filter generator’s output.

![Figure 2. The auditory scales available in LTFAT.](image)

### 2.2.3 Constant-Q transform

The constant-Q transform has been specifically designed with musical instruments in mind [4]. Based on the observation that the absolute positions of harmonics depend on the fundamental frequency while their relative distance remains constant on a logarithmic scale, it allows for the concise representation of time-frequency patterns, largely invariant to their fundamental frequencies. An example of that relative frequency invariance of instrumental patterns is depicted in Figure 3. A further benefit of logarithmically spaced filter center frequencies arises when audio data covers several octaves, since, as compared to equidistant frequency spacings, fewer frequency bins may be required to cover a given frequency range.

In LTFAT, the *cqtfilters* generator provides an invertible constant-Q transform. Similar as for the *audfilters* function, additional low- and highpass filters ensure stable invertibility.

### 2.2.4 Wavelet transform

Wavelets are an alternative to Fourier coefficients for mapping a time-domain signal to the time-frequency domain. LTFAT provides *waveletfilters*, a generator for redundant wavelet filter banks as introduced in [8]. Considering wavelet filters of the form

\[
\psi_{j,n}[n] = \frac{1}{\sqrt{s_j}} \psi \left( \frac{n - d(l + \delta_j)}{s_j} \right)
\]

(6)

with \( \psi \) the mother wavelet, \( s \) its dilation, \( d(l + \delta_j) \) its translation parameter, and with \( \delta_j \) corresponding to a small delay generated by a low discrepancy sequence \((\delta_0, \delta_1, \delta_2, \ldots)\), allows for greater freedom in adjusting the
frequency resolution than is usually provided by dyadically sampled transforms. At the same time, the resulting transform fulfills the frame property, allowing for flexible adjustments of its redundancy. Lowpass filters ensuring waveletfilters's invertibility are added to the generator's output.

2.2.5 Warped filter banks
The warpedfilters function proposed in [9] allows for the direct provision of a function freqtoscale that defines the center frequency to bandwidth ratio of the filterbank. Therefore, it offers the greatest amount of freedom in filter bank design. The function freqtoscale can be freely chosen, provided that it is an invertible, increasing real function. However, care needs to be given to the scaling of that function, which needs to take the desired number of frequency channels into account, and for the function not being too steep locally, as this may yield filters with very low bandwidths, potentially causing unsatisfactory results.

3 FINAL REMARKS
While LTFAT's main focus is on invertible representations, undersampled representations are equally easily constructed when desired. Besides most of the transforms presented here that emphasize freedom in choosing the center frequency to bandwidth ratio of the filters, LTFAT features a transform that allows the window to change with time [1]. This non-stationary Gabor transform provides flexibility in the time-domain while still allowing for an efficient fast Fourier transform implementation, and is available via the nsdgt function call. Moreover, many of LTFAT's time-frequency transforms can be used in real time settings via LTFAT's block processing framework, directly from Matlab and Octave. In addition to offering freedom in the design of time-frequency representations for music analysis, LTFAT also provides algorithms for their processing. They comprise efficient time-varying filtering via Gabor multipliers [5], tonal-transient separation via the time-frequency jigsaw algorithm [10] and the lasso shrinkage [3], as well as approaches for audio denoising and compression. Demos and examples are available directly from within the toolbox and on the homepage ltfat.org.

ACKNOWLEDGEMENTS
This work was partly supported by the Austrian Science Fund (FWF): LOFT (FWF - P 34624, Localized, Fusion and Tensors of Frames) and NoMASP (FWF P 34922-N, Nonsmooth Nonconvex Optimization Methods for Acoustic Signal Processing).
REFERENCES


Measurement of the ringing sound of Sacred Bell of Great King Seongdeok using multi-channel and an acoustic camera

Youngsun KIM\(^1\); Mina KIM\(^2\); Jiyoung OH\(^3\); Jeongho JEONG\(^4\)

\(^1\) Graduate School of Seoul National University, Seoul National University, Korea  
\(^2\) Diplom Tonmeister, MBI, Korea  
\(^3\) Sound Engineer, Sgosda, Korea  
\(^4\) Fire Safety & Building Environment Research Center, Korean Fire Protection Association, Korea

ABSTRACT

The Sacred Bell of Great King Seongdeok is a cast bronze bell for Buddhist temples, completed in AD 771 measuring 3.66 m in height, 2.27 m in diameter (at the rim), 11-25 cm in thickness (of the wall), weighs 18.9 tons, and an acoustical cylinder particularly found in Korean bells is on top. This research project employed multi-channel microphones at several points around the bell for measuring the variations of sound pressure levels over the time at each frequency band from the moment of tolling the bell and changes in the central axis of sound pressure was observed with an acoustic camera.

Commissioned by Gyeongju National Museum, the measurement was conducted as a part of the examination of the bell, so the ringing was taken with much less force than ordinary. As a result of the experiments, it is observed that the audible lasting time of the bell is 1 minute and 53 seconds, and the central axis of sound pressure moves over time.

Keywords: Korean Bell, Pavillion, Acoustic camera, Multi-channel microphone measurement

1. INTRODUCTION

The Sacred Bell of Great King Seongdeok was completed in 771 by King Hyegong of Silla as the largest bell in Korea. This bell is hung on a pavilion and makes a sound by bell tolling with a bell pounder by beating the spot called “Dangjwa,” decorated with lotus patterns.

The first place to be dedicated and installed was Bongdeoksa temple in Gyeongju, but it was moved several times when Bongdeoksa temple disappeared due to flood damage in 1460. Currently, it has been displayed at the Gyeongju National Museum since 1975. Unfortunately the original design of the pavilion of Bongdeoksa temple has not remained. For the design of the pavilion and the hemispherical cylindrical cavity system on the ground, a more specific academic basis is needed to restore the original construction.

2. CHARACTERISTICS

2.1 Shape

Measuring 3.66 m in height, 2.27 m in diameter (at the rim), 11 ~ 25 cm in thickness (of the wall), weighs 18.9 tons. An acoustical cylinder on top and a resonator at the bottom of the ground appear particularly in Korean bells as tradition.
The acoustic characteristic of The Sacred Bell of Great King Seongdeok yields 64.06 Hz as the lowest frequency, and 64.38 Hz is generated at the same time which yields beat. Table 1 shows the beat frequency of each natural frequency and the number of modes in each direction.

Table 1: Natural frequencies of the Sacred Bell of the Great King Seongdeok (Kim S-H, Jeong W-T, Kang Y-J, 2012)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Mode (m, n)</th>
<th>Frequencies (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0, 2)</td>
<td>64.07</td>
</tr>
<tr>
<td></td>
<td>(0, 2)</td>
<td>64.42</td>
</tr>
<tr>
<td>2</td>
<td>(0, 3)</td>
<td>168.52</td>
</tr>
<tr>
<td></td>
<td>(0, 3)</td>
<td>168.63</td>
</tr>
<tr>
<td>3</td>
<td>(1, 2)</td>
<td>189.34</td>
</tr>
<tr>
<td></td>
<td>(1, 2)</td>
<td>190.55</td>
</tr>
<tr>
<td>4</td>
<td>(1, 3)</td>
<td>227.99</td>
</tr>
<tr>
<td></td>
<td>(1, 3)</td>
<td>228.34</td>
</tr>
<tr>
<td>5</td>
<td>(0, 4)</td>
<td>281.93</td>
</tr>
<tr>
<td></td>
<td>(0, 4)</td>
<td>282.61</td>
</tr>
<tr>
<td>6</td>
<td>(1, 4)</td>
<td>349.4</td>
</tr>
<tr>
<td></td>
<td>(1, 4)</td>
<td>350.1</td>
</tr>
</tbody>
</table>
2.3 Duration of the audible bell sound

As a result of the measurement at dawn (03:20 AM) on October 28, 2020, the duration of the audible bell sound was 1 minute 53 seconds, and the background noise was about 30 dB at 500Hz, as shown in Figure 3. As the purpose of the test was to examine whether the bell was damaged, the intensity of the ringing was generally fragile compared to the intensity of striking the bell. As a result, the measured audible ringing time is relatively short.

![Figure 3. Background noise in the space around the pavilion](image)

2.4 Reverberation time of the pavilion

The reverberation time of the pavilion is shown in Figure 4. As the reverberation time of the pavilion, the rapid increase in the reverberation time below 125Hz at T20 is presumed to be due to the resonance of the bell.

![Figure 4. The reverberation time of the pavilion of the Sacred Bell of Great King Seongdeok](image)
2.5 Acoustic camera measurement results

In order to examine the movement path of sound energy, an acoustic camera was used to take pictures, and the results are shown in Figure 4. There is no change in position from 4 sec. to 32 sec. of the ringing. After 32 sec. the peak energy location changes, but the exact cause is a matter that requires further research.

Figure 5. Peak energy location movements over time after the bell tolling
3. DISCUSSION

The Sacred Bell of Great King Seongdeok is the largest in the Shilla period in Korea. The acoustic function of the sound tube is known as the resonance transmission and the high-frequency sound attenuator, and it is known that it has a function to increase the amplification to sound last longer in time by using hemispherical cylindrical cavity system on the ground.

This study introduces the central movement of negative energy shown in acoustic cameras, and it is considered that research on the structure of the species, the design of the pavilion, and the direct relationship with sound is needed.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Culture, Sports, and Tourism of Korea through Gyeongju National Museum.

REFERENCES

1. Seong, Myeong-Je, Haan, Chan-Hoon, Presentation of the Restoration of a Bell Pavilion of King Song-Dok Bell, The Regional Association of Architectural Institute of Korea, Fall conference [Internet], 2007:p.386-391.
2. Do, hakhoe, The Study on Bell Head of Sangwonsa bell and The Bell of King Sungdeuk, Korea Art Education Association [Internet], 2011;25(3):289-314.
A Custom Feature Set For Analyzing Historically Informed Baroque Performances

Nolan ELEY\(^{1,2}\), Théodora PSYCHOYOU\(^{3}\), Catherine LAVANDIER\(^{1}\), Brian F.G. KATZ\(^{2}\)

\(^{1}\)ETIS Laboratory, CY Cergy Paris University, ENSEA, CNRS, France, nolan.eley@dalembert.upmc.fr
\(^{2}\)Sorbonne Université, CNRS, Institut Jean Le Rond d’Alembert UMR7190, Paris, France, brian.katz@sorbonne-universite.fr
\(^{3}\)Sorbonne Université, Institut de Recherche en Musicologie UMR 8223, France

ABSTRACT
A common strategy for analyzing music performances is to extract a high number of low-level features followed by dimensionality reduction. While this can be an effective way of identifying differences in performances, interpreting these differences in a musically meaningful way is difficult. An alternative strategy is to select or design features which are appropriate for the task at hand, however, this must be done carefully and in accordance with musicological principles. In this study, a custom feature set was designed to capture baroque expressiveness in music performances, motivated by principles of historically informed performance practice. This feature set was verified on a small set of annotated performances then applied to a larger group of recordings of historically informed solo baroque performances in two halls, one historically appropriate and one modern. These features were examined to study the effect of room acoustics on the baroque expressiveness of the performances in order to gain insight into the role of acoustics in historically informed performances.

Keywords: Music performance analysis, chamber music, room acoustics, historically informed performance

1 INTRODUCTION
There is often a large variety among expressive musical performances of a single composition. The field of music performance analysis is concerned with systematically identifying and assessing these variations. There have been many strategies for analyzing different aspects of musical performances, which are well described in \[1\]. Because each genre may have a different set of intrinsic mechanisms which contribute to an idealized performance within that genre, it is important that the analysis framework take into consideration the musicological context of the task at hand.

One way to address this is through careful selection of low-level features. However, the available low-level features may be limiting and often lack a direct musical meaning. Furthermore, it is not always known which features will be most discriminative and one runs the risk of leaving out meaningful features whose capabilities may be currently unknown. Another strategy, therefore, is to extract a large number of low-level features then apply dimensionality reduction, followed by statistical analysis \[2, 3\]. While this can be effective at identifying similarities and differences among performances, one disadvantage is that the resulting dimensions can often be difficult to interpret, especially in a way that is musically meaningful. One way to overcome these problems is to design custom features for identifying specific musical parameters which are known to be of interest to the current task, based on musicological principles.

2 CONTEXT
There are many factors which can affect musical performance, one of which is the acoustic environment \[4\]. This study takes place as part of a larger effort to study the interactions between acoustics and musical performance from a historical perspective, merging the study of acoustics with the fields of historically informed...
performance (HIP) and historical musicology [5, 6]. More specifically, this study focuses on baroque HIP and whether the acoustics of spaces of the era facilitate HIP musicians’ play. To better understand this problem, it is helpful to know which musical parameters are most salient in the genre. Through the study of period musical treatises and manuals, musicologists have identified a number of characteristics which are responsible for defining baroque performance style and serve as the basis for HIP [7, 8, 9].

Musicologists have adopted the term “stylishness” as a way to describe musical expressiveness in terms of its appropriateness for a specific musical and historical context [10]. For example, the parameters which define an expressive performance of 19th century romantic music are very different from those which define an equally expressive performance of 18th century baroque music. While both performances may be perceived as equally stylish, their perceived expressiveness may vary widely due to this term’s strong association with mainstream romantic musical gestures. For this reason, researchers have started utilizing the term “baroque expressive” in order to differentiate a type of expressive performance which is deemed as being baroque stylish [11].

The goal of this paper is to design an analysis framework for identifying the parameters which make up baroque expressiveness among other performance styles of baroque music. This framework will be verified against a small set of recordings which have already been authenticated as representing distinct performance styles (see Section 3). Following this verification, the designed features will be applied to a set of recordings from a previous study [6] to shed light on the role of acoustics in facilitating baroque HIP practice (see Section 4).

3 FEATURE DESIGN AND VERIFICATION

3.1 Dataset

Three distinct schools of baroque performance practice are identified in [11], and we refer the reader to that paper for more information. The three schools are referred to as expressive-emotional, modern-literalistic, and baroque-expressive. In that study, listeners were able to identify these three schools of performance as evidenced by ratings in several musical categories. Relying only on the musical example used in that study (the first 8 bars of J.S. Bach’s Sarabanda from the D minor Partita for solo violin) to verify our analysis framework would not suffice as that would not provide enough data to yield meaningful results. So, the number of recordings was expanded to three full solo violin pieces within each performance school. These pieces, all composed for solo violin by J.S. Bach, were the aforementioned Sarabanda, the Gavotte en Rondeau from the E major Partita, and the Largo from the C major Sonata. While these additional recordings have not been verified by listeners as embodying specific performance parameters, it is assumed that additional recordings by the same musicians would carry similar performance characteristics to the one previously examined.

Below follows the development of features to identify baroque expressiveness in three major areas which listeners have already identified in [11], namely, phrasing, tone quality, and vibrato. The verification for each feature set is done using support vector machines (SVM) using a one versus one approach with cross validation (10 random folds) to predict the musician which serves as a proxy for performance style.

3.2 Phrasing

The phrasing style in a baroque expressive performance is often described as rhetorical, since it tends to mimic the patterns found in the art of oratory, in that it is varied, locally nuanced, and rhythmically flexible [12]. By contrast, other performance styles tend to use legato, or continuous phrasing, in order to highlight cadences and large phrase boundaries.

Researchers have found that phrasing is closely related to changes in tempo and loudness, and correlations are typically found between these two musical parameters [13, 14]. Our methodology for capturing phrasing was to separate the recordings into segments of different lengths (1, 2, 4, 8, and 16 bars) and extract features from the tempo and loudness curves of these segments as well as the entire recordings. In total, the dataset yielded 1029 segments. The tempo curves were made up of the note-level tempo values smoothed with a moving average filter since smoothed tempo curves have been found to be helpful in modeling expressive timing [15, 16]. The loudness curves were calculated using the frame-wise root-mean square (RMS) with a window size of 0.05s and a hop size of 0.02s. The features extracted from these curves are range, standard deviation
(SD), and coefficients of a parabolic function (2nd-order polynomial) which have been shown to be effective in modeling expressive performance [17, 18].

Using these features as input to an SVM tasked with predicting the musician, we were able to achieve an average F-score (the geometric mean of precision and recall) of 72%. As further exploration, the same classifier was used to predict the composition, and this achieved an average F-score of 87%. As the different compositions are stylistically different, they should exhibit very different phrasing characteristics, meaning this F-score may reflect somewhat the upper bound of the discriminability of this feature set.

3.3 Tone Quality
Tone quality or timbre is multidimensional and complex. One dimension of tone quality which was found to be representative of baroque expressive play is “lightness of tone.” Our methodology for identifying this tone quality was to use mel-frequency cepstral coefficients (MFCC) on the harmonic and percussive components of the signal which were extracted using the median filter approach [19]. This approach applies a median filter to the spectrogram either across successive bins or frames which enhances and suppresses either the harmonic or percussive components of the signal. A typical use case for such an algorithm might be to separate the drums in a pop song to facilitate remixing. However, when applied to solo instrumental music, the percussive components end up highlighting the nontonal, transient, and stochastic aspects of the signal which translate to elements such as bow and articulation technique which are likely correlated with one’s perception of tone intensity [20]. Additionally, the nontonal spectrum has previously been identified as being particularly discriminative in instrument classification [21]. MFCCs were extracted from these harmonic and percussive signals in addition to the original, unprocessed audio. The MFCCs were calculated frame-wise using a window size of 0.04 s and a hop size of 0.01 s, and 21 coefficients were used.

To quantify the effectiveness of these various MFCC groups in identifying the performer, four classification tasks were considered. One using each the percussive, harmonic, and unprocessed audio, and one which used the MFCCs from both the harmonic and percussive audio. Using the unprocessed audio, the classifier was able to achieve a class-weighted average F-score of 68%. Using the harmonic and percussive signals improved the classification to 72% and 77%, respectively. And finally, using MFCCs from both the harmonic and percussive signals together achieved the highest class-weighted average F-score of 83%. These results suggest that the harmonic, and especially the percussive components of the signal can be more discriminative than the unprocessed signal, although more testing is needed to see how robust these components are for different types of signals.

3.4 Vibrato
Vibrato is a musical ornamentation made up of semi-periodic fluctuations of pitch and is often described by two features: rate and extent [22]. The rate refers to the frequency of modulation which is typically in the range of 4 Hz to 8 Hz [23], while the extent refers to the depth or intensity of the vibrato, measured in cents. Baroque expressive performances tend to use less intense vibrato, less frequently than mainstream performance styles.

Most methods calculate extent in reference to a pitch center, usually the mean [24, 23, 25, 26]. However, this assumes that the fundamental frequency remains unchanged throughout the duration of the note. While this may be the case in controlled laboratory experiments and therefore would not pose a problem to those examining vibrato parameters under such conditions [27], in real performance situations, performers do not always start and end each note at the precise pitch center and may adjust their intonation while maintaining vibrato. One previous attempt to control for this used linear regression to model the pitch center [28].

The proposed methodology aims to improve upon previous methods in two ways. First, the mean removal step is modified by modeling the change in fundamental frequency over the duration of the note with a polynomial curve, aiming to improve the accuracy of the measured vibrato extent. Second, it models the evolution of the vibrato rate as it changes throughout the duration of the note using polynomial fit coefficients, rather than relying on a single measure (such as the median) for the vibrato rate.

The first step in the process is to extract pitch contours [29]; this was done using the Melodia plug-in¹ for Sonic Visualiser². Pitch contours are continuous curves of varying length and can be output as multi-pitch

¹https://www.upf.edu/web/mtg/melodia
²https://www.sonicvisualiser.org/
representations giving them an advantage to other fundamental frequency algorithms such as pYin \[30\] when applied to polyphonic contexts. The pitch contours were separated by note using note onset information. In cases where multiple contours were detected during the span of one note, the best pitch contour was chosen by summing the energy in each bin (representing 10 cents) and then selecting the bin with the most energy. This bin, along with the six bins above and below (equivalent to a range of 130 cents) were selected for further processing (see Fig. 1b).

Because the raw contour was quantized to the nearest 10-cent bin (see Fig. 1c), the contour was upsampled, smoothed using a moving average filter, and factored by a gain of 10 so that each bin would represent one cent (see Fig. 1d). The presence or lack of vibrato was detected if the signal met three conditions: a minimum length, zero crossing rate (on the pitch-centered signal), and extent. These thresholds were arrived at empirically through preliminary tests.

If the presence of vibrato was detected the pitch center was approximated using a polynomial curve of degree 5 (see Fig. 1e). This pitch center was then subtracted from each value, resulting in the final signal seen in Fig. 1f from which the vibrato parameters were extracted. The extent was calculated by taking the mean of the doubled absolute value of each peak and trough. The rate was calculated frame-wise by taking an auto-correlation of the signal followed by peak picking with a peak distance threshold set to stay within the typical rate of vibrato. The median value across all frames was selected as the rate. Polynomial fit coefficients
Table 1. Vibrato characteristics for each performance style.

<table>
<thead>
<tr>
<th>Style</th>
<th>Prevalence</th>
<th>Rate (mean &amp; SD)</th>
<th>Extent (mean &amp; SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expressive-emotional</td>
<td>32%</td>
<td>7.1 Hz ±0.75</td>
<td>16.8 cents ±5.8</td>
</tr>
<tr>
<td>Modern-literalistic</td>
<td>22%</td>
<td>7.3 Hz ±0.73</td>
<td>19.1 cents ±7.3</td>
</tr>
<tr>
<td>Baroque-expressive</td>
<td>13%</td>
<td>6.8 Hz ±0.87</td>
<td>12.7 cents ±3.9</td>
</tr>
</tbody>
</table>

of order 3 were taken of the frame-wise rate to model the evolution of the rate across the duration of the note.

In order to quantify the discriminability of these proposed vibrato features, four classification tasks were considered. Two tasks used mean removal while the other two used the polynomial curve modeled pitch center removal. These two groups either used only the baseline features (rate and extent) or the baseline features in addition to polynomial fit coefficients. Overall, 972 notes with vibrato were identified out of 4374 played notes.

In general, the different classification tasks did not yield very different results. Using the baseline features (vibrato rate and extent) yielded class-weighted average F-scores of 47% and 49% using the mean and polynomial curve as pitch center estimation, respectively. Adding the additional features meant to model the change of rate yielded class-weighted average F-scores of 46% and 52% using the mean and polynomial curve as pitch center estimation, respectively. These F-scores are all significantly above random chance (33%) and, overall, the proposed changes in methodology and additional features yielded an approximately 5% increase in classification accuracy.

