Acoust. Sci. & Tech. **22**, 2 (2001)

PAPER

Sound timbre interpolation based on physical modeling

Takafumi Hikichi and Naotoshi Osaka

NTT Communication Science Laboratories,
3–1, Morinosato Wakamiya, Atsugi, 243–0198 Japan

(*Received 12 April 2000, Accepted for publication 24 August 2000*)

Abstract: Our goal is to develop sound synthesis technology that users can synthesize arbitrary sound timbre, including musical instrument sounds, natural sounds, and their interpolation/extrapolation on demand. For this purpose, we investigated sound interpolation based on physical modeling. A sound-synthesis model composed of an exciter, a one-dimensional vibrator, and a two-dimensional resonator is used, and smooth timbre conversion by parameter control is examined. Piano and guitar sounds are simulated using this model, and interpolation between piano and guitar tones is investigated. The strategy for parameter control is proposed, and subjective tests were performed to evaluate the algorithm. A multidimensional scaling (MDS) technique is used, and perceptual characteristics are discussed. One of the axes of the timbre space is interpreted as spectral energy distribution, so the spectral centroid is used as a reference to adjust parameters for synthesis. By considering the centroids, smoothly interpolating timbre is achieved. These results suggest the possibility of developing a morphing system using a physical model.

Keywords: Sound interpolation, Physical modeling, Struck strings, Plucked strings, MDS, Spectral centroid

PACS number: 43.75.Wx

1. INTRODUCTION

Physical modeling synthesis is now becoming one of the most promising methods used to simulate musical instrument sounds. Since the artificial instrument can have the same control parameters as the real one, the users can control its timbre more intuitively than other abstract methods. Many artificial instruments have recently been developed [1–3]. Researches on flexible model structures [4, 5] and on cost-effective algorithms for real time implementation [6, 7] have been reported. Now there are various synthesis systems with expressive control.

Controlling timbres of different musical instrument sounds and between them using physical models, however, has not been attempted so far. Such techniques, called timbre morphing, have been attacked by signal-based methods. Several researchers, including one of the authors, have discussed morphing or sound interpolation techniques [8–11]. In the field of speech synthesis, speech morphing has been investigated [12–14]. Morphing based on the sinusoidal model or other analysis-based methods is done by first interpolating model parameters extracted from the two sounds, then resynthesizing using the interpolated parameters. The advantage of such an approach is that almost any sound can be handled and that a variety of timbre can be achieved. The main disadvantage is that too many parameters must be handled for resynthesis, and transient parts of a signal are difficult to treat.

Some morphing algorithms are already used in software tools [10], and some computer music composers are making use of them. However, evaluations in terms of timbre perception have not been necessarily performed. Jaffe [15] has outlined the evaluation criteria for synthesis techniques, but comparative study has been insufficient. As a result, the user must select parameters after a tedious trial-and-error procedure in order to obtain the desirable timbre. The situation is that intuitive control is difficult.

In the field of psychoacoustics, the perceptual space for timbre has been derived by subjective tests using various acoustic instrument and synthesized sounds [16–18], and several features which affect the timbre perception have been investigated. As an evaluation framework, a multidimensional scaling (MDS) technique is usually used. In this paper, MDS technique is utilized for evaluation.

Our goal is to develop sound synthesis technology that users can synthesize arbitrary sound timbre, including musical instrument sounds, natural sounds, and their interpolation/extrapolation on demand. We intend to utilize the controllability of physical models in order to apply a sound morphing system to various sounds. The morphing techniques include 1) extracting the model parameters from the original signals, 2) modifying the parameters, and 3) synthesizing the signals. We are primarily concerned here with smooth timbre control using a physical model.

According to an approach to physical modeling synthesis, sound sources are physically modeled, and used for synthesis. Here, interpolation of sound source is considered. For two sound sources having the same production mechanism, sound interpolation can be achieved by simply interpolating different physical parameters. For two sources having a different production mechanism, an integrated model that includes the different models is considered (referred to as "structural interpolation"). According to this approach, all intermediate sounds have the production mechanism, and as a result, sounds of natural and homogeneous quality are expected. Furthermore, the number of parameters needed for synthesis is generally smaller than that of signal-based methods. The major disadvantages are that a model or algorithm must be built first and that the model limits the timbre range that can be produced.