Various vibrato characteristics are shown in Table 1 and reveal that, as expected, the baroque-expressive style exhibits the least amount of vibrato, in terms of prevalence (the ratio of notes with vibrato to those without), rate, and extent. The expressive-emotional style predictably shows the most prevalent vibrato, and both it and the modern-literalistic style show somewhat comparable rates and extents.

4 APPLIED ANALYSIS

The dataset that this analysis framework was developed for was part of a larger study intended to help understand the role of acoustics in historically informed performance. A broad overview of the study can be found in [6], but a brief description follows. Musicians specializing in historically informed baroque performance practice played in two rooms, one historically appropriate to the musical era, and one modern. The historical room is the Salon des Nobles in Versailles, and the modern hall is the amphitheater from the Cité de la Musique in Paris. 10 musicians participated, 4 violists da gamba, 3 flutists, and 3 theorists. The musicians each played 3 pieces of music which were chosen for each instrument by musicologists, and they repeated each piece around 3 times in each room, rendering a total of more than 180 recordings. The same classification exercises used to verify the features were used when applying the analysis framework to these performances, but instead of being tasked with predicting the musician, they were tasked with predicting the room in which the performances occurred.

Using the phrasing set of features yielded average F-scores of 90%, 95%, and 83% for the viol, flute, and theorbo, respectively. While this is fairly high, it is possible that some of the features which make up the phrasing may have been influenced by the sound of the room captured by the microphone. Even though each instrument was recorded with a directional microphone situated 1 m away, resulting in a high ratio of instrument sound compared to room sound, the sound of the room can still be identified in most recordings. Some of the features which make up the phrasing set are related to the RMS level, which could be influenced by this incidental room sound. It is difficult to tell exactly which identified differences are due to this room sound and which are due to actual differences in the performance. In order to clarify this, the classification task was repeated on these three instruments using only the tempo-related features in the phrasing set. This yielded average F-scores of 64%, 73%, and 53%, respectively. This indicates that at least some of the phrasing expressiveness was changed due to the room, at least for the viol and flute. It was previously reported in [6]...
that the performances by the theorists included many mistakes, so it is not surprising that the results using only tempo-related features resulted in an accuracy that is only slightly above random chance.

The vibrato analysis is limited to the viol since it is the only instrument used in the study which is capable of vibrato. In total, 1145 notes were identified with vibrato out 6888 played notes, meaning almost 17% of the played notes were embellished with vibrato, compared to the 22% in the test set of Bach recordings. The results of the classifier using the vibrato feature set performed no better than random chance, indicating that the vibrato style did not change as a function of the room. This is not very surprising as it was the least discriminative feature in the test set. Despite the poor performance of vibrato as a predictor of room, some interesting findings can still be reported. 54% of the vibrato occurred in the Salon des Nobles. Furthermore, about 30% of the vibrato can be attributed to one of the four violists who had 3 years of performing experience, compared to an average of 37.3 (SD: 5.5) years of experience the other 3 violists had. This implies that perhaps the more liberal use of vibrato may be correlated with having less experience or different training. Lastly, the average vibrato rate was 6.8 Hz and the average extent was 15.2 cents.

In general, the features designed to describe tone quality were able to predict the room with fairly high accuracy, with class-weighted average F-scores among the 90% range using the normal, harmonic, or percussive components of the signal. There was not much of a difference between the unprocessed and the harmonic signals, while the percussive signal resulted in an F-score of about 5-7 points lower. However, it should be noted that these MFCCs are likely influenced by the room sound to some degree though it is difficult to say precisely how much. The percussive signal, with its emphasis on transients, is likely influenced by the room less, and therefore, its slightly lower F-scores may more accurately reflect an actual difference in tone quality.

5 Conclusions

The goal of this study was to simplify the analysis of music performance to a few musically meaningful dimensions which are relevant to the task at hand, in this case, historically informed baroque performance. We identified three major areas which are important to baroque expressiveness, namely, phrasing, vibrato, and tone quality, and developed an analysis framework for identifying differences in these areas. This framework was first successfully verified on recordings from a published study in which musicologists and listeners had identified significant differences among baroque performance styles.

Subsequently, this analysis framework was applied to an unrelated set of recordings of baroque performances, with the goal of identifying significant differences among the performances which may have been influenced by the room and its acoustics. Judging the success of the analysis framework on this dataset was difficult, as it was not previously known how these performances differed, nor to what degree. Overall, the phrasing and tone quality features identified significant differences between performances in the two rooms. However, it was noted that some of this may be due to the sound of the room incidentally picked up by the microphone. The vibrato feature set did not show any significant differences between the two rooms, however, it did reveal a difference between performers which may be correlated with their experience or training.

While this method of performance analysis was able to reduce the problem to a few components, these components are still described by high-dimensional feature sets which exhibit problems of interpretability. More work (and larger datasets) would be needed to reduce the dimensionality of these feature sets in a way that makes the results easier to interpret. For example, we are able to identify differences in phrasing, but cannot say definitively what this means on the continuum of “legato” to “separated”. A complementary listening test performed on a subset of these recordings may help identify what these differences are.

One issue with developing an analysis framework on a few recordings of Bach violin pieces is the major risk of over-fitting to such a small and unrepresentative dataset. Ultimately, without larger datasets of performances of baroque music annotated along specific dimensions of performance, it is difficult to create a truly robust analysis framework. Still, the approach outlined in this paper suggests that a tailored method of analyzing music performance can be effective and yield meaningful results that are more interpretable than broad methods of low-level feature extraction.
ACKNOWLEDGEMENTS

This work was supported by the Paris Seine Graduate School Humanities, Creation, Heritage, Investissement d’Avenir ANR-17-EURE-0021 – Foundation for Cultural Heritage Science. This project takes place as part of EVAA_Ver: Expérience Virtuelle en Acoustique Archéologique. We would also like to thank the research center of the Château de Versailles and the Musée de la Musique of the Philharmonie de Paris, as well as the musicians who participated in the study.

REFERENCES


Pipe organ buffet radiation patterns under different excitation strategies

Gonzalo VILLEGAS CURULLA\(^{(1)}\); Elliot K. CANFIELD-DAFILOU\(^{(1)}\); Piergiovanni DOMENIGHINI\(^{(2)}\); Benoit FABRE\(^{(1)}\); Christophe d’ALESSANDRO\(^{(1)}\); Brian F. G. KATZ\(^{(1)}\)*

\(^{(1)}\)Sorbonne Université, CNRS, Institut Jean Le Rond d’Alembert, F-75005 Paris, France
\(^{(2)}\)Università degli Studi di Perugia, Italy

**ABSTRACT**

Studying the directivity of musical instruments such as pipe organs is challenging because of their great size and because they typically cannot be moved from the locations where they are built to the laboratory. This study compares the radiation pattern of organ buffets under different excitation conditions, using a 19th century French pipe organ. We investigate the positive and the great organ independently due to their spatial separation and size. The excitation strategies were comprised of cylindrical and an omnidirectional electroacoustic sources. Measurements were carried out on each organ buffet along a horizontal line spanning the width of the church. Results in octave bands are shown and compared with a discussion on possible causes for observed differences. Significant variations in directivity were observed for the 2 kHz to 4 kHz octave-band regions where scattering from pipes is expected to have a predominant effect, while little variation from omnidirectional was observed for lower frequency bands for both sources.

Keywords: Pipe Organ, Source Directivity, Musical Acoustics

1 INTRODUCTION AND MOTIVATION

The interest of studying the radiation and directivity patterns of pipe organs is framed within the broader scope of modeling the acoustic behavior of individual components of pipe organs (e.g., pipes, windchests, buffets) both alone and in combination with one another. These goals are both theoretical—for understanding the underlying behavior of the instrument, and applied to case of modeling historical instruments such as the Dallery organ of Chapelle Sorbonne and the instrument in Notre-Dame by Thierry, Clicquot, and Cavaillé-Coll. An overview of the broader context of pipe organ modelling can be found in [1]. Results from recent campaigns of measurements and simulations are presented here as well as in [2, 3, 4, 5].

The current study considers the organ buffet as an acoustic obstacle (diffraction/scattering element). Horizontal directivity patterns of the positive and great organ of St. Élisabeth are presented when excited by idealized electroacoustic sources located within the buffets. These sources include an omnidirectional point-source (3D) and a linear-array cylindrical source (2D). These two approaches offer simplifications of the acoustic radiation of pipes within the organ buffets.

Different time scales are involved in the production and radiation of sound produced by pipe organs. Starting from the aero-acoustic sound production in each pipe, the sound is first radiated inside the buffet, where the characteristic time is given by \(L_{buf}/c\) (~5 ms in the current case), where \(L_{buf}\) is the characteristic length of the buffet (height of the facade’s slit) and \(c\) the propagation velocity of the sound. The next time scale is given by the room.

In a more detailed analysis, multiple regimes can be extended to the field inside the buffet with the additional aspects of: the establishment of modes, complex scattering due to the presence of hundreds or thousands...
of pipes of different heights, and the establishment of a diffuse incident field onto the facade from the inside, where the facade constitutes the interface between the interior and the exterior of the organ buffet. This facade can be initially described as an array of air-masses trapped between cylinders (pipes) that are brought to oscillation.

In [6], the explicit problem of an evenly-distributed array of cylinders is discussed. These cylinders can represent organ pipes with known diameters and center-to-center spacing in the facade, where a first-order estimation is proposed to predict the scattered pressure. In [2], a comparative study of the 1-slit system was proposed using a laboratory proxy organ buffet, with respect to theoretical models informing directivity and full-range SPL-vs-frequency about the slit and the cylindrical source used in the current study of the positive.

The presence of the facade is bound to contribute as a frequency dependent boundary condition, resulting in a reflection or transmission with frequency dependent behaviour and with small losses. As a first approximation, it may well be considered almost transparent in the low frequencies. In [4], indices are shown that the protruding towers of the facade of the positive (see Section 1.1) constitute a scattering contribution already for the region of middle center frequencies of 1kHz and 2kHz.

1.1 The Organ in St. Élisabeth, Paris

The organ in St. Élisabeth was dedicated on 28 April, 1853 and restored in 1999. It is a large historical organ, with a total facade area of approximately $10 \times 10$ m. The organ has three levels and hosts a total of 2322 pipes, of which 143 are visible in the facade. The organ has 42 stops on three manuals and pedal, with tracker action [7]. Figure 1 shows the locations of the three main radiating cavities and their wind-chest distributions.

The positive (420 pipes) stands on the floor of the organ loft, separated and in front of the main case and corresponds to the first manual keyboard. Its facade (inscribed in a $2.62 \times 2.64$ m surface) contains 33 pipes, organized in 3 towers of 5 pipes each, and 2 flat sets of 9 pipes. The tallest pipe in the facade is 6 feet, sounding F$_1$. It is relatively narrow and protrudes into the nave, radiating into free field without

Figure 1. The organ in St. Élisabeth, Paris and measurement schematics.
nearby reflective surfaces. Above the positive resides the main case 5.35 m above the floor level. It contains the great organ (GO) in the center, and the pédale on each sides behind the highest towers. The GO (1014 pipes) and pédale (180 pipes) correspond to the second manual keyboard and pedal board, respectively. Its facade is approximately 9.39 m × 9.6 m, and contains 75 pipes, organized in 7 towers and 8 flat segments. The tallest pipe in the facade is 12 feet (≈3.7 m), sounding F₀. The récit (708 pipes) crowns the instrument; it is enclosed in a swell-box and corresponds to the third manual keyboard. Its facade (inscribed in a 2.95 m × 3 m square) contains 35 mute pipes. The récit is not studied here.

The acoustics of the church show moderate reverberation times ($T_{20} = 2.14$ s to 3.68 s across the 125 Hz to 4000 Hz octave bands, see [8]).

1.2 Theory and Approximations

We make use of a number of approximations and assumptions in order to decompose the problem into smaller, less complex systems. Namely, we try to isolate and simplify the organ into studies on the sources, the pressure field inside the cavity, and the transmission and radiation into the building; the last of these being the subject of the current study. The sound-production sources are organ pipes with a non-flanged termination on one end (the so-called passive end of the resonator) and a mouth or window opening at the other end. The latter can be approximated in the low frequencies as a second monopole, but more elaborate characterizations are readily available: filtering has been described in [9] and general intensity around the pipe as a whole in [10]. The majority of pipes are assumed to be cylindrical and of open-open terminations. Each of them is not only a sound source, but also an obstacle to the propagation front of other sources within the buffet. In the condition of a fully closed buffet, the pressure field inside is mostly transmitted through the front face of the organ, the facade. The pipe-to-pipe spacings in the facade are the most salient transmission aperture outwards: they can be represented by an array of slits of resilient air-masses. Said slits are considered in the finite and infinitesimal thickness, and ideally as infinite in vertical length condition [6, 11, 12] or finite length, amounting to the full height of the cavity. These apertures become larger at the very bottom of the facade pipes due to the conicity of their foot, but this is considered irrelevant here. If acquisitions of the radiated field into the nave are made close enough to the facade, the infinitely tall slit [13] is a plausible approximation.

The effects of scattering will be expected to become apparent for $ka \gg 1$, where $a$ is the half-width of the spacing between facade pipes, and $k = \omega / c$ is the wavenumber. For comparability purposes, the cavity of the St. Élisabeth’s positive organ was chosen as a feasible approximation of the proxy organ used in previous laboratory studies [2, 3, 4].

2 Methods

The organ buffets under examination were excited by 6 s logarithmic sweeps spanning 60 Hz to 20 kHz, played from electroacoustic sources [14]. Signal acquisitions were made in several points inside the buffets for reference. The measurements outside the buffet were made along a set of equally-spaced locations in a straight line spanning the width of the nave (10 m). Acquisitions were made using the following audio hardware: RME UFX+, RME Octamic, omnidirectional measurement microphones BAM T1 class 1, and a Carver PM–175 Amplifier. The 3D source was a dodecahedron omnidirectional source, Dr. Three 3d-032. For the 2D cylindrical sources, the small 2D array was comprised of 18 Aurasound NSW2-326-8A speakers, with a 50 mm center-to-center spacing; the large 2D array was comprised of 10 Monacor SPH 165CP, with a 170 mm center-to-center spacing. The speakers were positioned near the center of the positive, and near the center of the left-half of the great organ. Groups of 4 to 5 microphones were arranged with 20 cm spacing on a transversal tension line, manually positioned using a pull-cord and measuring tape (see Figure 2(a)). See Figure 1(c) for source and receiver locations.

Deconvolution of the acquired signals was performed. Temporal alignment of the resulting IRs, compensating for arrival times, were calculated based on microphone positions along the acquisition lines, parallel to the measured facade (see Figure 1(c)), converting the linear measurement array to a polar propagation estimation relative to the source position. Octave-band frequency filtering of data was performed and relative pressure levels were calculated using fourth order Butterworth filters given by MATLAB’s octaveFilterBank. Subsequent
amplitude correction of time aligned IRs was applied dependant on a reference distance \( r'/r \), with \( r \) the shortest path at \( 0^\circ \). For the 3D omnidirectional source a \(-6\) dB per doubling distance was employed. For the 2D cylindrical source, the correction was more complicated, employing a simplified model of \(-3\) dB or \(-6\) dB for near or far-field corrections, based on an idealized finite array (see [15]) with a \( 2L_c^2/\lambda \) rule for the cut-off, where \( L_c \) is a characteristic length, here the height of the source. The IRs were windowed in time to avoid contributions from wall reflections. The length of this window was determined based on the distance between the microphones closest to the side-walls and the length of the direct path. An adequate estimation was determined to be 4.7 ms between the direct path to the microphone position closest to wall and its first reflection arrival, applied after time alignment of each position’s deconvolved IR. Results are presented for octave bands from 125 Hz to 4 kHz. Due to the microphone spacing (20 cm), spatial aliasing limits the interpretability of higher frequency data.

### 3 RESULTS AND DISCUSSION

The octave band directivity pattern results for the positive are show in Figure 3, and for the GO in Figure 4, for the two source excitation conditions (3D omnidirectional and 2D cylindrical). Note the increasing measurement point density with increasing angle, due to the projection of the linear measurement setup to polar coordinates. Results are first examined for each buffet separately, then cross-compared. Closer inspection of the positive microphone array showed a variance in calibration level of 0.87 dB (B&K calibrator type 4231); accounting for this would likely resolve the observed minor fluctuations. Similar variations were not observed for the grand organ microphone array.

#### 3.1 Positive Directivity

Inspection of the results for the positive identify a relatively omnidirectional directivity pattern across the octave band frequencies of 125 Hz to 1000 Hz (Figures 3(a) to 3(d)) for both 3D and 2D sources. Some minor attenuation can be observed at the extreme angles (\(|\theta| > 50^\circ\)), likely due to the enclosure effect of the buffet windowing the energy generated within the buffet, redirecting it to the more central region. Patterns are relatively smooth, with measurement variations on the order of a few dB.

Directivity in the 2000 Hz octave band (Figure 3(e)) shows a significant broadening of the directivity for the 3D source, enforcing radiation at \( \pm 45^\circ \) relative to on-axis. This effect is less pronounced for the 2D source, where the directivity is more erratic compared to lower frequency bands. This effect is attributed to the significant scattering effect of the pipes and facade interface.

Results for the 4000 Hz octave band (Figure 3(f)) are similar to those at the low frequency bands, though with higher variance. This variance is expected due to the effects of spatial aliasing at higher frequencies.
3.2 Great Organ Directivity

Directivity patterns for the octave band frequencies of 125 Hz to 500 Hz (Figures 4(a) to 4(c)) show relatively omnidirectional directivity for both source types, with pattern fluctuations occurring at the extreme angles. The slight increase at the extreme (θ ≥ 40°) could be attributed to the contribution of minor reflections from the
side wall at such proximity to the boundary. At the other extreme positions ($\theta \approx -60$), one can observe an increase in energy relative to the on-axis response. For 125 Hz to 250 Hz, the 2D source directivity exhibits a discontinuity, likely attributed to the idealized cylindrical propagation correction for mapping the measurement points from their linear configuration to polar coordinates relative to the source position (see Section 2). A smoother, more sophisticated estimation of propagation properties for the truncated linear source array could improve results, though it is noted that the corrected values asymptotically converge to the 3D source response in those frequency bands. A general increase in energy is seen with negative angle progression (approaching 5 dB), more pronounced for 125 Hz to 250 Hz with the 3D source and 500 Hz with the 2D source.

Results for the 125 Hz to 1000 Hz octave bands (Figures 4(a) to 4(d)) show increasing emphasis for the extreme negative angles, (on the order of 5 dB). It is worth noting that the GO has horizontal platforms separating levels. It is plausible that with increasing angle, and hence increased distance to the receiver array, reflections from these floor/ceiling surfaces arrive with diminishing delays relative to the direct sound, thereby becoming more prevalent within the 4.7 ms analysis window. As this behavior appears common to both sources, this seems a reasonable assumption.

Results at the higher octave bands, 2000 Hz to 4000 Hz (Figures 4(e) to 4(f)), exhibit large oscillations in directivity patterns for both sources, though the patterns do not align at all angles. It would seem plausible that such variations in directivity could be attributed to the scattering properties of the large pipes making up the GO in general.

3.3 Positive and Great Organ Directivity Comparison
Comparing the results between the two measured organ buffets (the positive and GO), similar behavior can be observed at the lower frequency regions, with markedly less measurement variance (noise) for the GO directivity patterns. This observation holds true for both sources, for which it is noted that the 3D was the same source while the 2D source was different due to the different dimensions/scale of the two organ buffets. For both buffets, the 2 kHz octave band showed the most pronounced deviations from free-field, omnidirectional directivity patterns, with patterns exhibiting increased variance at higher frequencies. The observed differences in these patterns between 3D and 2D sources suggest that the different excitation methods (spherical versus cylindrical waves) as well as the slight positional differences within the buffets and of the host of interior scatter objects induced different but still pronounced patterns.

While the GO contains more larger-diameter, low frequency pipes than the positive on the whole, examination of the two facades shows very similar pipe diameters in the center region (see Figure 1(b)), which would lead to an expectation of similar frequency behavior with regards to scatter properties. However, while these patterns were clearly evident for the GO in the 2 kHz to 4 kHz octave bands, they are only observed in the 2 kHz octave band for the positive. The complex response of the positive could be attributed to the generally small pipe diameter and spacing within the buffet, producing scattered patterns with spatial resolutions exceeding the current measurements spatial aliasing limits. Reproducing these measurements with a finer spatial resolution would allow further investigation of this phenomenon at these frequency ranges.

4 CONCLUSIONS AND FUTURE WORK
This study has presented measurement results of an investigation into the diffraction/scattering effects of the organ buffet and facade on the directivity pattern of the radiated sound field. Two ideal sound sources were employed, an omnidirectional point source and a linear (truncated) cylindrical source, as approximations of an organ pipe source. Two sections of the church organ at St. Élisabeth were studied, the small and compact positive and the larger great organ. Measurements were made linearly, spanning the nave of the church at a distance of roughly 1.4 m to 2 m, and subsequently processed to obtain polar directivity data, excluding contributions of wall reflections through short time windowing.

Octave band results highlighted generally omnidirectional directivity in the frequency range of 125 Hz to 1000 Hz. Prominent variations were observed in the 2 kHz octave band, attributed to the diffraction and scattering effects of the pipes comprising the interior and facade of the organ. Similarity between the two organ sections in frequency appears reasonable as similar pipe dimensions are present in both the positive and great
organ, despite the presence of larger pipes in the great organ. Differences in the details of the pattern are expected due to differences in the pipes distribution patterns within the organ.

While some differences observed between the two excitation sources can be attributed predominantly to assumptions regarding corrections for near-field versus far-field propagation correction with the cylindrical source, these results suggest the source model is not a predominant factor in the directivity behaviour. Other variations are attributed to slight positional differences (not precisely coincident) of the two sources within the organ.

It is interesting to note that the dissymmetry in the GO patterns, due to the lateral position of the source excitation (particularly in the 1 kHz octave band), is consistent with the auditory perception of the diatonic disposition of the instrument. In St. Élisabeth, as is typical in large instruments, the wind-chests are diatonic, divided in C and C# sides. This means that half of the notes are on the left (in whole tones, C - D - E etc.) and half on the right (C# - D# - F etc.). This is contrary to, for instance, piano strings which are chromatically ordered (in semi-tones C - C# - D - D# etc.). This ping-pong effect is clearly audible (when one pays attention to it) and is part of the organ listening experience. Here, the sources were positioned on the C side of the organ. Directivity patterns suggest that the buffet may emphasize this effect by enhancing lateral perception.

Further studies shall examine the case using real organ pipe excitation sources within different regions of the organ, rather than the idealized electroacoustic sources in the current study, providing a more realistic context for the results [5]. These results could also be further enhanced by analyzing in finer frequency resolution, such as 1/3-octave bands and by exploration with finite difference and finite volume scheme simulations [16, 17], well-suited to scattering and diffraction problems.

Higher spatial resolution would be necessary to provide further conclusions regarding higher frequency bands. Such high resolution measurements have been carried out on a proxy organ, comparable to the positive of the current study, comprised of a homogeneous collection of simulated pipes and using the same sources [3, 4]. These studies examined in more detail the effect of pipe density in an attempt to simplify the diffraction/scattering problem.

The perceptual relevance of various organ properties in the context of auralizations is a primary application of the current results. Directivity perception is currently being evaluated through a series of listening tests, followed by a study which will examine the number of spatially distributed virtual sources required for plausible-natural auralization renderings of the largest indoor musical instrument [18]. Such studies compliment previous works investigating the applicability of spatial perception metrics, such as the inter-aural cross correlation, as it pertains to the perceptual notion of apparent source width of such and extended instrument [8].

ACKNOWLEDGEMENTS

Funding for G.V.C. has been provided by the ISCD (grant no. FED 3 – 2019/7/2). Funding for E.C.D. and B.K. has been provided by the European Union’s Joint Programming Initiative on Cultural Heritage project PHE (The Past Has Ears, phe.pasthasears.eu, Grant No. ANR-20-JPIC-0002), and the French project PHEND (The Past Has Ears at Notre-Dame, Grant No. ANR-20-CE38-0014, phend.pasthasears.eu, Grant No. ANR-20-CE38-0014).

REFERENCES


Investigation of relationship between beat-timing and acoustics-parameter for method of beat-timing and tempo estimation

Madoka OKEMOTO¹ and Masanobu MIURA²

¹ Graduate School of Kunitachi College of Music, Japan
² Kunitachi College of Music, Japan

ABSTRACT
Research on beat-timing and tempo estimation has been studied in the field of musical acoustics. Nowadays, researchers are trying to use deep learning technology to explore the estimation of beat-timing and tempo. It is hard to explain the result of deep learning since it is an obscure method of factors. However, research on beat-timing and tempo is deep involvement with human perception and cognition of beat-timing and tempo. A computational method of beat-timing and tempo by explaining the factor of understanding the results can contribute to the investigation into a human's perception and cognition of beat-timing and tempo. The authors will investigate the relationship between beat-timing and acoustics parameters for beat-timing and tempo estimation. Beat-timing is given as timing data of classical music labeled by a student at the Kunitachi College of music graduate school in Japan. Acoustics parameters were calculated with the length of the time window by one-hundredth of a second. The results suggest that the most significant acoustic-parameter is RMS.

Keywords: beat-timing, acoustic-parameter, deep learning, beat-timing estimation, LSTM

1. INTRODUCTION
In this report, "beat" refers to the beat in music. The beat in music is a unit of musical rhythm. The beat-timing is treated as the switching time between beats when a piece of music is divided by beats. A musical term similar in nuance to beat, the tempo is a measure of the speed of the music and represents the number of beats present in a minute.

Methods for estimating beat-timing and tempo have been proposed in the field of musical acoustics. Nowadays, deep learning methods have been proposed to solve this problem. It is hard to explain the result of deep learning since it is an obscure method of factors. However, research on beat-timing and tempo is deep involvement with human perception and cognition of beat-timing and tempo. A computational method of beat-timing and tempo by explaining the factor of understanding the results can contribute to the investigation into a human's perception and cognition of beat-timing and tempo.

2. AIMS
In this report, we focus on beat-timing, investigate the relationship between beat-timing and acoustic-parameters, and examine the acoustic-parameters that are significant for beat-timing estimation.

3. METHOD
The first method is to confirm the correlation between beat-timing and acoustic-parameters. The
second method uses an LSTM (Long Short-Term Memory) network [1] to examine the estimation results when each acoustic-parameter is predicted as an explanatory variable and beat-timing data as an objective variable.

4. USED DATA

The beat-timing data used in this report is the beat-timing data obtained by labeling beat times by students of the Graduate School of the Kunitachi College of Music, and then using the data to assign a flag of 1 if it is a beat and 0 if it is not a beat for every one-hundredth of a second.

The acoustic-parameter data in this report are the data of RMS (rms), zero-crossing (zcs), spectrum-centroid (cnt), and spectrum-flux (flx), which are calculated every one-hundredth of a second.

The music pieces for which beat-timing data were labeled and acoustic-parameters were calculated were taken from 36 classical music pieces.

5. RESULT

5.1 The correlation between beat-timing and acoustic-parameters

Figure 1 shows the correlation between beat-timing data and acoustic-parameters data. Figure 1 confirms that all four acoustic parameters have a low correlation with beat-timing data. Although all correlations are low, the correlation between rms and zcs tends to be high among these four.

![Figure 1 - The correlation between beat-timing data and acoustic-parameters data](image)

5.2 LSTM network to examine

This section describes the estimation results when using an LSTM network to predict each acoustic parameter as an explanatory variable and beat-timing as an objective variable. The feedback of the LSTM network is 50. Since it is 50 in one-hundredth of a second division, it is 0.5 seconds in time.

Table 1 shows the classification results using the LSTM network. Table 1 summarizes the classification results into True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). Table 2 shows the values of Accuracy, Precision, Recall, Specificity, and F-measure calculated from Table 1.

Table 2 confirms that the highest F-measure values are obtained when rms is used as the explanatory variable. Although Precision is highest for cnt, F-measure tends to be low because Recall is low.

Table 1 - The classification results using the LSTM network

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>rms</td>
<td>390</td>
<td>34</td>
<td>6644</td>
<td>390456</td>
</tr>
<tr>
<td>zcs</td>
<td>248</td>
<td>20</td>
<td>6786</td>
<td>390970</td>
</tr>
<tr>
<td>cnt</td>
<td>198</td>
<td>11</td>
<td>6836</td>
<td>390979</td>
</tr>
<tr>
<td>flx</td>
<td>236</td>
<td>26</td>
<td>6798</td>
<td>390964</td>
</tr>
</tbody>
</table>
Table 2 - Accuracy, Precision, Recall, Specificity, and F-measure calculated from Table 1

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>rms</td>
<td>0.98320</td>
<td>0.91981</td>
<td>0.05544</td>
<td>0.99991</td>
<td>0.10459</td>
</tr>
<tr>
<td>zcs</td>
<td>0.98290</td>
<td>0.92537</td>
<td>0.03526</td>
<td>0.99995</td>
<td>0.06793</td>
</tr>
<tr>
<td>cnt</td>
<td>0.98280</td>
<td>0.94737</td>
<td>0.02815</td>
<td>0.99997</td>
<td>0.05467</td>
</tr>
<tr>
<td>flx</td>
<td>0.98286</td>
<td>0.90076</td>
<td>0.03355</td>
<td>0.99993</td>
<td>0.06469</td>
</tr>
</tbody>
</table>

6. CONCLUDING REMARKS

This report researched relationship between beat-timing and acoustic-parameter for method of beat-timing and tempo estimation. As a result, the results of 5.1 and 5.2 suggest that rms may be the acoustic-parameter most closely related to beat-timing data among the four acoustic-parameters. Future work will verify the effects of combinations of acoustic-parameter on the performance of beat-timing estimation.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP21J13539.

REFERENCE

timbre.fun:
A gamified interactive system for crowdsourcing a timbre semantic vocabulary

Ben HAYS(1), Charalampos SAITIS(2), György FAZEKAS(3)

(1)Queen Mary University of London, United Kingdom, [b.j.hayes, c.saitis, gyorgy.fazeaks]@qmul.ac.uk

ABSTRACT
We present timbre.fun (https://timbre.fun/), a web-based gamified interactive system where users create sounds in response to semantic prompts (e.g., bright, rough) through exploring a two-dimensional control space that maps nonlinearly to the parameters of a simple hybrid wavetable and amplitude-modulation synthesizer. The current version features 25 semantic adjectives mined from a popular synthesis forum. As well as creating sounds, users can explore heatmaps generated from others’ responses, and fit a classifier (k-nearest neighbors) in-browser. timbre.fun is based on recent work, including by the authors, which studied timbre semantic associations through prompted synthesis paradigms. The interactive is embedded in a digital exhibition on sensory variation and interaction (https://seeingmusic.app/) which debuted at the 2021 Edinburgh Science Festival, where it was visited by 197 users from 21 countries over 16 days. As it continues running online, a further 596 visitors from 35 countries have engaged. To date 579 sounds have been created and tagged, which will facilitate parallel research in timbre semantics and neural audio synthesis. Future work will include further gamifying the data collection pipeline, including “leveling-up” to unlock new words and synthesizers, and a full open-source release.

Keywords: Timbre, Synthesis, Crowdsourcing, Gamification, Timbre semantics

1 INTRODUCTION
The scientific study of musical timbre has historically focused primarily on understanding the perceptual phenomenon itself (28). This has variously involved seeking acoustic explanations for experimental data (13; 19; 33), studying its cross-modal associations (38; 15), and subjectively constructing systems for its description (25; 30). More recently, however, another approach has emerged which seeks to understand the relationship between timbre and sound production (15; 34) through prompted synthesis tasks. In these studies, participants are asked to produce a sound using a synthesiser in response to a prompt, such as a descriptive adjective. This yields a dataset of prompt-sound pairs which can be analysed to study the effects of various prompts on the sounds created.

This approach is of particular interest from the perspective of synthesiser control. The problem of providing meaningful controls for synthesisers has recurred in the literature since the late 70s (36; 27; 26), and has recently become of particular interest with the advent of complex neural audio synthesisers capable of being used in music production and sound design workflows (14; 21; 7; 2). However, the data collected in these studies have been limited in both scope and scale. Paired with the heterogeneity observed between timbral datasets (33), this has limited the usefulness of this data in such downstream tasks. To address these issues, we propose an approach for crowdsourcing timbre semantic data in a manner that is amenable to scaling and gamification. Through an exploratory analysis, we demonstrate that the data collected in a preliminary run of this study exhibit an emergent structure which is broadly congruent with the findings of prior timbre semantic research. We further observe, through the application of simple machine learning techniques, that the data collected in this manner hold sufficient predictive power to allow the affective connotations of semantic prompts to be classified from synthesiser parameters and acoustic features.

2 RELATED WORK

2.1 Prompted Synthesis
The standard paradigm for timbre semantic research involves listeners rating sounds along scales defined by descriptive adjectives (24). Controlling characteristics of the stimuli allows their perceptual influence on timbre semantic associations to be studied. This approach does not, however, provide insight into the inverse relationship: the influence of timbre perception and its semantic associations on the process of sound design. Despite significant effort having been invested into developing
adjective-controlled systems for these tasks (11; 12; 3; 32; 31; 9), this relationship has only recently begun to receive attention in the psychoacoustic literature (34; 15).

In the present authors’ prior study (15), experienced sound designers programmed an FM synthesiser in response to adjectival prompts, and rated their created sounds on semantic scales. In Wallmark, et al.’s study (34), classically trained musicians interacted with a simplified 2D control space mapped to FM synthesiser parameters. In both studies, clear structure emerged when the created sounds were analysed in terms of their acoustic characteristics, synthesis parameters, and relationships to prompt words. Both similarities to and deviations from the findings of conventional timbre studies were observed, suggesting that the prompted synthesis approach is a viable method for studying the specifics of the “inverted” relationship between timbre and sound design.

2.2 Crowdsourcing

Estellés-Arolas and González-Ladrón-de-Guevara (10), through a detailed meta-analysis, defined crowdsourcing as “a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task.” In this work, we view crowdsourcing as an alternative to lab-based or structured online methods of data collection, where increasing the flexibility of participation enables the collection of data at much greater scale.

Adjerid & Kelley (1) discuss how the allure of “big data” makes crowdsourcing attractive in experimental sciences, where logistical constraints often limit sample sizes. Where large scale datasets were once out of reach to all but the most highly resourced labs, they point out that a new generation of tools paired with the ubiquity of internet connections has brought this approach to data collection within the realm of possibility of more modestly financed projects. They argue that there are many benefits to collecting data at scale for multiple sub-fields of psychology, but go on to raise concerns around the ethics, quality, and appropriateness of these approaches to data collection.

Nonetheless, crowdsourced data has been shown to have benefits that extend beyond those induced simply by the size of the dataset. Casler, et al. (4) demonstrated that crowdsourcing a dataset resulted in greater socio-economic and ethnic diversity in the participant pool, whilst yielding almost indistinguishable test results from participants recruited in person or on social media. This suggests that crowdsourcing may be of particular interest in fields where cultural bias introduced through the homogeneity of participants is impactful. In timbre research, for example, it cannot be safely assumed that any perceptual phenomena observed are entirely devoid of cultural influence.

The similarity in group responses in Casler, et al.’s (4) study is encouraging in terms of the quality of data collected online. Nonetheless, the relince of psychoacoustic research on consistent audio reproduction sets the bar somewhat higher. To address this concern, Zacharakis, et al. (37) compared responses from a timbre dissimilarity study conducted in a laboratory, to those collected online. Timbre spaces computed from the two datasets showed very high configurational similarity, suggesting that both groups responded to similar acoustic cues despite variability in listening apparatus.

Crowdsourcing has previously been successfully applied to sourcing acoustic-semantic vocabularies in the context of music production. Specifically, Cartwright & Pardo (3) collected a set of descriptors of 40-band equalizer (EQ) curves from within an interactive EQ interface. Subsequently, Stables, et al. (32) captured descriptors from within the music production workflow by providing fully functional compressor, reverb, and EQ audio plugins.

2.3 Gamification

The term gamification broadly refers to the practice of imbuing a platform, application, or interface with motivational design features borrowed from games (20). It is of particular interest in the design of crowdsourced experiments, as these rely on a large body of participants being sufficiently motivated to provide responses. To address the ambiguity in terminology and praxis in crowdsourcing and gamification, Morschheuser, et al. (20) proposed a simple system of four archetypes for classifying gamified crowdsourcing systems, within which the present work would be described as a crowd-creating task.

Morschheuser, et al. state that a central challenge in designing a gamified crowdsourcing system is providing incentives that promote “the formation of positive motivations towards crowdsourcing work” whilst also fitting the type of activity. They further acknowledge that intrinsic motivation, such as that induced by tasks that “allow a participant to be creative and experience autonomy”, can be dominated by the extrinsic motivation of social or financial reward, or perceived motivational affordances (e.g. credits or in-game rewards) of the task. For these reasons, we were careful in designing the present study to avoid rewards that might overshadow the central activity, or incentivise users to act quickly and carelessly. We took particular inspiration from the approach of Stables, et al. (32), who motivated users to interact with their audio plugins by providing access to the responses of other users. Further details of the choices we made are given in section 3.3.

With the growing number of online platforms for gamified crowdsourced data collection, the popularity of this approach is rapidly growing despite limited data as to its effectiveness (17). One study even found that gamification did not improve participant attrition rates for “effortful” and “unengaging” tasks (18). Further, Deterding, et al. (5) point out challenging,
unanswered ethical questions, including whether the "playful veneer" of games and gamified systems leads users to share more data than they otherwise would, and to what extent users' effort might be considered an unfair use of their labour. Despite these concerns, however, Keusch & Zhang (17) did observe a clear effect throughout the literature on psychological outcomes such as fun, interest, and satisfaction.

3 METHOD

3.1 Participants
Due to the casual, gamified nature of the experiment, participants were not subject to any selection or screening process, nor was any demographic information collected. Each participant was, however, assigned a unique anonymous user ID on loading the page which was stored with each response they submitted. On their first visit, users were asked if they would allow a cookie to be stored on their computer to allow this user ID to persist between sessions. If the cookie was declined, a new ID would be generated at each visit. At the time of writing, a total of 95 unique user IDs have been recorded, and a total of 579 responses submitted.

3.2 Web interface
The prompted synthesis interface was hosted online. On arrival at the web page, visitors were given a brief explanation of the interface in informal language. Those who wished to learn more about the motivations behind the page were able to click a link to read some more detailed information. After reading the introductory directions, participants were presented with a rectangular control space containing a black square. Clicking and dragging the black square around the space produced a sound from a synthesizer built on the WebAudio API. As the square moved around the space, the sound of the synthesizer changed in accordance with a control mapping. Above the control space was a short textual prompt, instructing participants to create a sound matching a given adjective. The created sound was stored in a database alongside a time series of points representing the path taken through the control space.

3.3 Gamification & rewards
As discussed in section 2.3, we opted for a “light touch” in terms of gamification and rewards. In particular, to create a more relaxed, game-like experience, we opted for a more colourful user interface with informal language, including the use of emoji, in the explanatory text. To maximise the intrinsic motivation of the task itself, we used language that emphasised its creative and exploratory nature, and designed the control space mapping to be sufficiently nonlinear as to prevent users from easily employing prior knowledge of sound synthesis. Finally, to provide a reward for creating sounds, two further features were unlocked once five trials had been completed: (i) explore the responses of other users via heatmap visualisations, (ii) fit a kNN classifier in-browser to predict the prompts associated with different sounds.

3.4 Word stimuli
The descriptive adjectives used in the textual prompts were taken from the ModWiggler corpus collected in a prior prompted synthesis study (15). As the sounds visitors created lacked a clear amplitude envelope or sub audio rate modulation and were thus effectively time invariant, we removed two words from the original list of 27 that clearly implied a particular temporal evolution: percussive and plucky. For a detailed explanation of how these were sourced, see (15).

3.5 Synthesiser
To encourage participants to explore the synthesiser control space by listening to the sounds rather than relying on prior expectations, we aimed to create a synthesiser and accompanying parameter mapping that fulfilled the following criteria:

1. sounds do not obviously resemble familiar instruments
2. control dimensions do not obviously map to familiar (e.g. FM or subtractive) synthesiser parameters
3. control dimensions do not exhibit an obvious linear mapping to any specific attribute of the sound

The resulting synthesiser design consisted of three components connected in series: an interpolated wavetable space, a ring modulator, and a bandpass biquad filter. Details of these components and the parameters they exposed are given below, and an overall schematic is given in Fig. 1.

To allow the synthesiser to produce a variety of harmonic distributions without implying a familiar synthesis method, a wavetable oscillator was chosen as the initial sound source, with wavetables sampled using trilinear interpolation from a 2-dimensional grid. This allows for a direct mapping from the 2D control space to the wavetable.

The website is available at https://timbre.fun/
This space of wavetables was sampled from a multilayer perceptron with sinusoidal activations (29) which was trained to reproduce single cycles of audio from the NSynth dataset (8). The network’s inputs consisted of the x- and y-coordinates in the 2D wavetable space, and a time index within the wavetable, represented in two dimensions as a point on a unit circle to avoid discontinuities at the start and end of the wavetable. The x- and y-coordinate inputs for each training sample were derived as a learned matrix projection from a one-hot representation of sample identity, thus effectively allowing an unsupervised spatial organisation of wavetables to be learned. After training, the wavetable grid was sampled at a spatial resolution of 64 x 64 and a temporal resolution of 384.

The wavetable space allowed for the production of purely harmonic spectra, but was not capable of producing inharmonic signals. To address this, a ring modulator was added to the signal chain due to its ability to produce a finite number of inharmonic partials and its low dimensional parameter space. In a ring modulated signal, each frequency component \( \omega_i \) in the carrier signal is replaced by two frequency components \( \omega_i \pm \omega_m \) where \( \omega_m \) is the modulator frequency. Ring modulation can therefore be used to create a variety of effects from subtle beating through to complete destruction of the harmonic structure of the signal. The modulation frequency and gain parameters were mapped to the 2D control space.

Prior research into the semantic correlates of timbral attributes (15; 38) suggests that a relationship exists between certain verbal descriptors and acoustic features describing the shape or distribution of energy in the frequency spectrum. A notable example is the relationship between descriptions of brightness and the spectral centroid (23), an audio feature describing the weighted mean of frequencies present in a signal. With the wavetable space and ring modulator alone, however, participants would have been unable to shape the broader distribution of spectral energy beyond the pre-existing variation in harmonic amplitudes between the various wavetables. For this reason, we introduced a biquadratic bandpass filter with a fixed resonance of \( \sqrt{2} \) to the end of the signal chain. Participants were thus able to more closely control the spectral distribution of their created sound. The centre frequency parameter was mapped to the 2D control space.

### 3.5.1 Control mapping

Previous prompted synthesis studies have allowed participants to interact both through full sets of synthesiser controls (15) and low dimensional mappings (34), in which synthesis parameters correspond directly to the x- or y-axis. These approaches risk encouraging participants, especially those with existing synthesis experience, to rely on their intuition about the workings of the synthesiser in formulating their responses. To ameliorate this effect, we opted to obscure the relationship between the dimensions of the control space and the parameters of the underlying synthesiser (with the exception of the x- and y-coordinates in the wavetable space) by using a nonlinear mapping of the form:

\[
p_i = (0.5\cos(ax+by+cxy)+0.5)^2
\]

where the parameters \( a, b, \) and \( c \) were unique to each synthesiser parameter. These parameters were manually tuned to ensure a subjectively wide variety of sounds across the space, with no clear linear relationship between either axis and any synthesiser parameter. The values used in the experiment are listed in the source code repository\(^2\) alongside the final ranges of the resulting parameters.

### 3.6 Experiment design

Prior timbral studies have typically followed a design that facilitates statistical hypothesis testing, or the application of other relevant data analysis techniques. In particular, stimuli are usually applied equivalently to all participants, or at least to

\(^2\)https://github.com/ben-hayes/crossmodal-synthesis
those within a group. Whilst this is valuable from the perspective of drawing robust conclusions from the data, it constrains the scope of data that can be collected due to the redundancy that is necessitated by these statistical techniques. The utility of this data in applications that aim to use data-driven methods to build on research in timbre perception, such as synthesis control, is therefore limited.