The purpose of this paper is to propose a synthesis algorithm that achieves smooth and gradual timbre conversion from one timbre to the other, *i.e.* timbre interpolation. Interpolation approaches are twofold: one is "structural" and the other is "characteristic." Piano and guitar sounds are selected as two targets which are represented by a unified physical model, and the interpolation between them is investigated. Both have a similar mechanism: the strings are excited by an object, and the vibration of the strings propagate to a resonator, and radiate into the air. The key idea for interpolation is that by properly adjusting the parameters, two different timbre can be synthesized by one model, and that smooth transition from one timbre to another may be possible.

In the next section, the physical model is briefly described. Then the strategies for smooth interpolation are presented, and the algorithm is evaluated by subjective tests. The relationships between physical and perceptual characteristics are investigated. Modification of the algorithm is also introduced, and subjective tests are performed to evaluate this modification. We discuss the experimental results and criteria for timbre interpolation in section 4. Section 5 concludes the whole paper.

2. PHYSICAL MODEL

2.1. Model Structure

For the purpose of synthesis, cost-effective algorithms which are tuned for real time processing are well-known. However, in order to clarify the relationship between physical parameters and synthesized tones, we use the classical method based on numerical solutions of differential equations, and assume a simple model which is composed of an exciter, a vibrator, and a resonator connected in series (Fig. 1). The present model describes the transverse, one-dimensional vibration of a string/strings and that of a plate generated by an exciter. This can be regarded as one of the simplest models for the piano.

The exciter and the vibrator model used here is basically the same as that reported by Hiller and Ruiz [19,20], and further elaborated by Chaigne and Askenfelt [21, 22] for the simulation of the vibration of a piano string struck by a hammer. This struck-string model is modified to enable us to interpolate sound to obtain a timbre between striking and plucking. The resonator model represents the transverse vibration of the plate.

The next three subsections describe each component of the physical model.

2.2. A Vibrator Model

The present model describes the transverse motion of a one-dimensional vibrator with damping, which is struck/plucked by a nonlinear hammer/plectrum. The vibrator includes strings and bars, *i.e.* elastic media.

The vibrations are governed by the following equation:

$$\frac{\partial^2 y}{\partial t^2} = \frac{T}{\mu}\frac{\partial^2 y}{\partial x^2} - \frac{\kappa^2 ES}{\mu}\frac{\partial^4 y}{\partial x^4} - 2b_1\frac{\partial y}{\partial t} + 2b_3\frac{\partial^3 y}{\partial t^3} + f(x, x_0, t), \tag{1}$$

where y is string displacement, μ is line density, T is tension, E is Young's modulus, κ is the radius of gyration, S is the cross-sectional area, b_1 and b_3 are damping coefficients, $f(x, x_0, t)$ is force density, and x_0 is the distance of the hammer from one end of the string. Stiffness and damping terms are included. The two partial derivatives of odd order with respect to time simulate a frequency-dependent decay rate of the form,

$$d(\omega) = b_1 + b_3\,\omega^2,$$

where ω denotes angular frequency.

The force density term $f(x, x_0, t)$ represents the excitation by a hammer, a plectrum, or fingers. This term is limited in time and distributed over a certain width.

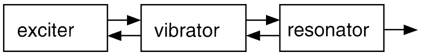


Fig. 1 A block diagram of a physical model used in the experiment.

The string is assumed to be hinged at one end and connected to a resonator at the other end, which corresponds to the following four boundary conditions:

$$y(0,t)=0, \quad y(L,t)=z(x_1,y_1) \tag{2}$$

and

$$\frac{\partial^2 y}{\partial x^2}(0,t)=\frac{\partial^2 y}{\partial x^2}(L,t)=0, \tag{3}$$

where $z(x_1,y_1)$ is resonator displacement and (x_1,y_1) is the point at which the plate is connected to the end of the string.