As the purpose of timbre.fun is not to directly test hypotheses, but rather to pilot a method for crowdsourcing timbre data at scale with a view to supporting downstream applications, we opted to relax these usual constraints in favour of collecting a greater amount of more diverse data. Specifically, participants were able to complete as few or as many trials as they wished and prompts were simply sampled from a uniform distribution, with no guarantee that each participant responded to any specific subset of prompts.

4 RESULTS & DISCUSSION

Before analysis, the dataset was filtered to reduce the influence of low effort and accidental responses by removing any sound for which three or fewer JavaScript mouse move events were recorded. This resulted in a dataset containing 468 responses from 93 participants. Due to the inherently unmatched and unbalanced nature of the collected data, we forgo conventional statistical analysis in lieu of an exploratory analysis.

4.1 Control space density estimates

As an initial data exploration step, we visualised 2D kernel density estimates – with a standard bivariate normal – of participants’ responses in the 2D control space, conditioned on the semantic prompt. These are shown in Fig. 2

For the majority of prompts, the resulting densities appear relatively diffuse and lack distinct modes, suggesting that participants found suitable responses across large regions of the control space. Certain prompts, however, do appear to have resulted in clearer agreement between participants: clear, dull, noisy, and sharp, for example, appear to have received a large proportion of their responses in small regions of the control space.

High level similarities, congruent with the groupings of descriptors observed in previous timbre semantic work (15; 38; 6), do also seem to emerge. For example, dark, deep, and dull appear to have most of their responses concentrated around a similar region, while thin and thick primarily occupy opposing regions of the space. The nonlinear mapping of synthesis parameters, however, means that any regions exhibiting similarity other than very distinct modes should be interpreted with caution, as neighbouring areas of the control space could result in dramatically different sounds.

4.2 Effect of prompt on synthesis parameters

Fig. 3 illustrates the distribution of responses for two of the five synthesiser parameters – filter center frequency and modulator gain – conditional on the prompt given. In both cases the x-axis is sorted according to the median value per prompt, and the colour scale is normalised to the full available range of the parameter. Plots for the remaining prompts
Very pronounced differences between prompts are visible for the Filter Centre and Modulation Gain parameters. In the case of Filter Centre, we see prompts including sharp, bright, thin, and harsh resulting in consistently high parameter values, and deep, smooth, dull, and mellow resulting in lower ones. For the Modulation Gain, we see hard, noisy, harsh, and big resulting in higher values, while clear, sweet, clean, and dull lead to lower values. These coarse groupings are, again, consistent with the factor structure of our prior prompted synthesis study, in which sharp, bright, and harsh all exhibited their highest loadings on the same factor, and clean, clear, and sweet all exhibited their highest on another factor. The grouping Effects on the remaining parameters (see online supplement for plots) are markedly weaker, with wider distributions and less difference between median parameter values. Whilst the prompt effects on wavetable x- and y-coordinates appear primarily to have resulted in diffuse distributions, some prompts do appear to have a resulted in a concentration of responses around a particular value, suggesting that the wavetables at this point in the space may agree particularly well with the descriptor. These prompts include sharp, gritty, and noisy.

### 4.3 Acoustic features analysis

Whilst synthesiser parameters offer a comprehensive description of how to reproduce a given sound from the dataset, they provide only indirect insight into a sound’s acoustic characteristics, owing to their perceptual nonlinearity and properties emerging from their interaction. We therefore extracted a set of acoustic features from all sounds created in the experiment to more completely describe them.

A four second long WAV file was rendered for each sound using the same WebAudio synthesiser as found on the timbre.fun website. Features were extracted from each audio file using the Timbre Toolbox (22). As the sounds lacked any temporal evolution, features computed on the Temporal Energy Envelope representation were excluded. All other representations and features included in the Toolbox’s default configuration were used. By default, the Timbre Toolbox reports two summary statistics aggregated over time: the median and interquartile range. Due, again, to the lack of temporal variation in the sounds, we discarded the interquartile range and retained only the median.

As many of the computed features measure very similar characteristics, we removed columns with high pairwise correlations in order to avoid the analysis being biased by over-representation of certain properties. In particular, for every pair of features with a Spearman correlation coefficient of greater than 0.9, we removed the feature with the higher mean absolute correlation across all remaining features. The list of remaining features is given in Table 1.

Principal components analysis was then applied to the remaining features. Parallel analysis (16) was used to select an appropriate number of components. This procedure involves repeating PCA on a number of randomly sampled, uncorrelated datasets of equivalent size, and retaining n components, where the n-th component in the real dataset is the component with the lowest eigenvalue that still exceeds a given percentile of the eigenvalues of the n-th components of the random
datasets. We ran the procedure for 480 iterations and found that three components exceeded the 99th percentile. Loadings of the first three components are reported in Table 1.

The resulting principal components were then subject to k-means clustering. A two-cluster solution was selected as optimal by computing the connectivity, Dunn index, and silhouette coefficient. For each of the 25 prompt words, we computed the proportion of sounds created in response to that word in each acoustic cluster. These are shown in Fig. 4. Note that the proportion was used in lieu of the raw count owing to the unbalanced design of the experiment. The majority of prompts appear to strongly associated with a specific cluster, with 19 prompts having at least twice as many responses in one cluster than the other.

We note that descriptors that have previously been associated with distinct semantic factors are grouped together in these clusters. For example, cluster 2 contains the majority of responses for sharp, thin, and rough, whilst these terms load strongly onto two distinct factors in our prior study (15) and onto three distinct factors in Zacharakis, et al.’s prior work (38). Instead of representing distinct timbral dimensions, the acoustic clusters instead appear to correspond to the opposite ends of an aggregation of timbral scales. This emergence of two timbral “poles” in this manner is somewhat consistent with the behaviour observed in our previous study (15) where, despite post-hoc semantic ratings supporting a five factor solution, correlations between semantic factors and synthesiser parameters appeared to form just two groups.

### 4.4 Word affect

Using the word affect norms collected by Warriner, et al. (35), valence, arousal, and dominance scores were obtained for all 25 prompts. In all cases except one the exact word form was found in the dataset. In the one case where this wasn’t possible (woody), the closest adjectival form that shared a lemma (wooden) was used in its place. Each of the three dimensions was then discretised into two classes: high for those above the mean, and low for those below. Two classes were used instead of the three used by Wallmark, et al. (34) to enable both direct comparison with our acoustic clusters and the formulation of binary classification problems (see Section 4.5).

To examine the relationship between these affect classes and the clusters obtained from acoustic features, prompts were
assigned to the acoustic cluster which contained a greater proportion of their responses, and three external validation metrics – the Rand, Jaccard, and Fowlkes-Mallows indices – were computed between these assignments and each of the three affect dimensions in turn. The results of these metrics are presented in Table 2. Whilst low to moderate values were observed for the valence and dominance classes, the arousal classes showed very high similarity to the separation of prompts between acoustic clusters. Whilst neither the data nor analysis are sufficient to establish a causal relationship, these results are strongly suggestive of an interaction between prompt arousal and the acoustic characteristics of the resulting sound.

### 4.5 Affect classification with support vector machines

As discussed previously, the structure of this dataset precludes conventional statistical hypothesis testing. Therefore, to further examine the relationship between the created sounds and the affective connotations of the prompt words we instead opted to fit a predictive machine learning model. In particular, we used a support-vector machine (SVM) with radial basis functions. We trained separate models on both acoustic principal components and the raw synthesiser parameters.

Data was partitioned into 80% training and 20% test subsets. Using this split, six models were fit — one for each combination of affect dimension and feature set (where synth parameters and acoustic principal components were the two feature sets). SVM hyperparameters were tuned using the adaptive random search implemented in the caret package for R, following a 10-fold cross validation with 5 repeats of each fold. To ensure an even class balance, training folds were upsampled by randomly repeating samples from the less represented class.

Table 2 lists the test-set accuracy of each trained SVM. Statistical significance of the accuracy was determined using a binomial test, where the no-information rate of the test dataset (i.e. the raw class proportions) was used as the null hypothesis. Consistently with the external validation metrics, the test accuracy observed for arousal classes was highly statistically significant, further supporting an affect-timbre interaction.

These results are distinct from those of Wallmark, et al. (34), who did not observe a significant effect of arousal on acoustic principal components. Deeper comparison of these results is challenging due to methodological differences, but such a clear deviation from the particular timbre-affect relation observed in their work is nonetheless noteworthy. In particular, it both lends further weight to the role of affect in prompted timbre production, whilst also suggesting that the mechanism of this interaction may to some extent be influenced by the method of synthesis available.

### 5 CONCLUSION

This paper presented a method for sourcing a timbre semantic vocabulary of prompt-sound pairs at scale using a gamified crowdsourced approach. A pilot run of the method, debuted as part of a digital exhibition on sensory variation and interaction at the 2021 Edinburgh Science Festival, resulted in a total of 579 sounds being created by 95 users. Exploratory analysis of the collected data suggests that responses showed logical consistency with prior timbre research, and exhibit sufficient structure to allow certain affective connotations of prompt words to be predicted from acoustic characteristics of the created sounds. These results support further work on gamified crowdsourced collection of timbral datasets, with a view to producing large datasets for use in downstream tasks, including descriptive synthesiser control.

In future work, we will augment the data collection pipeline by introducing new sound synthesis methods, new types of interaction (including tagging, dissimilarity rating, and more), and further gamification elements (such as a reward/points system, and the ability to unlock new features). We will also conduct a run of the study at a larger scale, and provide a full open-source release of the platform so that it can be adapted for other data collection tasks.

### ACKNOWLEDGEMENTS

This work was supported by UK Research and Innovation [grant number EPS022694/1].
REFERENCES


Aeroacoustic analysis of oboe reeds with compressible direct numerical simulation

Yuki NAKAHARA(1), Rei SUMITA(1), Ryoya TABATA(1), Sho IWAGAMI(1), Takeshi NANRI(2), Taizo KOBAYASHI(2), Yuji HATTORI(3) and Kin’ya TAKAHASHI(1)

(1)Department of Physics and Information Technology, Kyushu Institute of Technology, Japan, takahasi@phys.kyutech.ac.jp
(2)Research Institute for Information Technology, Kyushu University, Japan, tkoba@cc.kyushu-u.ac.jp
(3)Institute of Fluid Science, Tohoku University, Japan, yuji.hattori@tohoku.ac.jp

ABSTRACT
A two-dimensional model of an oboe reed is studied numerically with a direct numerical simulation (DNS) of the compressible Navier-Stokes equations to investigate the sound generation mechanism from the viewpoint of aeroacoustics. The numerical tool is extremely accurate due to the smallest mesh size on the order of micrometers and successfully reproduces the details of fluid motion and acoustic vibrations inside and outside the reed. Particular attention is paid to the effect of reed vibration on the sound generation mechanism. When the reeds are fixed and a periodically varying flow is injected through the fixed reed slit, an aerodynamics sound created inside the reeds is an almost monotone including a few overtones. On the other hand, when a flow is injected through periodically vibrating reeds from an oral cavity, more overtone components are observed and the pressure waveforms are similar to those observed in the experiment. This indicates that the richness of the overtones of the double-reed instrument is mainly attributed to the aerodynamic sound created by the flow injected through vibrating reeds and the bore, a linear resonator, just enhances characteristics of the instrument, e.g., formant.

Keywords: DNS, Double reed, oboe

1 Introduction
The study of sound generation mechanism of double-reed instruments such as the oboe and bassoon, is an important problem in the field of musical acoustics[1, 2, 3, 4]. However, the number of publications on double-reed instruments is much less than those on single-reed instruments and lip-reed instruments, and no efficient model used widely has been proposed such as the Schumacher model for single-reed instruments and the Adachi-Sato model for lip-reed instruments [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. This is mainly due to the complexity of sound generation process evolving nonlinear motions of compressible fluid. The double reed has a narrow channel in which an airflow must obey the governing equations of compressible fluid and the effect of fluid motions is not negligible, for example, Bernoulli’s force acting on the inner surfaces of the reeds. An inlet flow injected from an oral cavity passes through the slit of vibrating reeds and is changed into acoustic pressure due to the interaction with the vibrating reeds. Thus, acoustic waves generated in double-reeds are regarded as aerodynamic sound, and one needs to address the problem of where and how acoustic vibrations arise from a fluid flow injected through vibrating reeds.

One of the important tools for attacking this problem is a numerical analysis with a compressible fluid scheme. Several numerical schemes for compressible fluid, e.g., Lattice Boltzmann Method (LBM), Large Eddy Simulation (LES) and Direct Numerical Simulation (DNS), have been used for the study of wind instruments[12, 13, 16, 17, 18]. Compressible DNS is the most reliable tool, but requires a huge numerical cost[19]. In this paper, we numerically studied a two-dimensional model of an oboe reed with a compressible DNS solver to investigate the sound generation mechanism, when blowing a real oboe reed alone. To achieve this work, we constructed a 2D model of an oboe reed attached to an oral cavity. First, we checked the reproducibility of our
numerical method. Then, we focused on the effect of reed vibration on the sound generation process. Thus, to clarify the effect of vibrating reeds, we investigated the fluid motion and generated sound in detail comparing with the results when the reed was driven with a periodic flow injected through the fixed reed slit.

2 Model and Numerical Method

Figure 1 shows the dimensions of the two-dimensional model of the oboe reed accompanied by an oral cavity. We considered the following three cases by using this two-dimensional model. In Case 1, we determined the resonance frequency of the oboe reed in the following way. That is, a Poiseuille flow with a maximum velocity of $V = 10 \text{ m/s}$ was injected from the oral cavity, and the resonant frequency of the oboe reed was measured by observing the frequency of pressure-decay oscillations in an exterior domain. In Case 2, the model was driven by an alternating current-like periodic flow to mimic the flow injected through an oscillating reed slit. Then, the velocity of the flow at the left end of the oral cavity was given by

$$V = V_0 (\sin(2\pi f_H t) + 1),$$

where $V_0 = 10 \text{ m/s}$ and $f_H$ is the resonance angular frequency obtained in Case 1. In addition, the oboe reed was also driven at A5 (880 Hz), which was observed when blowing a real oboe reed alone. In Case 3, the effect of reed vibration was considered. As shown in Figure 2, to mimic the oscillations of the reed plates, portions of the plates of 1mm length from the reed tip were vibrated against each other with a period $T_H = 1/f_H$ with opening and closing the slit between them, and a constant flow of $V = 10 \text{ m/s}$ was injected from the left end of the oral cavity.

The compressible direct numerical simulation (DNS) developed by Komatsu et al. was employed for the numerical calculations[19]. The two-dimensional compressible Navier-Stokes equations are solved by using a higher-order difference scheme with the volume penalization (VP) method. The entire computational domain is $2000 \times 2000 \text{ mm}^2$ in all cases. Calculations were performed up to 0.01 second in each case. Table 1 shows the mesh parameters and time steps. Figure 3 shows a magnified view of the mesh near the oboe reed with observation point P. For convenience of constructing mesh, the oboe reed and oral cavity were embedded in a solid body. Table 2 shows the physical parameters of the compressible fluid. The simulations were performed on one node with 36 cores (36 parallel threads) of the high-performance computer cluster ITO (subsystem A) at Kyushu University, and each calculation took approximately 40 days.

![Figure 1. Two-dimensional model: $O_h = 10 \text{ mm}, S_h = 1 \text{ mm}, L = 71 \text{ mm}, D = 5 \text{ mm}$](image)

Table 1. Mesh parameters and time step

<table>
<thead>
<tr>
<th>Number of grids</th>
<th>Minimum grid size</th>
<th>Time step $\Delta t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2806 \times 2480$</td>
<td>$0.737 \times 10^{-2} \text{ mm}$</td>
<td>$1.8 \times 10^{-9} \text{ s}$</td>
</tr>
</tbody>
</table>


Table 2. Physical parameters of compressible fluid

<table>
<thead>
<tr>
<th>parameters</th>
<th>Density at rest $\rho_0$</th>
<th>Viscosity $\mu$</th>
<th>Sound speed $c_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>values</td>
<td>1.162 kg/m$^3$</td>
<td>$1.846 \times 10^{-5}$ Pa·s</td>
<td>347.127 m/s</td>
</tr>
</tbody>
</table>

3 Numerical results

3.1 Case 1

Figure 4 shows (a) the velocity distribution and (b) the vorticity distribution near the reed at $t = 0.01$s. Figure 5 shows (a) the pressure distribution in the whole computational domain at $t = 0.01$s, (b) the change in pressure at observation point P, and (c) a frequency-resolved graph of the pressure at observation point P. As shown in Figures 4 (a) and (b), the incoming airflow formed a wavy jet inside the reed and vorticity took large values near the tip of the reed. The wave jet should be transformed into a sound wave. However, as shown in Figure 5 (a) and (b), the acoustic wave radiated from the open end of the reed was not a regular spherical wave and the acoustic oscillation observed at the point P was not stationary, but rather gradually decayed being accompanied by noise components. This is because the constant flow cannot maintain a steady oscillation with a constant pitch. Nevertheless, the resonance frequency could be calculated from the Fourier spectrum of the pressure fluctuation and was estimated as $f_H = 2514$ Hz, which is 200 Hz higher than the resonance frequency theoretically estimated for a conical pipe of the same length, 2314.18 Hz. Figure 6 shows the pressure distributions along the
center line of the reed at representative observation times. Large negative values of pressure were observed near the tip of the reed. These negative values were mostly caused by strong vortices near the tip. Since vortices permanently existed near the tip, the envelopes of these graphs did not significantly change depending on the observation times.

Figure 4. Velocity and vorticity distributions at $t = 0.001s$ for Case 1. The x-axis and y-axis are in mm. (a) Velocity distribution. The scale of the color bar is in m/s. (b) Vorticity distribution. The scale of the color bar is in 1/s.

Figure 5. Pressure oscillations and power spectra for Case 1. (a) Pressure distribution at $t = 0.001s$. The x-axis and y-axis are in mm. The scale of the color bar is in Pa. (b) Pressure oscillation at the observation points P. (c) Power spectra of the pressure oscillation observed in (b).
Figure 6. Pressure distributions along the center line of the reed for Case 1.

3.2 Case 2
Figure 4 shows (a) the velocity distribution and (b) vorticity distribution near the reed at \( t = 0.01 \) s, when the driving frequency is set at \( f_H = 2514 \) Hz. Figure 5 shows (a) the pressure distribution in the whole computational domain at \( t = 0.01 \) s and at \( f_H = 2514 \) and 880 Hz, (b) the changes in pressure at observation point P at \( f_H = 2514 \) and 880 Hz, and (c) power spectra of pressure oscillations observed in (b). Compared to Case 1, no significant differences were observed in the flow velocity and vorticity distributions, but the behavior of pressure oscillations was much different from that for Case 1. Indeed, a regular and stationary spherical wave is radiated from the open end of the reed. The pressure oscillations observed at the point P were regular and periodic with large amplitude for both driving frequencies and their wave forms were similar to a sinusoidal wave. The oscillation at \( f_H = 2514 \) Hz was larger in amplitude than that at \( f_H = 880 \) Hz, because the former was the resonance frequency of the reed. From the Fourier spectra, for the case of \( f_H = 2514 \) Hz, the peak of the fundamental mode was dominant and the peak of the second harmonic was one digit less than the fundamental peak. For the case of \( f_H = 880 \) Hz, the peak of the fundamental mode was dominant and the second harmonic was not observed clearly. Figure 9 shows the pressure distributions along the center line of the reed at representative observation times for the case of \( f_H = 2514 \) Hz. Except in a neighborhood of the tip, where strong vortices existed, wave forms observed are regarded as snapshots of the standing wave of the fundamental mode.

Figure 7. Velocity and vorticity distributions at \( t = 0.001 s \) for Case 2, when the driving frequency is \( f_H = 2514 \) Hz. The x-axis and y-axis are in mm. (a) Velocity distribution. The scale of the color bar is in m/s. (b) Vorticity distribution. The scale of the color bar is in 1/s.
Figure 8. Pressure oscillations and power spectra for Case 2. (a) Pressure distributions at $t = 0.001\text{s}$ and at $f_H = 2514\text{Hz}$. The x-axis and y-axis are in mm. The scale of the color bar is in Pa. (b) Pressure oscillations at the observation points P at $f_H = 2514$ and 880Hz. (c) Power spectra of the pressure oscillations in (b).

![Figure 8](image)

Figure 9. Pressure distributions along the center line of the reed for Case 2.

3.3 Case 3

Figure 10 shows (a) the velocity distribution and (b) the vorticity distribution near the reed at $t=0.01\text{s}$ and at $f_H = 2514\text{Hz}$ for Case 3. Figure 11 shows (a) the pressure distribution in the whole computational domain at $t=0.01\text{s}$ and at $f_H = 2514\text{Hz}$ (b) the changes in pressure at observation point P at $f_H = 2514$ and 880Hz, and (c) Fourier spectra of the pressure oscillations in (b). As shown in Figure 10 (a) and (b), the flow injected from the oral cavity formed a wavy jet and vortices were observed in a neighborhood of the tip of the reed, but the wave jet gradually decayed and seemed to change into an acoustic oscillation in the middle of the reed. Actually, vorticity decayed quickly. As shown in Figure 11 (a) and (b), an acoustic wave radiated from the open end of the reed was stronger than that for Case 2 at $f_H = 2514\text{Hz}$, and pressure oscillations at the point P were also larger in amplitude than those for Case 2 at $f_H = 2514$ and 880Hz. Note that the oscillation at $f_H = 2514\text{Hz}$ was larger in amplitude than that at $f_H = 880\text{Hz}$ due to the same reason for Case 2. Furthermore, the wave form at $f_H = 2514\text{Hz}$ was apparently deformed from a sinusoidal wave due to the effect of the vibrating reed plates, and the wave at $f_H = 880\text{Hz}$ was much different from a sinusoidal wave and seemed to include the 2nd and 3rd harmonic components. As shown in Figure 11 (c), such wave forms included rich harmonic components and are often observed for conical instruments. Thus, the vibration of reeds together with the oral cavity plays an important role in the generation of harmonic components. Figure 12 (a) shows the pressure distributions along
the center line of the reed at representative observation times for the case of \( f_H = 2514 \text{Hz} \). Restively sharp kinks of the wave were observed and seemed to travel along a semi-oval shape in time evolution. These wave forms are similar to the right half of the Helmholtz wave shown in Fig.12(b), which is observed in the vibration of a string of string instruments. Such wave forms similar to the Helmholtz wave were also numerically obtained with the delay differential model of instruments with a conical bore [11].

Figure 10. Velocity and vorticity distributions at \( t = 0.001s \) and at \( f_H = 2514 \text{Hz} \) for Case 3. The x-axis and y-axis are in mm. (a) Velocity distribution. The scale of the color bar is in m/s. (b) Vorticity distribution. The scale of the color bar is in 1/s.

Figure 11. Pressure oscillations and power spectra for Case 3. (a) Pressure distributions at \( t = 0.001s \) and at \( f_H = 2514 \text{Hz} \). The x-axis and y-axis are in mm. The scale of the color bar is in Pa. (b) Pressure oscillations at the observation points P at \( f_H = 2514 \) and 880Hz. (c) Power spectra of the pressure oscillations in (b).
Figure 12. Waves in the reed. (a) Pressure distributions along the center line of the reed for Case 3. (b) Schematic view of Helmholtz wave.

4 CONCLUSIONS
In this paper, a two-dimensional model of an oboe reed was numerically studied using a two-dimensional Direct Numerical Simulation (DNS) of the compressible Navier-Stokes equations to investigate the sound generation mechanism in terms of aeroacoustics. Particular attention was paid to the effect of reed vibration on the sound generation process, and the following facts were found concerning the composition of harmonic components in sound waves. When the reed was driven with a periodic flow injected through the fixed reed slit with its resonance frequency or a frequency less than it, the sound wave emitted from it contained negligibly small harmonic components except the fundamental. On the other hand, when the inlet flow was injected from the oral cavity through the vibrating reed slit, the sound wave radiated from the open end contained rich harmonic components, and similar waves are often observed for real conical instruments with a double reed. This fact indicates that the richness of harmonic components in sound waves generated by double reed instruments such as the oboe is mainly attributed to the nonlinear interaction between the injected flow and vibrating reed plates, which causes an aerodynamic sound. The bore of the instrument, a linear resonator, only enhances the characteristics of the an individual instrument such as the formant for generated tones.

Unfortunately, the problem of where and how acoustic vibrations arise from a fluid flow injected through vibrating reeds has not been addressed. To solve this problem, an approach based on aero-acoustic theory, such as Lighthill’s acoustic analogy, is needed. Furthermore, the vibrating parts of the reed model take unrealistic shapes for convenience of numerical setup. Therefore, the model should be improved to reproduce more realistic reed vibrations and to investigate sound generation processes in detail. Further development in these directions will be the subject of future work.

ACKNOWLEDGEMENTS
The present work was supported by Grant-in-Aid for Scientific Research (C) Nos. 19K03655 from the Japan Society for the Promotion of Science (JSPS) and “Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures” and “High Performance Computing Infrastructure” in Japan (Project ID: jh220001). Part of the work was carried out under the Collaborative Research Project of the Institute of Fluid Science, Tohoku University.

REFERENCES


Numerical approach for aerodynamics around two tone holes of woodwind instruments

Shota TAKANAMI(1), Ryoya TABATA(1), Sho IWAGAMI(1), Taijiro OHNO(1), Takeshi NANRI(2), Taizo KOBAYASHI(2), Kin’ya TAKAHASHI(1)

(1) Kyushu Institute of Technology, Japan, takahasi@phys.kyutech.ac.jp
(2) Kyushu University, Japan, tkoba@cc.kyushu-u.ac.jp

ABSTRACT

In this paper, we discuss the numerical reproducibility of the compressible fluid behavior around two tone holes of woodwind instruments by using compressible Large Eddy Simulation (LES). In particular, we focus on the situation that the tone holes are opened and closed with moving pads above the tone holes, which is regarded as a moving boundary problem with topology change, and reproduce the change of the pitch when opening and closing the tone holes. Our two-dimensional model of a "recorder" has two tone holes. To reproduce the opening and closing the tone holes, the pads are moved continuously. That is, the position of the pads were continuously changed in the order of "open - close". Our numerical results are consistent with the Keef’s experimental results. We solved the moving boundary problem with topology change under the situation of acoustics of fluid-structure interaction, and reproduced the pitch change in the opening and closing the tone holes of the recorder like woodwind instrument model.

Keywords: Flue instruments, Tone hole, Moving boundary conditions, Numerical simulation

1 Introduction

The transient phenomena around tone holes of woodwind musical instrument have been studied as a long standing problems in the field of musical acoustics[1, 2, 3, 4, 5, 6, 7, 8]. The main difficulty in studying the function of tone holes comes from the highly complicated interaction around compressible fluid dynamics (non-linear dynamics), aerodynamic sound as a linear wave motion and a moving boundary problem with topology change.

In order to achieve a rigorous study on it, we had developed an compressible LES solver resolving the moving boundary problem[9]. Then we had studied the function of a tone hole by using a 2D model of a small recorder-like-instrument with a tone hole[10]. In that study, the pad above a tone hole is set to a fixed position, that is, the distance between the pad and the tone hole is taken at fixed values, 0, 0.5, 1, 2, 3, 5, 10, 20mm, and the change of pitch depending on the position of the pad is observed. The results is consistent with Keefe’s experimental results. In the next study, we had developed numerical models in order to manipulate moving mesh without topology change[11].

In this paper, we give a short review of recent results about this series of study. We reproduce numerically the compressible fluid behavior around two tone holes of woodwind instruments by using compressible LES. We focus on the situation that the tone holes are opened and closed with moving pads above the tone holes, which is regarded as a moving boundary problem with topology change. In order to achieve seamless numerical calculation of moving boundary problem with topology change, we develop compressible LES solver and a method to move and remap the mesh structure. The numerical results of our models are consistent with Keefe’s experimental results, and error rates of pressure between mesh remapping with/without topology change are in the same order \( \sim 10^{-9} \).
2 Model and Numerical Method

We have developed a two-dimensional “recorder” like model based on a model which is a 1/4 scaled 2D analog of Keefe’s experimental model (see Figure 1 upper), which has two tone holes placed at the center of the resonance tube and the point of 1/4 length from the end of the resonance tube. The pipe length \( L \), pipe diameter \( H \), height of tone holes \( d \), diameter of tone holes \( b \), embouchure distance \( D \), embouchure inlet diameter \( h \), edge angle \( \theta \), are set to \( L = 138.5 \text{mm} \), \( H = 10 \text{mm} \), \( d = 1.5 \text{mm} \), \( b = 6.6 \text{mm} \), \( D = 5 \text{mm} \), \( h = 1 \text{mm} \), \( \theta = \pi/7 \text{rad} \), respectively. To reproduce the opening and closing the tone holes, the pads are moved continuously.

Figure 1. (upper) Dimensions of the model. (lower) typical mesh structure. “dgap” is the distance between the pad and the top of the tone hole.

We employ a compressible Large Eddy Simulation (LES) with the one-equation sub-grid-scale (SGS) model[12]. In practice, we have developed an extended compressible LES solver “rhoPimpleMyDyMFoam” which includes a function of “DynamicMesh” for the numerical calculation of moving mesh. This solver “rhoPimpleMyDyMFoam” is based on “rhoPimpleFoam” in the open source software, OpenFOAM ver.5. The pressure and temperature at the rest are taken as \( p_0 = 10^5 \text{Pa} \) and \( T_0 = 300 \text{K} \), respectively. The smallest mesh size around the tone holes is \( \Delta x = 0.17 \text{mm} \) and the time step of the numerical integration is \( \Delta t = 1 \times 10^{-7} \text{s} \). The number of mesh cells is 243,528. To excite an acoustic wave in the resonance tube at a given frequency, the jet velocity “\( U \)” at the embouchure is taken as \( |U| = 20 \text{m/s} \) in the steady state.

To numerically solve the moving boundary problem with topology change, we use a combination of “DynamicMesh” and “mapFields”, which controls the pads movement and the mesh manipulation. In practice, we configure moving velocity of mesh in the configuration file “pointMotionU” and “cellMotionU”. In order to relieve the distortion of mesh structure, we use “mapFields” every interval of 0.2mm for the pad movement. We configure this mesh structure remapping in the configuration file “mapFieldsDict”.

The actual pad movement is shown in Figure 3. The quantity “dgap” as a function of time must be continuously differentiable in order to avoid shock waves. And furthermore, at the point (dgap=0mm) where the topological change of the mesh-boundary arises, the velocity of the pad as a function of time must be also continuously differentiable. The maximum velocity of the pad is set to 0.1m/s.

To ensure that the number of meshes between the tone hole and the pad does not change significantly when the pad is moved, default mesh structures for every 0.2mm from 0mm to 1mm are constructed. If the dgap is
Figure 2. Mesh structures around the pad and tone hole. These are the default “dgap” mesh structures, respectively. The distance between the pad and the top of the tone hole is 1 to 0mm.

Figure 3. Time evolution of the pad movement. (a) the velocity, (b) “dgap”: the distance between the pad and the top of tone hole.
between 1mm and 0.8mm, a default “dgap” mesh of 1mm is used. When the dgap narrows to 0.8mm, switch to the 0.8mm default “dgap” mesh. From a dgap of 0.8mm to 0.6mm, use a 0.8mm mesh. The process is repeated in the same manner until it is completely closed at 0 mm.

Table 1. Mesh dimensions and corresponding time of them

<table>
<thead>
<tr>
<th>Region</th>
<th>Time[s]</th>
<th>Time of mapping mesh[s]</th>
<th>Default “dgap”</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>0–0.053</td>
<td>0.053</td>
<td>1mm</td>
</tr>
<tr>
<td>(2)</td>
<td>0.053–0.055</td>
<td>0.055</td>
<td>0.8mm</td>
</tr>
<tr>
<td>(3)</td>
<td>0.055–0.057</td>
<td>0.057</td>
<td>0.6mm</td>
</tr>
<tr>
<td>(4)</td>
<td>0.057–0.059</td>
<td>0.059</td>
<td>0.4mm</td>
</tr>
<tr>
<td>(5)</td>
<td>0.059–0.06198</td>
<td>0.06198</td>
<td>0.2mm</td>
</tr>
<tr>
<td>(6)</td>
<td>0.06198–0.07</td>
<td></td>
<td>0mm</td>
</tr>
</tbody>
</table>

Table 1 shows the mesh dimensions and the time of mesh structure remapping.

3 NUMERICAL RESULT

3.1 Resonance frequencies of typical four conditions
Our numerical model has two tone holes. Thus, the model has four typical conditions by the combination states of the tone holes, which are open and/or closed. That is, the combination states of tone holes “left-right” can be set to “open-open”(Fig. 4 (1)), “open-close”(Fig. 4 (2)), “close-open”(Fig. 4 (3)), “close-close”(Fig. 4 (4)). Since, the resonance frequency is depend on the effective pipe length, these four conditions make different resonance frequencies each other. The shorter the effective resonance pipe length, the higher the resonance frequency; thus the resonance frequency increases in the order “close-close”, “close-open”, “open-close”, “open-open”. Numerical results of the power spectra of the pressure oscillations are shown in Figure 5.

As shown in Figure 5, the resonance frequencies of “close-close”, “close-open”, “open-close”, “open-open” are 1097Hz, 1317Hz, 1427Hz, 1586Hz, respectively. Thus our numerical model of a recorder with two tone holes was successfully reproduces the resonance frequencies according to the state of the tone holes.

3.2 Transition of the resonance frequency
Two tone holes are simultaneously closed from 1mm to 0mm by applying a constant jet velocity “U” of 20m/s at the embouchure. The default “dgap” mesh is switched at the times shown in Table 1. Figure 6 (left) shows the time evolution of the pressure distribution. The coloring corresponds to the default “dgap” meshes. The amplitude suddenly increases from dgap=0.2mm in Region 6. This is due to the fact that the jet velocity “U” of 20 m/s is too large for a pipe length of 138.5 mm.

Figure 6 (right) shows the results for resonance frequency. As with the left graph, the regions are color-coded. It can be seen that the resonance frequency decreases from Region 6 to 1. Furthermore, the resonance frequency is the same from Regin 2 to 4, and then decreases abruptly at Region 5. This indicates that the resonance frequency does not decrease continuously in synchronization with the decrease in the distance between the tone hole and the pad, but rather decreases at a certain fixed distance. This phenomenon has been proven in the Keef’s experiments, and we thought that the same phenomenon might have occurred in the numerical analysis.

However, in the condition that this pads move in succession, the period of time in each Region is as short as 0.002 sec. Thus the frequency resolution of the Fourier transform is rough. Therefore, it is necessary to determine the resonance frequency at each region with high accuracy. The resonance frequencies at each “dgap” were determined by numerical calculations for 0.1 sec at appropriate values of “dgap” from 1 mm to 0 mm.
Figure 4. Mesh structures: (1) the both tone holes are open, (2) the right side tone hole is closed and the left side tone hole is open, (3) the left side tone hole is closed and the right side tone hole is open, (4) the both tone holes are closed. Effective pipe lengths are shortened in the order (1), (2), (3), and (4). The distance between the open pad and the top of the tone hole is set to 1mm.
Figure 5. Power spectra of the pressure oscillations.

“Open-Open”: 1586Hz  “Open-Close”: 1427Hz

“Close-Open”: 1317Hz  “Close-Close”: 1097Hz

Figure 6. The distance between the tone holes and the pads are varied continuously from 1 mm to 0 mm, (left) time evolution of pressure $p$ and (right) resonance frequency transition. In the range of “Region” from 2 to 4 mm, the resonance frequency is almost no change.
Figure 7. Power spectra of pressure $p$. In the range of “$dgap$” conditions from 0.8 mm to 0.3 mm, the resonance frequency is almost no change. These results are consistent with Keef’s experimental results.
In order to find out from which position the resonance frequency changed to increase, we prepared several conditions with the distance between the tone hole and the pad ranging from 0.8 mm to 0 mm, fixed the conditions at each position, performed the analysis, and observed the change in the resonance frequency. The series of resonance frequencies at each point is shown in Figures 7. In the range of “dgap” conditions from 0.8 mm to 0.3 mm, the resonance frequency is almost no change. However, it begins to decrease from 0.2 mm to 0.1 mm of the “dgap”. Although there is a difference in this frequency change, the resonance frequency value continues to decrease up to 0.05 mm. From this result, it is found that the resonance frequency decreased after 0.2 mm of the distance between the sound hole and the pad. These results are quantitatively consistent with Keef’s experimental results.

3.3 Error estimations of mesh manipulation

<table>
<thead>
<tr>
<th>Mesh manipulation</th>
<th>Default mesh dimension</th>
<th>Pressure [Pa]</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>decrease mesh (1mm → 0.8mm)</td>
<td>1mm (Before)</td>
<td>100089.0219</td>
<td>$1.0 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.8mm (After)</td>
<td>100089.0338</td>
<td></td>
</tr>
<tr>
<td>topology change (Open → Close)</td>
<td>0.2mm (Before)</td>
<td>99953.51932</td>
<td>$4.0 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>0mm (After)</td>
<td>99953.51892</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows errors of pressure at the “Center” sampling point. In the case of “increase mesh (0.8mm → 1mm)”, error is bigger than the case of “decrease mesh (1mm → 0.8mm)”. We assume that this increasing of error rate is caused by a higher order complement of physical quantities when they are mapped from on small mesh onto large one.

Other error rates are in the almost same order $\sim 10^{-8}$. Thus, the function of moving boundary calculation with topology change in our numerical model is ensured.

4 CONCLUSIONS

In this paper, we study the numerical reproducibility of the compressible fluid behavior around two tone holes of woodwind instruments by using compressible LES. In particular, we focus on the situation that the tone holes are opened and closed with moving the pads above the tone holes, which is regarded as a moving boundary problem with topology change.

Our numerical model is a two dimensional “recorder” like model based on a model which is a 2D analog of 1/4 scaled Keefe’s experimental model. For seamless numerical calculation of moving boundary problem with topology change, we have also developed our own compressible LES solver “rhoPimpleMyDyMFOam” on an open source software OpenFOAM ver.5. This solver is applied to our numerical tone holes model and ensured that the functions of mesh manipulation as moving boundary problems with/without topology change work very well. Error rates of pressure at the mesh manipulation is also consistent with this reproducibility. From the view point of numerical reproducibility, the most important result is as follows. When the position of the pads are continuously changed in the order of "open - close", the transitions of resonance frequency is observed in the range of “dgap” value between 0.2 and 0 mm. These results are quantitatively consistent with Keefe’s experimental results.

It is necessary to check that the same results can be reproduced for a 3D model, and can be also reproduced the pitch change in the opening and closing the tone holes of more practical situation. For example, by opening and closing multiple tone holes independently of each other, the conditions of the numerical calculation can be brought closer to the fingering of actual instrument performances. This allows to discuss the situation as fluid sound around the tone holes and the pitch change during actual instrumental performance.
ACKNOWLEDGEMENTS
The present work was supported by JST CREST Grant Number JPMJCR1501, Japan, Grant-in-Aid for Scientific Research (C) Nos. 16K05477 and 19K03655 from the Japan Society for the Promotion of Science (JSPS) and “Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures” and “High Performance Computing Infrastructure” in Japan (Project IDs: jh180007-MDH, jh190010-MDH, and jh220001). Part of the work was carried out under the Collaborative Research Project of the Institute of Fluid Science, Tohoku University.

REFERENCES
A framework for the analysis of bowing actions with increased realisticness

Montserrat PÀMIES-VILÀ¹; Anna SCHEIBLAUER¹;
Alexander MAYER¹; Vasileios CHATZIOANNOU¹

¹ Department of Music Acoustics - Wiener Klangstil (IWK), mdw - University of Music and Performing Arts Vienna, Austria

ABSTRACT

The action of the bow on the string is one of the key aspects of the player-instrument interaction in bowed string instruments. For the analysis of this interaction, it is common to use custom-built artificial excitation mechanisms, where the human arm is substituted by a mechanical setup. Such artificial means allow recreating the excitation mechanism repeatedly. In this study, a robotic arm with 6 degrees of freedom is used to achieve highly precise motion of a cello bow. Preliminary tests have achieved repeatable reproduction of plucked and bowed string sounds using linear trajectories of a plectrum or a bow. In order to take the nuances of the player actions into account, a more realistic bow movement is required. To that aim, the robot might be instructed with bowing trajectories obtained from a real playing situation with expert cellists. These bowing trajectories are recorded using optical motion capture. The motion capture system consists of infrared cameras that track the position of a set of reflective markers attached to the bow, musician, and instrument. The combination of motion-capture data with a fine control of the robotic arm creates a framework for the accurate analysis of the player-instrument interaction in bowed string instruments.

Keywords: Bowed strings, Robotic arm, Cello gesture, Playing technique

1. INTRODUCTION

In music acoustics, artificial playing machines are frequently used to evaluate musical instruments under controlled laboratory conditions. In bowed-string instruments, most machines have been dedicated to explore the sound production in steady state [1, 2] and, only lately, some studies included the analysis of articulation and transitory phenomena [3, 4]. Since the early experiments by Raman [1], bowing devices have been complemented with measurements of several relevant variables, such as bow force and velocity [3, 4, 5]. As an example, Schoonderwaldt [3] and colleagues used a computer-controlled linear bowing machine to bow on a monochord in order to investigate how bowing manipulations affect the violin sound.

In the discipline of performance science, motion capture devices are frequently employed to record body movements so as to analyze the musicians’ gestures during musical performance [6]. Optical motion capture uses infrared cameras to track the position of reflecting markers attached to musicians’ body parts and instruments. To better understand and incorporate the player actions in the analysis of bowed-string instruments, the body motions have been measured and taken into account in few studies considering violin and cello playing [7, 8, 9], particularly those regarding the motion of the bowing arm. Such studies made use of motion capture systems to locate the position of the human body parts and the position of the bow and instrument during playing.

Current research at the Department of Music Acoustics (IWK) seeks at providing a framework for the analysis of the acoustics of bowed-string instruments that considers repeatable and realistic bowing patterns. To that aim, we first use state-of-the-art motion capture technology to track the motion of the bow during real performance with experienced cellists. This data is then used to instruct a robotic arm that holds the bow. In a first test-case, linear bowing patterns that resemble the player actions are used to instruct the robot. This paper shows ongoing research towards achieving...
artificially-articulated bowing strategies in which the motion of the bow is as close as possible to the real-playing situation.

2. METHODS

2.1 Playing tests with cellists

Six advanced cello students were invited to the motion-capture laboratory of the University of Music and Performing Arts Vienna and were asked to play some exercises on an instrumented cello and bow while wearing reflective markers (on their body, on the bow and on the cello), as seen in Figure 1 (left). These tests served at gathering information on their playing style and posture [10] and at collecting 3D data regarding the motion of the bow. Besides motion-capture (frame rate of 240 Hz, OptiTrack infra-red cameras), the sound was recorded with a microphone in front of the cello as well as with a piezoelectric-sensor on the bridge (frame rate of 41000 Hz). Signals were synchronized using a clapperboard. Video recordings are also available to support the analysis of the 3D data.

2.2 Robot control

The robotic arm used in this study is the model UR5e by Universal Robots, which has a weight of 20.6 kg, a reach of 850 mm and it can hold objects up to 5 kg. The bow is attached to the holding-joint of the robot using 3D printed parts. The robot tool-position is given by its 3D position in space (with respect to the origin at its base) and the inclination of the held object is given in axis-angle representation. These are three degrees of freedom for the position and three degrees of freedom for the orientation. The robotic arm can be controlled via a user-friendly environment that allows for a set of custom-made instructions. Alternatively, the tool-position of the robot or the robot joint angles can be driven externally via Real-Time Data Exchange (RTDE) [11].

Using the custom-made instructions approach, one can create motions that may be repeated according to several requirements. As such, this study presents a test-case where linear bowing patterns including velocity and acceleration changes were tested, while accurate increases in bow-to-string distance were introduced, increasing the bowing force on the string.