2.3. Excitation Model

According to Ref. [21], the motion of the piano hammer and the collision process of the hammer with the string are described as

$$M_\mathrm{H}\frac{d^2\eta}{dt^2}=-F_\mathrm{H}(t), \tag{4}$$

$$F_\mathrm{H}(t)=\begin{cases} K|\eta(t)-y(x_0,t)|^p, & \eta(t)\geq y(x_0,t) \\ 0, & \eta(t)<y(x_0,t) \end{cases} \tag{5}$$

where η is hammer displacement, $F_\mathrm{H}(t)$ is hammer force, and M_H is hammer weight. Coefficients K and p are determined experimentally. The relationship between $f(x,x_0,t)$ and $F_\mathrm{H}(t)$ is

$$f(x,x_0,t)=\frac{F_\mathrm{H}(t)g(x,x_0)}{\mu\int_{x_0-\delta x}^{x_0+\delta x} g(x,x_0)dx}, \tag{6}$$

where $2\delta x$ is hammer width, $g(x,x_0)$ is force distribution along the string, and μ is line density of the string. This is the struck-string model used in Ref. [21]. The duration of contact between the hammer and string is derived automatically by solving Eqs. (1)–(6). We refer to this as force duration time in the struck-string case t_st.

On the other hand, the most primitive model for plucking is described by specifying the initial shape of a string. Recently, more elaborate physical models of the plucking process have been reported [23, 24] and some have been used for sound timbre control [25, 26]. They are based on mass-spring representation. Since they are developed independently with a hammer-string interaction model, the struck-string model and the plucked-string model have little in common. Here, we will explain how to extend the struck-string model to the plucked-string model.

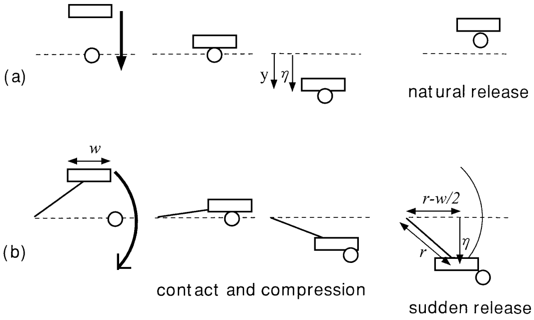


Fig. 2 Behavior of an exciter and a string in the striking and the plucking motions. The rectangle represents an exciter, and the circle represents a string. The dotted line means the equilibrium position of the string. (a) Struck case. The hammer moves straight down (or up). (b) Plucked case. The solid line represents a rod, and the plectrum moves in a circular motion with the rod.

The most significant difference between striking and plucking appears at the end of an excitation force signal applied to a string. When a string is struck, the string is pushed downward by a hammer, as shown in Fig. 2(a). Compressive force given by Eq. (5) is exerted on both the string and the hammer, as depicted by “contact and compression.” Then, the string pushes back the hammer, and when the distance between the string and the hammer becomes zero, the hammer releases from the string naturally, as shown by “natural release” in Fig. 2(a). When a string is plucked, on the other hand, the finger or plectrum is pulled off the string suddenly and the excitation force becomes zero at the end of the contact period, as depicted by “sudden release” in Fig. 2(b). Hence, another force duration time t_f, specified by users, is introduced to the conventional model. The proposed excitation model is expressed as

$$F_\mathrm{H}(t) = \begin{cases} K|\eta(t)-y(x_0,t)|^p, & \eta(t)\geq y(x_0,t) \text{ and } t<t_\mathrm{f} \\ 0, & \text{others} \end{cases} \tag{7}$$

If the value of t_f is large enough to satisfy $t_\mathrm{f}\geq t_\mathrm{st}$, where t_st denotes force duration in the struck-string case, Eq. (7) expresses the conventional struck-string model shown in Eq. (5). On the other hand, when t_f is set so that $t_\mathrm{f}<t_\mathrm{st}$, force signal is truncated before natural release, and becomes zero at $t=t_\mathrm{f}$ like a step function. This represents one of the characteristics of plucked excitation. In the experiment reported later, it will be shown that the proposed model can treat both cases and intermediate conditions continuously by controlling t_f and other parameters. We refer to this as “structural interpolation,” which means

structural differences between struck and plucked string are focused on. Frictional force, which may be exerted on the exciter and the string, is not considered in this model.

2.4. A Resonator Model

A resonator model used here is a rectangular plate with supported boundaries. This is one of the simplest approximations for a piano soundboard. The soundboard is connected to one end of a string/strings meeting at a point.