2.3 Experimental procedure

Inspired by some of the exercises that players performed, we designed a bowing pattern using linear trajectories, which aims at testing the capability of the robotic arm in such bowing tasks (taking player num. 5 as the reference). The robot was instructed with linear motions along the longitudinal axis of the bow (being held at the last joint of the robot). This axis was set to be perpendicular to the D string, the only string used. Following the axis perpendicular to the string and to the holding joint, increments of 0.2 mm where applied and the bowing pattern was repeated 12 times. This means that the bowing pattern was repeated with increments of bow force on the string. Accelerations from 2 m/s² to 15 m/s², velocities from 0.15 m/s to 0.45 m/s and trajectories from 0.09 m to 0.5 m were tested. These tests were carried out in the same room, with the same bow, cello, sensors and reflective markers.
as for the tests with real players. All 3D motion-capture data were low-pass filtered at 6 Hz using a third-order Butterworth filter, which is commonly used when considering slow motions (bow strokes are longer than 1 second) to significantly reduce measurement noise and it is necessary to compute the kinematics of each marker (velocity and acceleration). The motion-capture data showed in this study corresponds to the movement of one marker along the x-axis, which is an axis perpendicular to the strings where the movement of the bow during playing on the D string mainly takes place. The cello was held immobile in the same position for both cases (Figure 1), with its longitudinal axis perpendicular to the x-axis.

3. RESULTS

Two examples of the performed bowing patterns have been chosen, containing four bow strokes starting at the frog. Figure 2 shows position (x), velocity (v), acceleration (a) of one marker mounted on the bow, in front of the index finger (for players) or in front of the holding part of the robotic arm. On the left plots, the portion of the bow used is constant and every stroke is performed at a similar velocity, using the whole bow (50 cm). On the right, the players were instructed to play from pianissimo to fortissimo along four bow strokes. The robot was accordingly instructed to play four strokes using a trajectory of 20, 30, 40 and 50 cm, and every stroke at a higher velocity (from 0.17 m/s to 0.41 m/s), yet the bow-to-string distance, i.e., the force of the bow on the string was kept the same between strokes (7th tested bow-force increment).

![Figure 2: Position (x), velocity (v), acceleration (a) of one marker on the bow, together with the signal of the piezoelectric sensor on the bridge, comparing one player (orange) and the robotic arm (blue). On the left, whole bow strokes, on the right, ascending velocity pattern.](image)

In the whole-bow example (left) one can observe the regularity of the bowing pattern for both player and robotic arm along the x-axis. Regarding velocity, players keep gently increasing velocity during the long note, indicating a preparation for the bow-change [12]. The acceleration peaks reached at the bow change are set at 2 m/s² for the robot, while for the player slightly higher values are observed.

In the velocity-increasing exercise (right), the acceleration peaks are very similar in both player and robot, and are higher than for the whole-bow exercise, achieving values up to 10 m/s². With high acceleration, an additional challenge seems to be the bouncing of the bow held on the robot (see...
changes in amplitude in the piezoelectric signal). One can also observe that players might smoothen or shorten the note release, and achieve quieter note attacks at pianissimo.

Figure 3 shows the superposition of signals obtained at the 12 tested bow forces. These were obtained by adjusting the distance between the bow and the D string, with increments of $\Delta z = 0.2$ mm on the $z$-axis, an axis normal to the plane defined by the bow-hair. Doing so, at every tested condition, the wood of the bow moves slightly towards the string, thus the hair of the bow exerts a higher force on the string. The upper plot shows the position of one marker of the bow along the $x$-axis (as in Figure 2), the middle plot shows the signal of the piezoelectric sensor for all 12 repetitions, and the bottom plot shows the root-mean-square envelope (rms) of these signals. The bowing pattern begins with the frog on the string. The increase in amplitude in the low bow-force plots shows a difference when playing next to the frog or playing next to the string. This indicates a difference in bow force along the bow, which was not directly imposed in these tests. The difference in amplitude disappears at higher bow force, where the sounds are steady after the attack transient (around 0.4 seconds long).

Figure 3: Four bow strokes at 12 bow-to-string distances considering increases of $\Delta z = 0.2$ mm (high bow force in darker color, low bow force in lighter color). The upper plot shows the position along the $x$-axis of one marker on the bow, next to the holding joint at the frog of the bow. The signal of the piezoelectric sensor on the bridge and its rms-envelope are shown with superposition of all tested conditions.

4. DISCUSSION AND PERSPECTIVES

The presented preliminary results show the capability and potential of the proposed framework for the analysis of bowed-string instruments under controlled laboratory conditions. The custom-made instructions approach allows for rapid testing of repetitive motions. The trajectories of the bow were mimicked using linear bow motion, and some of the challenges of the setup were encountered. High acceleration results in bouncing of the bow, showing a lack of damping or bowing pressure. Higher bowing pressures and lower accelerations show more steady sounds. For long notes using the whole bow, as in Figure 3, particularly at high bow force, the setup achieved more steady sounds.

After 1.2 mm increase of bow position pressing on the string, the security limit for the applied force was met. This limit was sufficient to play from soft to loud sounds at the several tested velocities and accelerations. Yet the given security limit might be adapted if higher bow forces are required. With the control of the bow motion in the real-time approach, which allows driving the robot sample-
by-sample, playing nuances as the ones observed in the data of the cellists may be included in the artificial playing setup. The presented methodology, which is for now purely kinematic, might be enhanced if the dynamics of the bow-string interaction are also considered.

ACKNOWLEDGEMENTS

The authors acknowledge the cellists who took part in the experiments, as well as Sara D’Amario for her assistance regarding the motion capture system. This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [T1295-G, P34852-N].

REFERENCES

1. Raman CV. On the mechanical theory of the vibrations of bowed strings and of musical instruments of the violin family, with experimental verification of the results, part I, Indian Association for the Cultivation of Science, 15:243–276, 1918.
A Comparison of Player-Dízi Performance Parameters across Four Members of Dízi Family

Hsien Han NG; Jer-Ming CHEN1
Singapore University of Technology & Design, Singapore

ABSTRACT

The Dízi is a traditional Chinese flute made from bamboo with characteristic bright timbre arising from non-linear interactions due to a membrane stretched over a membrane-hole. Being keyless, the Dízi is also available in different octave and keys, resulting in a large family with numerous representative members. To understand how different Dízi family members compare in terms of player-instrument interaction, we survey the performance parameters of four different Dízi members (two Dízi each of “High” C, G, D & “Low” G keys, played by the same expert performer) by quantifying how pitch and SPL (‘loudness’) vary according to four musical dynamic markings, when sounded across each instrument’s standard range. We also identify regions of the parameter space that is easy or difficult to control (and at what expense: pitch deviation or loudness control?), revealing the expected consistency (and variability) of the Dízi’s performance as played by an expert. The results suggest that certain performance features may scale with the particular Dízi’s size/octave, design and construction.

Keywords: Performance Parameters, Transverse Flute, Player-Instrument Interaction

1. INTRODUCTION

The Dízi is a traditional Chinese flute with characteristic bright timbre arising from non-linear interactions due to a membrane stretched over a membrane-hole. For traditional and cultural reasons [1], it retains a simple design and construction: a hollow bamboo tube (stopped at one end) with six finger-holes (no keys) and is played in a transverse manner like the western flute.

Its construction differs from the western flute, however, featuring a keyless bamboo body integrated with embouchure-, membrane- and finger-holes, such that its precise geometric dimensions are not necessarily standardized (the bamboo stock is a ‘found object’ and remains largely ‘unworked’ for traditional reasons). Being keyless, the Dízi is thus available in different octave and keys, resulting in a large family with numerous representative members, presenting considerable practical challenge to makers, performers, composers, teachers and students.

Consequently, we investigate how output performance parameters of pitch (‘deviation’) and sound pressure level (SPL, ‘loudness’) vary across four different Dízi members (two Dízi each of “High” C, G, D & “Low” G keys, played by the same expert) when performing across each instrument’s standard range at four musical dynamic markings (pp, mp, f, ff), thereby ‘mapping out’ gross similarities and differences of these four Dízi family members, and expectations of scalability (or lack thereof). While previous studies have investigated the effect of dynamic markings on sound pressure output produced by Chinese instruments [2-5], they do not report the relationship between loudness and pitch.

2. METHODOLOGY

Two sets of four Dízi family members (“High” C, G, D & “Low” G) were chosen for this study; the names refer to the reference tonic (‘Do’) sounding respectively C6, G5, D5 and G4 when played using the standard ‘Do’ fingering xxx ooo. All Dízi used were concert-grade instruments, made by different reputable masters and thus we may consider them idealized representations of that family member (this was also anecdotally confirmed by several expert players available to the authors). An expert Dízi player was engaged to play these eight instruments, and the player is a professional engineer.

1 jerming.chen@sutd.edu.sg
member of a traditional musical ensemble, frequent concert soloist, winner of international performance competitions and has over twenty years of performance and teaching experience – again we may thus consider him to be an idealized representation of a typical expert Dízi player.

To understand the relationship between each Dízi’s SPL (‘loudness’) and pitch output, four dynamic level where used: pp, mp, f and ff (where pp here means “as soft as possible” and ff means “as loud as possible” musically). The player was instructed to deliver the full standard tessitura of the Dízi (two octaves plus a major second; Wang, 2014) by performing four sets repeated ascending and descending diatonic scales, holding each note least three seconds. The player was asked to play each note with respect to equal temperament (an electronic tuner provided, A4 = 442Hz) while delivering the assigned musical dynamic level; we do this because pitch and loudness is determined not only by the Dízi, but rather arises from complex interactions in the player-instrument system, especially since the air jet depends on the player’s embouchure and breath control [6, 7]: e.g. distance from the lips to the sharp edge, shape of the lips/embouchure and speed of the jet. Such parameters contribute to the timbre, loudness and fine pitch control in wind instruments [8, 9], and so we wish to investigate how difficult (or easy) it is for an expert to deliver ‘on-pitch’ while assigned a particular musical dynamic level (whilst maintaining musicality).

Audio measurements were collected in a recording studio treated for noise isolation and absorption, with the player located 1 m from the microphone (Zoom H5 Recorder) co-axial with the player’s face. The prevailing sound pressure level (SPL) is calibrated against a constant excitation source of known SPL sited at the same position as the player with respect to the microphone. Pitch and SPL of each note played (segment of stable pitch identified by ear) was extracted using Adobe Audition CC 2018, and its pitch and SPL then extracted to a resolution of 1 cent and 0.1 decibel respectively.

3. RESULTS and DISCUSSION

For the purposes of this extended abstract, we limit ourselves only to the nominal overall behavior of the four Dízi members studied: accordingly we collated the pitch deviation (cents) and SPL (dB) of all notes played and plotted it as Figure 1, with each of the 8 Dízi represented as a separate curve (same color for each family member), and each instrument’s “curve” of 4 data points representing, from left to right respectively, pp, mp, f and ff.

The standard deviation for each data point is bounded by the ‘convex’ orange shaded area and indicates nominally the effective performance parameter space (pitch deviation vs SPL) of the four Dízi members: the typical loudness achieved across all Dízi members studied range from 72–89 dB SPL (each Dízi member’s SPL range is fairly consistent, ~9-10 dB for, regardless of size) while

![Figure 1: Pitch Deviation vs SPL for the four Dízi members (“High” C, G, D & “Low” G keys, 2 Dízi each) studied, collated for four dynamic levels pp, mp, f and ff (points going left to right); Orange shaded region represents nominal boundaries of the mean ± standard deviation for these points.](image-url)
accompanying pitch deviation varied from 0 to +24 cents.

In general, increasing musical dynamic level (increasing SPL) attracted a ‘penalty’ of increasing pitch deviation (‘sharper’) – the largest ‘pitch penalty’ arises going from $f$ to ff (steepest slope). In terms of practical performance, this may suggest it is musically ‘risky’ to deliver ff if pitch is an important consideration; hence if we exclude ff data point, we note only a modest pitch deviation of <10 cents for all Dizi studied.

We also note that High C, G and D share an overlapping parameter space (higher SPL, ranging from ~80–90 dB) than the Low G (lower SPL, from ~72–84 dB, with somewhat greater pitch deviation). This is not surprising, as High C, G and D Dizis often play a “soprano” role as soloist and handling main melodic motifs which cut through the orchestra, while Low G plays “alto” role in accompaniment and harmonic function. The ~4-5 dB difference in gross SPL levels may also provide some insight in terms of dynamic balance constraints often faced when Low G Dizi has to play against ‘soprano’ instruments. Among the three ‘soprano’ instruments, however, we note that in fact it is the G Dizi which is loudest (most “dynamically efficient”), indicating that size (key) of the instrument does not necessarily scale monotonically with loudness; G Dizi also seems to enjoy the least “pitch penalty” compared with the other family members.

In regard to each family member, the two representations (Dizi from different masters) show remarkable consistency in terms differences in SPL range (<1.8 dB) and pitch deviation range (<3 cents). (Granted, the “pitch-centre” of both instruments may not be in agreement, as that depends on the maker’s preference of pitch reference; something even an expert player may not entirely account for.) This observation also implies that Dizi members (of the same key) are rather more similar than Dizi members across different keys, for performance parameters such as pitch and SPL (timbre is not explored in this study).

These insights drawn may inform how performers, composers, and arrangers may balance practical considerations and priorities when they need to choose the most optimal Dizi member to play on. Of course, pitch and loudness are only very simple performance parameters to consider, and we have not yet explored timbre, ease of playing (avoiding cross-fingering or half-holing), note range required, balance and blending, musical context etc., and can be the focus of future studies.

4. CONCLUSION

In this extended abstract, we summarized and report the overall pitch deviation and sound pressure level of how four Dizi family members (two concert-grade examples of each members) behave when an expert player is asked to perform the instrument’s standard tessitura at four dynamic levels – the operating performance parameter space is thus identified and outlined. When playing from soft (pp) to very loud (ff), SPL rises as expected, but pitch deviation also rises (sharpen). We identified that the SPL rise from $f$ to ff may be accompanied with an unfavorably greater pitch deviation (‘pitch penalty’ effect) which is more severe that other dynamic levels. We also identified that the parameter space of Dizi family members do not scale simply with size (key), but instead the soprano instruments overlap and behave somewhat distinctly from the alto instrument, which show an overall 4-5dB lower sound level. These insights to the performance parameter spaces of Dizi family members will be useful for both performers and composers.

REFERENCES
Performance Parameter Survey of the Chinese Sheng

Samuel GOH, Christopher CLARKE, Da Yang TAN; Vincent TAN; Jer-Ming CHEN

Singapore University of Technology & Design, Singapore

ABSTRACT

The Sheng is a Chinese free-reed instrument (mouth organ) with documentation dating back to 1100 BCE, and has extensive related counterparts in other East Asian cultures, e.g. Saenghwang in Korea, Shō in Japan and Khene in Laos. In this study, we survey the full range of a contemporary “alto” Sheng for its performance parameters in terms of loudness vs pitch variation: across its 3 octave chromatic range, for 3 dynamic levels (pp, mf and ff), and for blow vs draw gestures. Additionally, we identify notes showing differences due to playing range, blow vs draw and dynamic levels. Lastly, we relate the output performance parameters in terms of the player’s input of blowing pressure and how it changes with respect to musical dynamic, output sound level and pitch deviation. By characterizing these multidimensional output parameters, we provide for the first time, a systematic description of the Sheng’s performance parameter space, thereby facilitating further documentation, improvement and propagation of the instrument’s design, performance and pedagogy.

Keywords: Performance Parameter, Free-Reed

1. INTRODUCTION

The Sheng is a Chinese free-reed instrument (mouth organ) with documentation dating back to 1100 BCE, and has extensive related counterparts in other East Asian cultures, e.g. Saenghwang in Korea, Shō in Japan and Khene in Laos. As a mouth organ it consists of a series of pipe resonators coupled, with a free-reed coupled to each pipe, and the reeds share a common pressure reservoir (“wind chamber”) which may be driven by both inhaling (draw) and exhaling (blow).

The acoustics of free-reed oscillators have been studied [1-3] and their mechanisms in interaction with the corresponding pipe resonators are now somewhat well-understood. However, investigations of the Sheng as a system of free-reed and pipe resonators, sharing a common pressure reservoir and homogenous design, have not been previously studied under performance contexts.

In this study, we survey the full range of a contemporary “alto” Sheng for its performance parameters in terms of loudness vs pitch variation: across its 3 octave chromatic range, for 3 dynamic levels (pp, mf and ff), and for blow vs draw gestures. We accordingly relate the output performance parameters in terms of the player’s input of blowing pressure and how it changes with respect to musical dynamic, output sound level and pitch deviation. This complements earlier surveys of other Chinese musical instruments [4-6].

2. METHODOLOGY

A concert-grade “alto” (Zhongyin) Sheng was chosen: it plays 3-octaves chromatically from C3 to B5, and is tuned nominally to equal-temperament (A4 = 442 Hz). As a concert-grade instrument, we thus consider this instrument to be an idealized representations of a typical “alto” (Zhongyin) Sheng. An expert Sheng player was engaged to play the instrument, and the player is a professional member of a traditional musical ensemble, frequent concert soloist, winner of international performance competitions and has over thirty years of performance and teaching experience – we may thus similarly consider him to be an idealized representation of a typical Sheng expert player.

To understand the relationship between the Sheng’s relative SPL (‘loudness’) and pitch across its full tessitura, the player was asked to play C3 to B5 chromatically (36 notes), repeating each pitch for 4 sets of draw vs blow gestures, sounded at three dynamic levels: pp, mf and ff (where pp means “as soft as possible” and ff means “as loud as possible” musically, while mf was “moderately loud”),

1 jerming_chen@sutd.edu.sg
holding each note at least three seconds. In many wind instruments, pitch and sound pressure level (“loudness”) are inherently interrelated [7-10] so we wish to understand how these performance parameters interact on the Sheng – a free-reed instrument – and investigate how difficult (or easy) it is for an expert to deliver ‘on-pitch’ while assigned a particular musical dynamic level whilst maintaining musicality.

Audio measurements were collected in a recording studio treated for noise isolation and absorption, with the player-instrument facing the microphone (Røde NT5) 1 m away. The prevailing sound pressure level (SPL) is calibrated against a constant excitation source of known SPL sited at the same position as the player with respect to the microphone. Pitch and relative sound pressure of each note played (segment of stable pitch identified by ear) was extracted using Adobe Audition CC 2018 and extracted to a resolution of 1 cent and 0.1 dB respectively.

3. RESULTS & DISCUSSION

3.1 Dynamic Levels and Pitch Deviation

The pitch deviation for all sounded notes (C3 to B5 chromatically, repeating each pitch for 4 sets of draw vs blow gestures, played at three dynamic levels: pp, mf and ff) are summarized in Figure 1.

![Figure 1. Pitch Deviation (cents) vs Chromatic Notes sounded from C3 to B5 on the alto Sheng, for three dynamic levels (pp, mf, ff) and for four repetitions of draw vs blow gestures.](image)

Of the three octaves surveyed, most of the notes observed – regardless of dynamic level – in the upper octaves have a sounding range of pitches deviating by about Δ10 cents, while broader ranges (Δ20 cents) can be seen in the lowest octave. Some notes seem to suffer from poorer pitch stability, e.g. C#3, G3, D4, E5, where the pitch range approach Δ40 cents.

In terms of ‘tonal’ center, the second octave seems generally sharper, centering around +10 cents, in contrast to the first and third octaves which center around 0 cents.

For wind instruments in general, playing at louder musical dynamics is usually associated with slightly raised pitch, however for the Sheng this dynamic-pitch bias is not generally observed here: in fact, the reverse can be seen, where softer dynamic results in pitch sharpening (e.g. C#3, G3, F4, B4, C5, C#5, F5, B5, etc). At the same time, louder dynamic results in pitch sharpening can still be seen in some instances (e.g. C3, E3, F3, F#3, D4, D#4, A#4, C#5, etc). Thus, as both effects are observed and well-spread across the instrument’s tessitura, it is possible any dynamic-pitch bias is idiosyncratic for that particular reed-pipe configuration, rather than a function of the instrument maker’s deliberate design.
In terms of pitch differences arising from blow (blue) vs draw (red) gestures, there are no clear trends observed, except for the note A4, and D4, where a loud draw is consistently sharper (~10 cents) than a loud blow.

In the second and third octaves, it is also interesting to observe modest local pitch deviations that repeat with octave-level correspondences, for example between D4-G4 and D5-G5, there is a slight rise in pitch centered around E and F. Although the Sheng is expected to nominally tune for equal-temperament [11], there may be some evidence of a non-equal temperament being favored here.

### 3.2 Dynamic Levels and Sound Pressure Level

Figure 2 summarizes the sound pressure levels for all notes (C3 to B5 chromatically, repeating each pitch for 4 sets of draw vs blow gestures) sounded at three dynamic levels: pp, mf and ff.

![Figure 2](image)

**Figure 2.** Relative Sound Pressure (dB) vs Chromatic Notes sounded from C3 to B5 on the alto Sheng, for three dynamic levels (pp, mf, ff) and for four repetitions of draw vs blow gestures. ‘Max’ and ‘Min’ refer to the maximum positive pressure and minimum negative (AC) pressure signal.

In general, we note a clear separation of sound pressure levels according to the musical dynamic delivered – this is not unexpected, as musical dynamic and perceived loudness are inherently correlated. However, it is satisfying to see that the player is able to achieve fairly discrete dynamic regions without much overlap (first octave is the exception, where particularly C3- D3 have extensive overlap between ff and mf). The second octave plays the loudest while the third octave is the softest, suggesting that this alto sheng is designed to be most efficient for notes in the middle of the instrument’s tessitura.

Each note on the Sheng has a dynamic range of ~20-25 dB; within that range, each dynamic level (pp, mf, ff) has up to ~7 dB of variability, indicating the extent to which an expert player has dynamic control.

In terms of blow vs draw gestures, no clear relationships are observed – neither blow nor draw have a systematic influence on sound pressure and in many instances, they can be comparably loud. Furthermore, different notes show different responses with regards to the direct of draw vs blow: “positive bias” (where ‘max’ > ‘min’ pressure magnitude signal, e.g. D#3, F4, G#5 ff), “negative bias” (where ‘max’ < ‘min’ pressure magnitude signal, e.g. D4, D5) or “zero bias” (where ‘max’ ≈ ‘min’ pressure magnitude signal, e.g. F#3, F#4, B4).

### 4. CONCLUSION

In this extended abstract, we surveyed and report the pitch deviation and relative sound pressure
level of a typical concert-grade “alto” (Zhongyin) Sheng when an expert player performed the instrument’s standard tessitura at three dynamic levels – the operating performance parameter space of the instrument is thus identified and outlined.