The vibration of the plate connected with the string/strings is described as

$$\frac{\partial^2 z}{\partial t^2} = -\frac{\kappa^2 E}{\rho(1-\nu^2)}\nabla^4 z - 2b_1\frac{\partial z}{\partial t} + 2b_3\frac{\partial^3 z}{\partial t^3} + \frac{F_s(t)}{\rho h}\delta(x-x_1)\delta(y-y_1), \quad (8)$$

where z is plate displacement, ρ is density, ν is Poisson's rate, E is Young's modulus, κ is the radius of gyration, h is thickness, $F_s(t)$ is the force exerted from the end of the string/strings on the plate, $\delta(x)$ is Dirac's delta, and (x_1,y_1) is the junction between the strings and the plate. The force $F_s(t)$ is derived by

$$F_s(t) = T\left.\frac{\partial y}{\partial x}\right|_{x=L}. \quad (9)$$

2.5. Numerical Solution

Equations (1)–(4) and (6)–(9) can be digitized by using an explicit finite difference scheme, which lead to the recurrence equations. The velocity signal at the junction between the string/strings and the plate, which corresponds to a bridge, is calculated and used as synthesized sound. The recurrence equations are not shown here for simplicity.

When digitizing the continuous equations, the appropriate number of segments N must be chosen. For a standard explicit finite different scheme, stability and numerical dispersion requirements determine the appropriate value. Using the Fourier's method, Chaigne has shown this value in the lossless case ($b_1 = b_3 = 0$) for a string ($T \neq 0$) [23]. Nakamura has derived the appropriate N when considering frictional damping coefficient b_1 [27]. Here, both b_1 and b_3 are considered and the optimum N value is determined. After some calculations, the solution of the following equation

$$\frac{\kappa^2 SE\Delta t^2}{\mu L^4}N^4 + \frac{T\Delta t^2}{\mu L^2}N^2 - 1 + b_1\Delta t - \frac{3b_3}{2\Delta t} \leq 0 \quad (10)$$

gives the maximum number $N_{\max}$, and the optimum N value is determined as a maximum integer which is less than $N_{\max}$. The value of $N_{\max}$ varies with other parameters, so N is calculated by other physical parameters for synthesis. The number of plate segments is also determined so as to fulfill the stability requirement.

The next section describes attempts to control parameters in order to achieve smoothly interpolated sounds.

3. EXPERIMENT

3.1. Parameter Fitting

The problem of parameter estimation from acoustic signals is beyond the scope of this paper. Here, some parameters are found in Ref. [21], gained experimentally, and others are selected by trial-and-error procedures. Subjective listening tests were carried out, and the FFT spectra of recorded sounds and those of the synthesized sounds were compared. In fitting the resonator parameters, our intent was not the precise modeling of the piano's plate or the guitar's body. Instead, overall frequency and time characteristics are imitated. The comparison of the FFT spectrum of piano sounds with that of guitar sounds showed that the piano's spectrum has denser peaks, and reaches higher frequency. It also showed that, in terms of time characteristics, the piano has longer reverberation than the guitar. Therefore, relatively small values were used for the damping coefficients of the plate for the piano to simulate its longer reverberation. For the guitar, larger damping coefficient values were used to ensure the frequency modes of the resonator are sufficiently damped. After several trials, parameter sets for a piano-like struck-string sound and a guitar-like plucked-string sound were determined.

Then, the piano tone was simulated using the fitted resonator and three slightly detuned strings. This resulted in the production of beats, which is one of the most salient features of real piano tones.

For the simulation of the guitar, the lowest mode frequency of the resonator was fitted to 100 Hz by modifying density and damping coefficient parameters. This corresponds to the Helmholtz resonance of the guitar body. The stability condition requires a smaller N value than that used in synthesizing the other sounds in consideration, a different N was used for guitar synthesis.

3.2. Parameter Control

Consider two different media having different physical parameters, and interpolated sounds are synthesized using these two parameter sets. Here, the interpolation ratio $\alpha (0 \leq \alpha \leq 1)$ is an index showing how close or far the synthesized sound is to the timbre of the two endpoints. Here, $\alpha = 0$ and $\alpha = 1$ represent the timbre of the endpoints, respectively. Interpolation was controlled by parameter α.