Unlike other wind instruments, the Sheng – being a free-reed instrument – does not necessarily show sharpening pitch deviation with increasing dynamic level; instead, the reverse can be observed. Nonetheless, we observe that the middle and upper octave of the Sheng has generally very good pitch consistency independent of dynamic level. No systematic difference in pitch was also observed with regard to draw vs blow gestures.

In terms of sound pressure level, we observe the Sheng to be most efficient at the middle of its tessitura. A typical Sheng each note has a dynamic range of ~20-25 dB and an expert player can deliver each dynamic level to ~7 dB. Again, no systematic difference between draw and blow gestures are observed, in terms of dynamic levels. Further, we note three types of ‘pressure bias’ responses are possible, regardless of direction of draw/blow.

These survey of the performance parameter space of the Sheng will be useful for both performers, instrument makers and composers.

REFERENCES
Bowing virtual strings with realistic haptic feedback

Marius-George Onofrei(1), Federico Fontana(2), Silvin Willemsen(3), Stefania Serafin(4)

(1) Multisensory Experience Lab, CREATE, Aalborg University, Denmark, mariusgono@gmail.com
(2) Human Computer Interaction Lab, University of Udine, Italy, federico.fontana@uniud.it
(3) Multisensory Experience Lab, CREATE, Aalborg University, Denmark, sil@create.aau.dk
(4) Multisensory Experience Lab, CREATE, Aalborg University, Denmark, sts@create.aau.dk

ABSTRACT
We present a music interface implementing a bowed string. The bow is realised using a commercially available haptic device, consisting of a stylus attached to a robotic arm. While playing the virtual strings with the stylus reproducing the bow, users feel both the elastic force from the strings and the friction resulting from the interaction with their surfaces. The audio-haptic feedback is obtained by a physical model: four stiff strings are simulated using a finite difference time domain method, modelled as 1-D elements in the virtual 3-D space. The bow is simply modelled as a rigid cylinder that can move free in this space, and interact with the strings. Finally, the frictional interaction between such elements is modelled by a nonlinear friction model capable of reproducing the characteristic stick-slip phenomenon observed during string bowing. Moreover, the model can be dynamically controlled in one parameter so as to become more sticky or slippery. By turning on and off the frictional feedback, users can appreciate the significance of this interaction. A real-time visualisation of the bowed strings complements the audio-haptic display.

Keywords: Musical interface, Digital music instrument, Haptic feedback, Bowed string model, Finite difference

1 INTRODUCTION
In traditional musical performance, there exists a physical coupling between the musician and their mechanical instrument. This follows, as highlighted by Luciani et al. [1], from the energetic exchange between the human instrumental gestures and the sound being produced. According to O’Modharin and Gillespie [2], this energetic flow starts from the mechanical energy injected into the instrument by the performer, whose aim is to produce acoustic energy via the instrument’s resonance. Some of this injected energy is stored in the instrument itself, which is to be reflected back to the performer at some point, e.g. in the form of haptic feedback. They conclude that this back and forth coupling of energy informs “the idea that an instrument becomes the extension of a player’s body”, a feeling that is shared by many musicians according to Nijs et al. [3].

In contrast to this energetic feedback loop interaction, the performance of digital musical instruments (DMIs) typically has a unidirectional energy path, namely from the musician towards the instrument. Such instruments are defined according to Miranda and Wanderley [4] as devices split between two independent units: a control surface and a sound generation unit, linked together by mapping strategies. However, this independence does not have to be the norm and it follows perhaps due to an aim toward generality in terms of possible sounds for a given control surface—at least for the case of software sound generating units, e.g. illustrated by the ubiquity of the MIDI keyboard as a possible controller for almost any software instrument.

DMIs are increasingly becoming a larger part of music production. In the case of pop music, the backbone of a music market sitting at an all time peak [5], synthesizers and samplers provide most if not all of the sounds on many recordings, according to Warner [6]. The accessibility of such instruments as software applications gives rise to large communities of learners and creators who share, play and teach music on web platforms further redefining notions of music making [7] and reinforcing the position of DMIs as part of the general
musical canon. Therefore, one could argue that there should be a greater emphasis on DMIs in the field of
musician-instrument interaction.

As part of a more general trend of haptic based applications in multimedia [8], audio-haptic musical applica-
tions based on interactions with virtual strings have shown a growing interest in the sound and music computing
community. Willemesen et al. [9] used an older version of the Touch haptic device used in the current project,
to control a physical model of a tromba marina—a bowed single-string instrument, in a virtual reality (VR)
environment. The interaction included the elastic response of the string (through the collision between VR ob-
jects) but no frictional feedback during bowing was included. Keeping in the realm of VR, Fontana et al. [10]
as well as Passalenti et al. [11] investigated the effect of haptic feedback with respect to plucked guitar string
simulations. Concerning bowed strings, Sinclair et al. [12] included frictional force feedback in their digital
waveguide based model, but however, it resulted from a different model than the one used for auditory feed-
back. Other custom-made haptic interfaces, including bowing, have been designed by researchers at ACROE in
Grenoble to control physical models built from systems of interconnected mass-springs [13, 14, 15]. A project
presenting a user study comparing the experience of rubbing virtual mass-spring-dampers by means of different
control strategies, including the use of a 3D Systems haptic device, was carried out by Onofrei et al. [16]. The
aim of the different controls was to mimic with increasing accuracy the gestural action of rubbing, culminating
with the presence of frictional haptic feedback resulting from the actual simulation of the frictional interaction,
i.e. the haptic friction force was the same as the force which excited the sound synthesis model. The music
interface appearing in this paper is rooted on the design just described.

In this paper we present an implementation of a real-time physically-modeled music application simulating
the bowing of strings, controlled via an interface which allows not only for intuitive instrumental gestures but
also for realistic mechanical feedback as a result of the physical simulation. This is achieved by making use of
the features of a 3D Systems Touch haptic device: in particular, by mapping the virtual bow-string interaction
point on the tip of its stylus and by rendering both the elasticity and the frictional force of the strings, as
resulting from the interaction of the bow with their surfaces. The strings are placed in a 3-D virtual space in a
trapezoidal cross section, similar to how they would be placed on the neck of a violin. This allows for the bow
to interact either with individual strings but also with pairs of strings at a time, as they lie on the same plane.
Additionally, the displacement of the elements in the 3-D space is visualized at runtime, complementing the
audio-haptic interaction. The setup aims at preserving a consistent link between the control and the synthesized
sound, as suggested by Cadoz [17], with the energetic feedback loop of the instrumental gesture together with
the haptic feedback giving rise to subtle nuances in the timbre and the dynamics of the sound. This link is
greatly facilitated by the use of physical modelling synthesis, as all the different modalities—sonic, tactile and
visual—result from the solution of the same mechanical system.

The rest of this paper is structured as follows: Section 2 presents the physical model behind the audio-haptic
synthesis, Section 3 describes the control interface and introduces the haptic device, then Section 4 presents the
real-time application illustrating the graphical user interface (GUI) and details the mapping between the app
(particularly the physical model behind it) and the haptic device. A discussion follows in Section 5 focused on
the playability of the app, while the final Section 6 concludes the paper and introduces possible future work.

2 PHYSICAL MODEL

This section describes the physical model used in this project and provides some considerations for its discretisation.

Consider a damped stiff string with a circular cross-section of length $L$ (in m) defined for time $t \geq 0$ (in s)
and space $x \in \mathcal{D}$ (in m) with domain $\mathcal{D} = [0, L]$. One can describe its transverse displacement by state variable
$u = u(x,t)$ (in m), and after adding a bow excitation, its dynamics can be described by the following partial
differential equation (PDE) [18, 19]

$$\partial_t^2 u = c^2 \partial_x^2 u - \kappa^2 \partial_x^4 u - 2\sigma_0 \partial_x u + 2\sigma_1 \partial_x \partial_x^2 u - \delta(x-x_B) \frac{f_B}{\rho A} \Phi(\nu_{rel})$$

(1)

where $\partial_t$ and $\partial_x$ describe a derivative with respect to time and space respectively. The parameters are as
follows: wave speed $c = \sqrt{T/\rho A}$ (in m/s) with tension $T$ (in N), material density $\rho$ (in kg/m$^3$), cross-sectional
area \( A = \pi r^2 \) (in \( \text{m}^2 \)), radius \( r \) (in \( \text{m} \)), stiffness coefficient \( \kappa = \sqrt{EI/\rho A} \) (in \( \text{m}^2/\text{s} \)), Young’s modulus \( E \) (in \( \text{Pa} \)), area moment of inertia \( I = \pi r^4/4 \), frequency-independent loss coefficient \( \sigma_0 \) (in \( \text{s}^{-1} \)) and frequency-dependent loss coefficient \( \sigma_1 \) (in \( \text{m}^2/\text{s} \)). The final term in Eq. (1) describes the bow frictional force excitation, \( F_{fr} \) (in \( \text{N} \)), using the following friction model [19]:

\[
\Phi(v_{rel}) = \sqrt{2\pi} v_{rel} e^{-av_{rel}^2 + 1/2},
\]

characterized by the friction parameter \( a \) (in \( \text{s}^2/\text{m}^2 \)). Equation (2) is nonlinearly dependent on the relative velocity between the string—at externally supplied bow location \( x_B = x_B(t) \in \mathcal{D} \) (in \( \text{m} \))—and the bow:

\[
v_{rel} = v_B - \frac{\partial}{\partial t} u(x_B, t) - v_B,
\]

where \( v_B = v_B(t) \) is the externally supplied bow velocity (in \( \text{m/s} \)). The bow is located along the string using the spatial Dirac delta function \( \delta(x - x_B) \). Finally, the externally supplied bow force is \( f_B = f_B(t) \) (in \( \text{N} \)). In this work, the boundary conditions of Eq. (1) are chosen to be simply supported such that

\[
u = \partial_t^2 u = 0, \quad \text{at} \quad x = 0, L.
\]

### 2.1 Discretisation

This work uses finite-difference time-domain (FDTD) methods to approximate the equations above. These methods discretize the continuous system into a grid in space and time according to \( x = lh \) with spatial index \( l \) and grid spacing \( h \) (in \( \text{m} \)), and \( t = nk \) with temporal index \( n \) and time step \( k = 1/f_s \) (in \( \text{s} \)) and sample rate \( f_s \) (in \( \text{Hz} \)). The state variable \( u(x, t) \) then becomes a grid function \( u^n_l \) which describes a grid point with spatial index \( l \) at temporal index \( n \). A full discretisation of the bowed stiff string will not be given here for brevity, but is well covered in the literature (see e.g. [19]). A full derivation of the scheme used in this work can be found in [20, Sec. 8.4]. Instead, several considerations for discrete implementation will be given here. Notice that the various parameters receive a superscript \( n \) to indicate that they are time-varying.

For approximating first-order temporal derivatives there are various options, including the forward, backward and centred differences [19]. The latter can be proven to be second-order accurate and is defined as

\[
\delta_l u^n_l = \frac{1}{2k} (u^n_{l+1} - u^n_{l-1}).
\]

In this work, this approximation is used to discretise Equation (3) yielding

\[
v^n_{rel} = I_l(x^n_B) \delta_l u^n_l - v^n_B
\]

where \( I_l(x^n_B) \) is an interpolation operator retrieving the state of the discrete string at location \( x^n_B \), and is defined as [19]

\[
I_l(x_l) = \begin{cases} 
-\alpha_i(\alpha_i - 1)(\alpha_i - 2)/6, & l = l_i - 1, \\
(\alpha_i - 1)(\alpha_i + 1)(\alpha_i - 2)/2, & l = l_i, \\
-\alpha_i(\alpha_i + 1)(\alpha_i - 2)/2, & l = l_i + 1, \\
\alpha_i(\alpha_i + 1)(\alpha_i - 1)/6, & l = l_i + 2, \\
0, & \text{otherwise},
\end{cases}
\]

with \( l_i = \text{floor}(x_l/h) \) and \( \alpha_i = x_l/h - l_i \). As \( v^n_{rel} \) is used in a nonlinear function (see Equation (2)), the scheme is now implicitly dependent on \( u^n_{l+1} \). In this work, the iterative Newton-Raphson method is used at every sampling step to solve the nonlinear bow function with a maximum of 99 iterations per sample.

The output of the discrete system can be retrieved by selecting a grid point and ‘listening’ to this over time. As the displacement of a medium is related to sound pressure through a temporal derivative [21], the output is retrieved as follows:

\[
\text{out}[n] = I_l(x_o) \delta_l u^n_l
\]

where \( x_o = 0.7L \) is the output location (in \( \text{m} \)).

Finally, the spatial Dirac delta function in Eq. (1) is discretised using a cubic spreading operator \( J_l(x_B) = I_l(x_B)/h \) making changes in the bow location smooth.
Figure 1. (a) Simulation results for the bowing of a stiff string tuned to 196 Hz. Top: String displacement at bowing position, \( w_{bp} \) (in m). Middle: Relative velocity between the string and the bow at bowing position, \( v_{rel} \) (in m/s^2). Bottom: Resulting frictional force at bowing excitation point, \( F_{fr} \) (in N). (b) Shape of the dimensionless nonlinear friction characteristic, \( \Phi(v_{rel}) \) using the friction parameter \( a = 80 \) (in s^2/m^2).

2.2 Prototype model results
A prototype implementation of the physical model was first carried out in an offline setting using Matlab\(^1\). This allowed for more flexibility with respect to investigating the results and calibrating the physical parameters, with their choice based on \([20]\).

Of notable interest was the behavior of the friction model and its capability of producing the stick-slip behavior typical of frictional interaction, which gives rise to the experimentally observed Helmholtz motion of bowed strings \([22]\).

Figure 1a shows the displacement of the string at the bowing position, \( u(x_B) \), with \( x_B = 0.25L \), together with the relative velocity between the bow and the string \( v_{rel} \) as well as the resulting frictional force resulting from the bowing simulation of a steel string tuned to a fundamental frequency of 196 Hz, at a sampling frequency \( f_S \) of 44100 Hz. The tuning is achieved by setting the relevant string tension \( T \) in the model. Other physical parameters of the string are: \( L = 0.5, \rho = 7850, r = 0.0005, E = 2 \cdot 10^{11} \), typical to a steel violin string. Furthermore, the damping parameters are set to zero and the friction parameter is taken as \( a = 80 \), resulting in the shape of the nonlinear friction characteristic \( \Phi(v_{rel}) \) shown in Figure 1b. The string is bowed with a constant bow velocity \( v_B = 0.2 \) and bow force \( f_B = 0.15 \). Units for these parameters were introduced in the previous section. The stick-slip behavior can be seen when looking at \( v_{rel} \), with the value hovering over zero for the most part of the periodic motion, indicating that the bow and the string are stuck together, followed by an abrupt slip in which the magnitude of \( v_{rel} \) greatly increases. This results in a triangular wave shape in the displacement of the string. Additionally, during the slipping phase, the resulting frictional force can be observed to be zero, as there is essentially a loss of contact between the bow and the string, i.e. no interaction.

3 CONTROL INTERFACE
To control the physical model system described in Section 2, a suitable interface needs to be found. The requirements of this interface are: 1) to enable flexibility in terms of instrumental gesture motion, and 2) to allow for the corresponding instantaneous haptic feedback.

\(^1\)https://www.mathworks.com/
3.1 Haptic Device

We found that the Touch professional haptic device manufactured by 3D Systems [23] is an ideal candidate for this task. It is a 6-degree of freedom haptic system equipped with silent internal motors which can provide force feedback in 3-degrees of freedom. The device is illustrated in Figure 2 together with its local coordinate system (CSYS) and the pivot joints B1, B2 and B3 which enable the translation in x, y, z directions of the gimbal joint (equivalent to pivot joint B2) and the 3 possible rotations of the stylus pen. This flexibility of motion in the 3-D space allows for replicating the gestural motion of bowing. In fact, one can hold the stylus in a similar fashion as they would hold an actual bow. Furthermore, the pivot joints A1, A2 and A3 provide force feedback, as opposed to the B joints which are only for rotational degrees of freedom. The combination of the force feedback in the A joints is enough to produce unique haptic response at the gimbal position in each of its three translation directions.

The maximum workspace dimensions of the device, resulting from its mechanical limits are [-210, 210] (in mm) in x direction (width), [-110, 205] (in mm) in the y direction (height) and [-85, 130] (in mm) in z direction (depth). These represent the bounds of the haptic domain, in which the position of the gimbal joint will always lie. This positional information as well as the orientation of the stylus can be accessed from the device using the OpenHaptics API [24]. Another functionality of interest part of the API, relevant to the mapping strategy of the device to the physical model used for synthesis—described in more detail in the next section—is that haptic force feedback can be set at the location of the gimbal joint. This is achieved by the use of callback functions inside a high-priority scheduler thread. For the current application, the refresh rate of these actions is set to 1 kHz, by adjusting the number of times the scheduler ticks its callbacks every second.

4 REAL-TIME APPLICATION

The real-time application was developed in C++ using the JUCE framework. It is open-source and available at [25] and a demo video can be found at [26].

A total of four simply supported damped stiff strings, modelled as per the details given in Section 2, are placed in a virtual 3-D space in a trapezoidal cross-section. By adjusting the tension parameter $T$ the strings are tuned to the fundamental frequencies of the strings of a violin, i.e., 196.0, 293.6, 440.0 and 659.2 Hz, which correspond to the musical notes: G3, D4, A4 and E5 respectively. The approach for tuning as well as typical parameters for such strings can be found in [20]. The strings can be excited by a perfectly rigid bow, with their interaction described by the friction model given in Equation (2). The CSYS of this virtual space is identical to the one of the Touch haptic device. Both the strings and the bow are modelled as infinitely thin lines.

---

2https://www.juce.com/
4.1 Graphical user interface (GUI)

A snapshot of the application during use together with its external control, taken from the demo video, can be seen in Figure 3. The window is divided into 3 areas of interest. On the right, a top view of the strings can be seen, i.e. the strings are projected onto the xz horizontal plane. The long transparent orange rectangle is the bow, whose color is grey when not in contact with any of the strings, e.g. when hovering above them in the 3-D virtual space. Its opacity is directly proportional to the externally supplied bow force, \( f_B \), meaning that when the bow exerts more pressure onto the strings the transparency of its color is reduced, until becoming fully opaque when the force becomes largest. A copy of this same bow and strings can be seen on the top left part of the GUI window, viewed as projected in the frontal plane xy. The maximum displacement of the strings in this view is illustrated as red ellipses placed underneath the static cyan ones, which represent the “at rest” position. Finally, in the bottom left part of the application window, a number of knobs controlling various parameters in the physical model can be found as well as a button which turns on or off the frictional haptic feedback in the control interaction. Going from left to right the knobs control the: damping amount, which is a combination of the frequency in-dependent damping, \( \sigma_0 \in [0.1,10] \), and the frequency dependent damping, \( \sigma_1 \in [0.0001,0.015] \), introduced in Section 2. The choice of combining them is for ease of use for users who are unfamiliar with the details of the physical model underlying the simulation. Their ranges are chosen such that the control of the knob gives either a very large release of the sound or almost none at all. The second knob controls the friction parameter, \( a \in [0.01,15000] \), mapped logarithmically such that the friction interaction between the bow and the strings smoothly goes from “slippery” to “sticky”. Lastly, a dimensionless global volume gain knob is included, controlling the master volume of the app.

Both the strings and the bow elements are depicted in the GUI as having some spatial volume: a width and height in addition to their segment length, forming for example the cross-section of the strings in the upper-left illustration. This is not the case for the virtual 3-D space of the physical model, where they are in fact infinitely thin (1-Delements), thus simplifying their interactions.

4.2 Haptic mapping

As previously mentioned, the stylus of the Touch device is an excellent control interface for a virtual bow. The position of the gimbal joint is mapped to the center of the bow in the virtual physical model space. Its orientation is extracted from the interface of the program with OpenHaptics API as the rotation angles of the stylus relative to the haptic CSYS, shown in Figure 2. For ease of control, the orientation of the bow in the physical model is fixed to be parallel with the frontal plane, i.e. xy, meanwhile allowed to rotate with respect
to the horizontal plane xz. This means the bow is always perpendicular to the strings in the horizontal plane projection. Moving the stylus in the haptic space is equivalent to moving the bow in the physical model space.

In the haptic feedback callback function code, the shortest distance between the bow and each string is calculated at every haptic frame as the shortest line/segment perpendicular to both elements. The exact end point coordinates of this segment can also be found, keeping track of which point lies on the string and which one lies on the bow. Knowing this, the distance can be calculated relative to a direction vector going from the point on the string towards the point on the bow, therefore, when the distance is negative we know the bow is pushing onto the string. An elastic spring force equal to some heuristically chosen spring stiffness times the magnitude of this distance is then sent to the gimbal joint point, with an orientation given by the direction vector. This is how the user can feel each of the strings in the virtual space. When the bow is in contact with multiple strings, the reaction forces from all of them are summed up and sent as haptic feedback. Through linear mapping, this force is directly proportional to the externally supplied bow force in the physical model, $f_B$, with a maximum value $f_B = 2$, hence the force with which the user presses onto the haptic strings gives the first audio synthesis model input. The second input, the externally supplied bow velocity $v_B$ (in m/s), is similarly retrieved from the gestural motion of the user. It is again a linear mapping to some reasonable physical values: between -0.4 and 0.4 m/s, see [27], of the change in position of the gimbal joint in the 3-D haptic space between subsequent scheduler ticks. Essentially it is retrieved from the velocity of the stylus.

During the interaction of the bow with any of the strings, the frictional force resulting from the physical model is sent as haptic feedback to the gimbal joint, split in x, y, z components given by the opposite direction of motion of the bow. As for the case of the elastic feedback, frictional forces resulting from multiple synchronous interactions are summed up.

5 DISCUSSION

The DMI as presented in the previous sections, i.e. the software application as sound generation unit and the Touch device as interface, is capable of producing "good" sounds without much practice. This is most likely due to the mapping choice of the gestural motion of the user to the input parameters of the friction model: the bow force $f_B$ and the bow velocity $v_B$, which are limited in ranges that are relevant for the current physical model. Additionally, the haptic feedback, helps restrain the user from applying large bow forces that result in raucous motion of the string, corresponding to scratchy sound. This is achieved both by the elastic feedback of the strings, as well as the fact that the magnitude of the frictional force—and the haptic feedback it produces, increases with applied bow force. With some practice, the user can achieve also a "whistling" type of sound, resulting from multiple slipping of the bow, with two or more slipping phases per fundamental period [27].