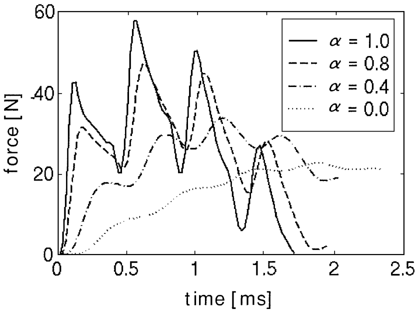


Fig. 3 Force signals with various values of α. $\alpha = 1.0$ corresponds to striking (solid line), and natural release occurs at 1.7 ms. $\alpha = 0.0$ corresponds to plucking (dotted line), and sudden release occurs at 2.4 ms. By changing parameter α, smooth interpolation is achieved.

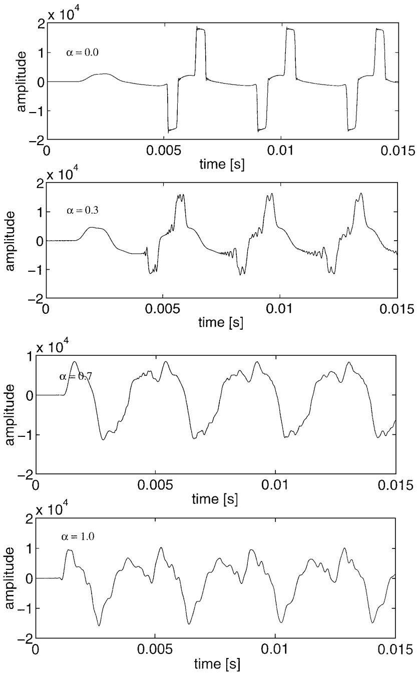


Fig. 4 Velocity waveforms at a point on a string for $\alpha = 0.0$, 0.3, 0.7, and 1.0 are shown. $\alpha = 0.0$ corresponds to a plucked-string case, and $\alpha = 1.0$ corresponds to a struck-string case. For $\alpha = 0.0$, a pulse-like feature is clearly seen. This feature is one of the characteristics of a plucked-string waveform.

3.2.1. Interpolation of excitation condition

Subscripts st and pl mean struck and plucked, respectively. There are four parameters for the excitation condition, namely, M_{H}, K, p, and t_{f}. All the four parameters are related to the amplitude, shape, and the duration of the force. Since proper values of these parameters are unknown, and since it is desirable that the number of control parameters be small, M_{H} and K were fixed and p and t_{f} were interpolated for simplicity. For the plucking condition, t_{pl} was simply fixed to $t_{\mathrm{st}}/2$. This condition was determined after several trials using various values. Strictly speaking, however, the duration time t_{pl} could be determined by the displacement of the exciter η, the width of the exciter w, and the radius of circular motion r, as shown in Fig. 2(b). When the horizontal displacement of the exciter exceeds $w/2$, release occurs and the time duration t_{pl} is calculated. When the interpolation parameter $\alpha = 0.0$ and $\alpha = 1.0$ correspond to plucking and striking, interpolated parameters are calculated by

$$p(\alpha) = \alpha p_{\mathrm{st}} + (1-\alpha) p_{\mathrm{pl}}, \qquad (11)$$

$$t_{\mathrm{f}}(\alpha) = \alpha t_{\mathrm{st}} + (1-\alpha)\frac{t_{\mathrm{st}}}{2} = \frac{1+\alpha}{2} t_{\mathrm{st}}. \qquad (12)$$

An example of force signals for various α values is shown in Fig. 3. By changing parameter α little by little, the force signal varies gradually. For $\alpha = 1.0$, several peaks occur as a result of multiple impact with reflected waves from the end of the string. This feature has been found in conventional hammer-string interaction models [21]. For $\alpha = 0.0$, it is reasonable that the release occurs around the maximum point in the force function, as depicted in Fig. 3. Figure 3 shows that an interpolation is achieved in the force domain.

Velocity waveforms at a point on a string for $\alpha = 0.0$, 0.3, 0.7, and 1.0 is shown in Fig. 4. For $\alpha = 0.0$, a pulse-like feature is clearly seen in the velocity waveform. This feature is one of the characteristics of a plucked-string waveform.

3.2.2. Interpolation of vibrator and resonator parameters

Damping coefficients of the vibrator b_3, and those of the resonator b_1 and b_3 were interpolated linearly between the two endpoints. In order to avoid variations of frequency partials, the other parameters were fixed.

In the next subsection, effects of changing the damping parameters on the synthesized tones are investigated, and strategies to implement smooth interpolation are examined.