What does take some practice is the ability to easily change from bowing individual strings to pairs of strings. With this, more interesting sounds can be achieved especially by modulating the applied bow force individually for different strings. The ability to bow the strings at a fixed location along their length also requires some rehearsing, as the user needs to get familiar with the mechanical response of the haptic device, independent of the haptic feedback. This reinforces the necessity to use the cubic interpolation operator given in Equation 7, used to place the bow on the string, in order to avoid clicks which would otherwise result from sudden jumps in the excitation location. The real-time visual feedback shown in the GUI helps the user better evaluate their movements in the 3-D virtual space.

Also, once the user becomes more accustomed to the control, variations in the attack and timbre of the sound can be achieved effortlessly. The modulation of the "damping amount" knob can be used to change the release of the sound in real-time and the "friction amount" knob can be used to increase the "stickiness" of the interaction and achieve a more plucked type sound.

6 CONCLUSION AND FUTURE WORK

This paper presents the implementation of a bowed strings musical interface. The audio-haptic synthesis is achieved by physical modelling, with the transverse displacement of a string being described by a PDE encompassing the string's physical properties and an external excitation. This excitation is provided by the contact with a perfectly rigid bow, and is described by a nonlinear frictional model, capable of exhibiting physically
consistent behavior observed in bowed string experiments, particularly the stick-slip interaction between the two elements. The numerical simulation of this system is carried out using FDTD methods and is implemented in a real-time software application written in C++. This app is then controlled via an external haptic device, the 3D Systems Touch, which allows for the inclusion of the haptic feedback resulting from the simulation. The actual frictional force which is used to excite the strings, resulting from the physical model, is felt by the user when bowing the strings. Both the audio feedback and the frictional feedback are directly linked to the instrumental gesture, allowing for a more personal experience with the instrument as opposed to typical DMIs whose control is often disconnected from the underlying audio synthesis model.

Future work can involve including an additional control with the hand which is not holding the stylus of the haptic device. This control could be for instance a haptic glove with which the user could press the different strings against a virtual neck, changing the notes being played or slightly damping the strings and introducing harmonics. A sound-board could be added to the strings, simulating the body of the violin, or perhaps the output sound could be convolved in real-time with the impulse response of the body of a real violin, using partitioned convolution algorithms. Also, spring connections—potentially nonlinear—could be added between various strings, thus creating a more unphysical instrument which could be played in a physical manner. Also of great interest may be developing cheaper alternative haptic devices for the control, as the price of the Touch currently limits the accessibility of the current instrument.

REFERENCES


Presenting a Compendium of Dizi Impedance Spectra

Hsien Han NG; Da Yang TAN; Teck Seng NG; Jer-Ming CHEN

Singapore University of Technology & Design, Singapore

ABSTRACT

The Dizi is a popular Chinese transverse flute with a long-documented history, featuring a distinctive membrane hole (‘mokong’), a bamboo body and typically six keyless fingerholes. Despite its construction from raw bamboo stock with minimal shaping and geometric manipulation, skilled masters are able to consistently create concert-grade instruments. We measured one such Dizi’s acoustic impedance spectra for all standard (and non-standard) fingerings to characterize the instrument’s acoustic response over its tessitura. Consequently, we describe how instrument resonances vary with fingering choice (standard vs alternatives), the effect of the membrane on resonance, and offer practical considerations with regards to performance practice, instrument design, conservation and transmission of the Dizi.

Keywords: Acoustic Impedance, Transverse Flute

1. INTRODUCTION

The Dizi is a popular Chinese transverse flute with a long-documented history, featuring a distinctive membrane hole (‘mokong’), a bamboo body and typically six keyless fingerholes (the Dizi uses a seven-note diatonic scale and relies on “movable Do” solfège notation).

For traditional and cultural reasons [1], it retains a simple design and is constructed from raw bamboo stock with minimal shaping and geometric manipulation. Because of the nature of bamboo (an organic material), the maker is limited by the “found” state of the bamboo blank available to him/her: its internal and external dimensions, taper, uniformity, surface conditions, stiffness and density all vary one from the other. Consequently, skilled masters must rely on experience (prior knowledge, trial-and-error and ‘intuition’) to approach every bamboo blank uniquely to shape it (within construction/design parameters and geometric/material constraints) such that they can consistently create concert-grade instruments that perform and interact with the player with consistent acoustic characteristics (e.g. conforms to standard playing range and fingerings, similar embouchure and breath control responses) and reliable manner (delivers consistent playability, etc.).

To understand the dizi’s linear acoustics, its geometric design and performance implications, we measure a Dizi’s acoustic impedance spectra for all standard (and non-standard) fingerings to characterize the instrument’s acoustic response over its tessitura. Consequently, we describe how instrument resonances vary with fingering choice (standard vs alternatives), the effect of the membrane on resonance, and offer practical considerations with regards to performance practice, instrument design, conservation and transmission of the Dizi.

The impedance spectra measured for these fingerings constitute a database [2,3] of Dizi’s acoustic impedance and sound spectra (here: https://acoustics.sutd.edu.sg/dizi-impedance/), and complements an earlier database of impedance spectra presented on western classical and modern flutes [4,5].

2. METHODOLOGY

For this study, a typical concert-grade Dizi in the key of G (note: being keyless, the Dizi presents as a family of instruments made in different keys) made by a Singaporean master Dizi-maker (co-author T.S. Ng) was chosen. To measure the Dizi’s input acoustic impedance $Z(f)$, the three-microphone two-calibration (3M2C) technique [6] is used to achieve high precision over a wide dynamic range. An impedance measurement head of 7.8 mm internal diameter coupled the Dizi at the embouchure hole [7] in accordance with previous studies on western classical flutes [4,5].

1 jerming_chen@sutd.edu.sg
The fingerings measured in this study were “played” on the setup by one of the authors, an experienced Dízi performer. A total of 35 fingerings (both standard and alternative – Table 1) for 24 semi-chromatic notes D5 to G7 (i.e. typical playing range) were investigated on the G-key Dízi, both with the membrane attached and without (the membrane hole is sealed over using stiff adhesive tape). The impedance measurements were completed in a single 3-hour session, such that ambient conditions and measurement calibration are not expected to vary. To complement the impedance measurement of each fingering studied, an accompanying audio recording was made using that fingering on the Dízi measured, played by the same Dízi performer, and the resulting pitch extracted for its fundamental frequency using Adobe Audition. Importantly, the player was asked to perform each note to produce the most ‘natural’ pitch elicited by the Dízi “as is”, without particular pitch adjustments to accommodate any particular temperament preference in question.

Table 1. Fingering chart for standard and alternative fingerings measured on the G-Key Dízi; ● – closed tone hole, ○ – open tone hole, both ◐ & ● indicate the hole is partially covered (half-hole).

<table>
<thead>
<tr>
<th>Musical Note</th>
<th>Standard Fingering</th>
<th>Alternative Fingering</th>
</tr>
</thead>
<tbody>
<tr>
<td>D5</td>
<td>●●● ●●●</td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>●●● ●○○</td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>●●● ●○○</td>
<td></td>
</tr>
<tr>
<td>F#5</td>
<td>●●● ●○○</td>
<td></td>
</tr>
<tr>
<td>G5</td>
<td>●●● ○○○</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>●●○ ○○○</td>
<td></td>
</tr>
<tr>
<td>A#5</td>
<td>●●○ ○○○</td>
<td>○○○ ●●●</td>
</tr>
<tr>
<td>B5</td>
<td>●○○ ●●●</td>
<td>○○○ ○○○</td>
</tr>
<tr>
<td>C6</td>
<td>○●● ○○○</td>
<td>○●● ○●●</td>
</tr>
<tr>
<td>C#6</td>
<td>○○○ ○○○</td>
<td></td>
</tr>
<tr>
<td>D6</td>
<td>●●● ●●●</td>
<td>○●● ●●●</td>
</tr>
<tr>
<td>E6</td>
<td>●●● ●●○</td>
<td></td>
</tr>
<tr>
<td>F6</td>
<td>●●● ●○○</td>
<td></td>
</tr>
<tr>
<td>F#6</td>
<td>●●● ●○○</td>
<td></td>
</tr>
<tr>
<td>G6</td>
<td>●●● ○○○</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>●●○ ○○○</td>
<td></td>
</tr>
<tr>
<td>A#6</td>
<td>●●○○ ○○○</td>
<td></td>
</tr>
<tr>
<td>B6</td>
<td>●●● ○○○</td>
<td>○●● ●●●</td>
</tr>
<tr>
<td>C7</td>
<td>○●○ ●●●</td>
<td>○○○ ○○○ (Alt1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>●●○ ○○○ (Alt2)</td>
</tr>
<tr>
<td>C#7</td>
<td>○●● ●○○</td>
<td>○○○ ○○○ (Alt1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○○○ ●●○ (Alt2)</td>
</tr>
<tr>
<td>D7</td>
<td>●●● ●●●</td>
<td>○●● ●●●</td>
</tr>
<tr>
<td>E7</td>
<td>●●○ ●●○</td>
<td>●●○ ○●●</td>
</tr>
<tr>
<td>F#7</td>
<td>●●○ ●●○</td>
<td></td>
</tr>
<tr>
<td>G7</td>
<td>●●○ ○○○</td>
<td></td>
</tr>
</tbody>
</table>
3. RESULTS & DISCUSSION

For this extended abstract, we selectively focus on the impedance spectra measured for fingerings associated with three notes as an example: D5/D6 and A#5 (half-hole vs cross-fingering).

Figure 1 shows $Z(f)$ measured for D5 and D6 (‘Sol’ in G key) fingering. Since the Dizi is largely cylindrical, its impedance spectrum for this fingering (all holes closed) is nominally equivalent to a simple open cylinder, hence we observe regularly spaced resonances (minima) to arise (oscillations are sustained at $|Z_{\text{min}}|$ because the embouchure hole opens to the atmosphere, so the air jet excites resonances with large air flow at [low] acoustic pressure [4]). To select either D5 or D6 (circled in Figure 1), the musician adjusts the speed, separation, length and shape of the air jet, as well as adjusting the extent of covering the embouchure hole [8,9]. For this fingering, the first two resonances are often played, although the third and fourth resonances here (A6 and D7 respectively) are also occasionally used for reasons of pitch preference, ease of playing/fingering and tone colour.

**Figure 1.** $Z(f)$ measured for a G-key Dizi for the fingering sounding the notes D5 and D6, with operating resonances ($|Z_{\text{min}}|$) 589.5 Hz (D5+6 cents) and 1205 Hz (D6+44 cents) respectively (circled).

**Figure 2.** $Z(f)$ comparing between half-hole (red) vs cross-fingering (blue) both sounding the note A#5. The four red curves are ‘snapshots’ reflecting the range of continuous pitches available to the player at various degrees of half-holing, showing pitch deviation of Δ53 cents here.
Being keyless, the Dízi allows non-binary finger control over tone hole closure: many subtle modes of partial closure and control are possible (not just open/close). Figure 2 compares half-hole fingering and cross-fingering, for sounding the note A#5. The first resonance (~940 Hz, circled) associated with cross-fingering is flatter in pitch when compared to the half-hole fingering. Figure 2 further shows a representative range of resonance frequencies made available by adjusting the degree of half-hole closure, thereby allowing continuous pitch deviation (‘glissando’) of Δ53 cents (a quarter-tone) to be performed; wider deviations are possible but at the expense of limited timbre and dynamic control. Thus, half-holing can be a useful technique to elicit pitch-bending effect. Another practical consideration: half-hole fingering is sometimes preferred as it requires only engaging one finger (good for fast passages), whereas cross-fingering requires careful coordination across fingers and hands. The trade-off, however, is that fine finger control is required to ensure precise control of pitch during half-holing and consistency over each repetition, whereas cross-fingering delivers consistent pitch.

4. CONCLUSION

35 standard and alternative Dízi fingerings representing the instrument’s tessitura were measured for their acoustic impedance spectra on a concert-grade Dízi, allowing insights on how fingerings determine the operating resonances available on the Dízi and show how each fingering can support several resonances and corresponding sounding pitches; conversely, a number of sounding pitches can be generated using several fingering alternatives. Together, these insights contribute to an appreciation of how expert Dízi players navigate a complex ‘map’ of fingering possibilities, performance parameters (ease of sounding and pitch control) and musical requirements. The compendium of Dízi impedance spectra is available here https://acoustics.sutd.edu.sg/dizi-impedance/ alongside the sound spectra, allowing acousticians, Dízi performers and makers a systematic visualization of the linear acoustics of the Dízi and insight into how fingerings and dizi design translate to acoustic response and its implications on ease of playability, pitch, and practical performance considerations.

ACKNOWLEDGEMENTS

The impedance measurements were conducted at the Acoustics Lab, School of Physics, UNSW, Sydney Australia. The authors are grateful to John Smith and Joe Wolfe for hosting us. The acoustic impedance measurement hardware and software were developed by Paul Dickens and Noel Hanna.

REFERENCES

Violinist, violin, bow: What can we hear and recognize?
Claudia FRITZ; Romain VIALA; Victor SALVADOR-CASTRILLO

Institut Jean le Rond d’Alembert, Sorbonne University, France
Pôle Innovation, ITEMM, France

ABSTRACT
The Science and violin making workshop organized every year by researcher Claudia Fritz and violin maker Paul Noulet in France gathers researchers, instrument makers and musicians to make progress in our understanding of the bowed string instruments. One of the daily sessions consists in critical listening, to train our ears, learn how to express and share what we hear, and search for correlations with construction parameters or study various modifications on the instrument, the bow or the playing technique.

It is generally accepted that the sound of a bowed string instrument depends on three parameters: the player, the instrument and the bow. During the 2021 workshop, formal listening tests were conducted to explore what a group of trained listeners is able to hear and recognize when one of these three parameters varies. Results challenge common ideas among violinists about the sound quality associated to the bow. A large difficulty of comparing instruments when listening was observed, as the musician keeps adapting to the instrument and only few cues are available to the listeners to decouple the player’s actions and the instrument behavior with respect to their influence on the resulting sound.

Keywords: Violin, Bow, Perception

1. INTRODUCTION
Many listening tests have been recently designed in violin acoustics to better understand how listeners evaluate violins as well as explore the influence of various parameters (the age, the model, the soundpost, the thickness of the plate, the effect of playing, etc.) on the perceived sound. While being live or via recordings, they all involved at least one player, playing many times the same short musical excerpt on various instruments. These studies involved a lot of discussions and pilot tests to design experimental conditions and protocols which could counterbalance the influence, on the quality evaluations of the instruments by the listeners, of the player, the reproducibility of the player, the bow that was used, the chosen excerpt, the order of the violins, the types of comparisons, … But to which extent? This study aims at investigating to which extent an instrument has a sound quality per se, independently of the player and the bow, by exploring what can be heard and discriminated within these experimental conditions (beyond the qualitative evaluations for which a large disagreement between listeners has been shown by these studies).

We took advantage of the gathering of violin and bow makers and scientists during a European workshop ‘Science and lutherie’ organized yearly by the first author and violin maker Paul Noulet. Critical listening sessions are part of the workshop daily routine, which thus make this group of people a relatively well trained audience. These sessions have many goals: ear training, negotiating some common vocabulary, comparing instruments and testing the influence of various modifications which are done during the workshop in order to correlate them with measured acoustical differences. We used these sessions during the 2021 workshop to design a series of listening tests to investigate what, among the three factors that are the player, the bow and the violin, can be recognized and discriminated.

1 claudia.fritz@sorbonne-universite.fr
2. LISTENING TESTS

2.1 General design

All the listening tests were designed as N successive trials, during which one or two violins were played by one or two musicians, behind an acoustically transparent screen, on stage of a 340-seat auditorium. For each test, the participants (between 11 and 12 violin/bow makers and violin acousticians) had to answer the same yes-no question. The probability of a correct answer being 50% for each trial, the distribution of answers for a given participant follows a binomial law with $p=1/2$. Based on how this law varies with N, we considered that N=8 was an interesting number as the probability of making one or no mistake is below the usual 5% chance threshold. Therefore, we can rule out that a participant did the test by chance if they made no more than one mistake.

The downside of such design is the repetition. Both the audience and the players can get tired over the length of the test. This tiredness can lead the players to be less reproducible. But at the same time, the repetition can make them adapt more and more to obtain the sound that they want, thus smoothing the differences. This will be discussed more later.

2.2 Violin recognition

A first series of tests was conducted to investigate whether a trained audience could reliably distinguish between two violins and could recognize specific characteristics of an instrument across players and trials.

Tests 1a et 1b

For each trial, two violins were played successively. They were chosen randomly among a group of 8 violins. The participants had to tell whether they were two different violins or the same one played twice. In test 1.a, the two violins were played by two professional orchestra players, using their own bow. In test 1.b, they were played by only one of these two players, the same one for all trials. After a few pre-tests, the chosen musical excerpt consisted in only 9 notes, a G Major diatonic scale on the G string with 5 notes up and 4 notes down, to keep the length of the test reasonable and, so we thought, to reduce the difficulty with a short excerpt, which is not too musical.

Test 2a and 2b

Here, only one violin is played per trial, which shortens the test by more than half and thus reduces the tiredness and the problems of reproducibility. The violin played was selected randomly among two violins, labelled A and B, which were chosen collectively before the test as being the most different among the 8 that were available. So for each trial, participants had to tell whether it was violin A or not (i.e. violin B). The violins were played by a semi-professional violinist, who played an ascending and descending A Major scale (covering the register of the violin) followed by the opening of the Tchaikovsky concerto. We thought that the test being reduced, we could afford hearing a bit more and thus have more cues to make our judgment on. In test 2a, the player played with a different bow (selected randomly among a pool of 10 bows) for each trial to limit the adaptation of the player while in test 2b, the player always used his own bow, to reduce the source of variation.

2.3 Bow recognition

Talking about the famous recording by Janine Jansen of 12 Stradivari for which she chose carefully which bow to use among her two bows for each violin, Pauline Harding wrote: “Her two bows had a dramatic effect. It was interesting how everything changed.” (6) This resonates with beliefs in the community of violinists and bow / violin makers who claim that changing bow can completely change the sound of an instrument, and even that a bow has a sound per se. We thus designed two tests to study whether an audience could reliably recognise a bow. As pre-tests made us realise that it may be much harder than expected based on the beliefs, we decided to use two extreme bows: a violin bow made by a renown bow maker and a middle-range viola bow (about 10g heavier).

Tests 3a and 3b:

For each trial, a violin was played with either the violin or the viola bow, by the same semi-professional violinist as in test 2. The musical material included a Ab Major scale covering the register...
to the violin, with one bow stroke per note on the way up and with an alternating of six rapid and staccato up and down bows (to illustrate the attacks allowed by the bow) on the way down followed by a very short excerpt from a Bach sonata. Similarly to what was done in test 2, test 3a was done with a different violin for each trial (chosen randomly among the pool of available violins) while test 3b was done on the player’s own violin.

3. RESULTS AND DISCUSSION

The number of participants among the total number of participants who did each test at better than chance level is reported in table 1.

Table 1: Proportion of participants who succeeded better than 5% chance level, i.e. who made no more than one mistake for each test.

<table>
<thead>
<tr>
<th>Test</th>
<th>1a</th>
<th>1b</th>
<th>2a</th>
<th>2b</th>
<th>3a</th>
<th>3b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants who succeeded better than chance level</td>
<td>1/11</td>
<td>1/11</td>
<td>5/11</td>
<td>5/12</td>
<td>3/12</td>
<td>2/12</td>
</tr>
</tbody>
</table>

Table 1 shows that for each test, only a very limited number of participants were able to succeed at better than chance level. The results of test 1.a seem to indicate that the same violin played by different players can have very different sound qualities which makes it difficult to tell whether it is the same violin or two different ones. Test 1.b’s results show that it is not easier when it is the same player. This seems to indicate that the way a violinist play can have as much influence on the timbre than playing different violins. It is difficult for a violinist to play in a strictly reproducible way along the course of the test because of tiredness but more importantly because of adaptation. A good player is always trying to achieve “his/her” sound and so will modify his/her technique to get the sound he/she wants. Therefore, the differences decrease along the test and so it becomes harder for the audience to tell apart the violins. We thought that this adaptation could be limited by playing less the violins, as in test 2 (only one violin per trial instead of 2) and by playing them with different bows (2a) rather than one bow (2b). This seems to be the case, though not to a large extent, with still fewer than half of the participants being able to succeed. The protocol with only one violin in each trial allows a shorter test, but at the same time, may be more difficult because participants need to remember which violin is violin A. If a participant loses track at some point during the experiment, he may well be wrong for the rest of the test. This could be possibly solved in future tests by providing participants with the correct answer after each trial.

The last two tests, about bow recognition, have not been able to show any “bow sound” with only very few people having been able to distinguish two very different bows at better than chance level. Though players claim that the bow can have a large influence on the sound, one can expect, from a mechanical point of view, that this influence will be rather on temporal aspects (like the precision of the attacks or the articulation between notes for instance) rather than on the timbre of the instrument (7). Therefore, one may expect a smaller influence on the resulting sound from the bow than from the violin, and it is thus not surprising that participants could not recognize bows if they were not able to recognize violins.

4. CONCLUSION

These series of listening tests are a first attempt at exploring, beyond any qualitative judgement, what can possibly be recognized in a reliable way by a trained audience when violins are played in similar experimental conditions as previous published studies on violins sound quality, involving the repetition of rather short excerpts on various violins. The results raise interesting questions about the timbre of an instrument: can a violin have some intrinsic sound qualities that can be recognized across players and bows and if yes, how can we access/perceive them in a reliable way? Further work and
more listening tests are definitely needed. They could be of significant importance in the current context about the shortage of the Pernambuco wood (traditionally used for bows), which is still used, despite more and more CITES regulations attempts (8), since makers and musicians claim it gives a “better sound”.

ACKNOWLEDGEMENTS

We would like to thank the “Villefavard team”, the dedicated group of violin makers who took part in all the discussions and pre-tests to design the protocols and then participated in the tests.

REFERENCES

8. CITES. Consideration of Proposals for Amendment of Appendices I and II. Nineteenth Meeting of the Conference of the Parties, 2022. https://cites.org/sites/default/files/eng/cop/19/prop/as_received/E-Pauabrasilia_echinata.pdf