3.3. Strategies for Smooth Interpolation

This subsection describes the strategies to implement timbre interpolation between simulated piano and guitar tones. Subjective similarity tests were conducted in order to evaluate the strategies, and the relationship between physical characteristics of synthesized tones and perceptual spaces derived from the tests were examined. Preliminary test results show that a simple linear interpolation of all the parameters does not work well, as expected.

In order to implement smooth interpolation, the tran-

Table 1 Values of the control parameters used for sound synthesis. t_{nr} shows time duration until natural release occurs, and determined by simulation. The symbol — means that the value is the same as in the left cell.

Parameters	piano	struck	plucked	guitar	plucked 2
Exciter striking position i	0.12	—	—	—	—
hammer mass M_{H} [kg]	2.97×10^{-3}	—	—	—	—
stiffness coefficient K	4.5×10^{9}	—	—	—	—
initial velocity V_{H} [m/s]	5.0	—	—	—	—
stiffness exponent p	2.5	—	3.5	—	—
force duration t_{f} [s]	t_{nr}	—	$t_{\mathrm{nr}}/2$	—	—
Vibrator length L [m]	0.62	—	—	—	—
radius a [m]	5.0×10^{-4}	—	—	—	—
density ρ [kg/m^3]	8.07×10^{3}	—	—	—	—
tension T [N]	1,058.9	—	—	—	—
Young’s modulus E [N/m^2]	2.0×10^{11}	—	—	—	—
damping coefficient b_1	0.5	—	—	—	—
number of string segments N	61	—	—	—	—
tuning difference [cent]	0.9	0	—	—	—
damping coefficint b_3	3.0×10^{-9}	—	4.0×10^{-8}	—	1.0×10^{-9}
Resonator length (L_x, L_y) [m]	(1.0, 1.0)	—	—	—	—
thickness h [m]	2.6×10^{-2}	—	—	—	—
Young’s modulus E [N/m^2]	1.0×10^{10}	—	—	—	—
Poisson’s rate ν	0.3	—	—	—	—
position of connection (x_1, y_1)	(0.24, 0.36)	—	—	—	—
density ρ [kg/m^3]	6.0×10^{4}	—	—	1.4×10^{3}	6.0×10^{4}
damping coefficient b_1	10.0	50.0	—	5.0	5.0
damping coefficient b_3	1.0×10^{-7}	—	1.0×10^{-5}	—	—
number of plate segments N	50	—	—	20	50
sampling frequency f_{e} [kHz]	48	—	—	—	—

sition path was divided into three domains:

1. From a piano to a struck-string tone,
2. from a struck-string to a plucked-string tone, and
3. from a plucked-string to a guitar tone.

First, the gradual change in timbre from a piano tone to a struck-string tone was created by the following procedure.

1. The piano tone was simulated using three slightly detuned strings. The difference between strings in cents was gradually changed to 0. This decreases beats.
2. The damping coefficient of plate b_1 was gradually varied to the larger value. This avoids creating a sharp resonance and an undesired change in the time envelope.

Next, the transition from a struck-string sound to a plucked-string sound was synthesized by:

1. interpolating excitation condition, and
2. interpolating damping coefficients of the string and plate b_3 linearly on a log scale, while keeping the damping coefficient of plate b_1 constant.

For the third domain, the lowest mode frequency of the resonator was made to approach the Helmholtz resonance of the guitar body by adjusting density and damping coefficient. A smaller number of plate segments N was used to satisfy the stability condition.

The numbers of samples selected from the three domains are the following. Piano to struck-string uses 4 tones. Plucked-string to guitar uses only 2 tones, because there is very little difference between a guitar tone and a plucked-string tone. For intermediate tones between struck and plucked, 6 sounds are synthesized, at physically equal intervals, *i.e.* 0%, 20, 40, 60, 80, and 100%.

Using the above procedure, 10 sounds were synthesized and their timbres were gradually changed from piano to guitar (referred to as Series A). Each sound is identified by indices from 1 to 10. The target fundamental frequency was set to 329.63 Hz (middle E), and the duration was 2.0 s. The parameters used for synthesis are summarized in Table 1. Waveforms and FFT spectra of the synthesized sounds are shown in Fig. 5.

3.4. Subjective Evaluation

Subjective evaluation is performed using the synthesized tones. The last 10 ms of the stimuli was linearly tapered in the case that pulse-noise might be heard. The

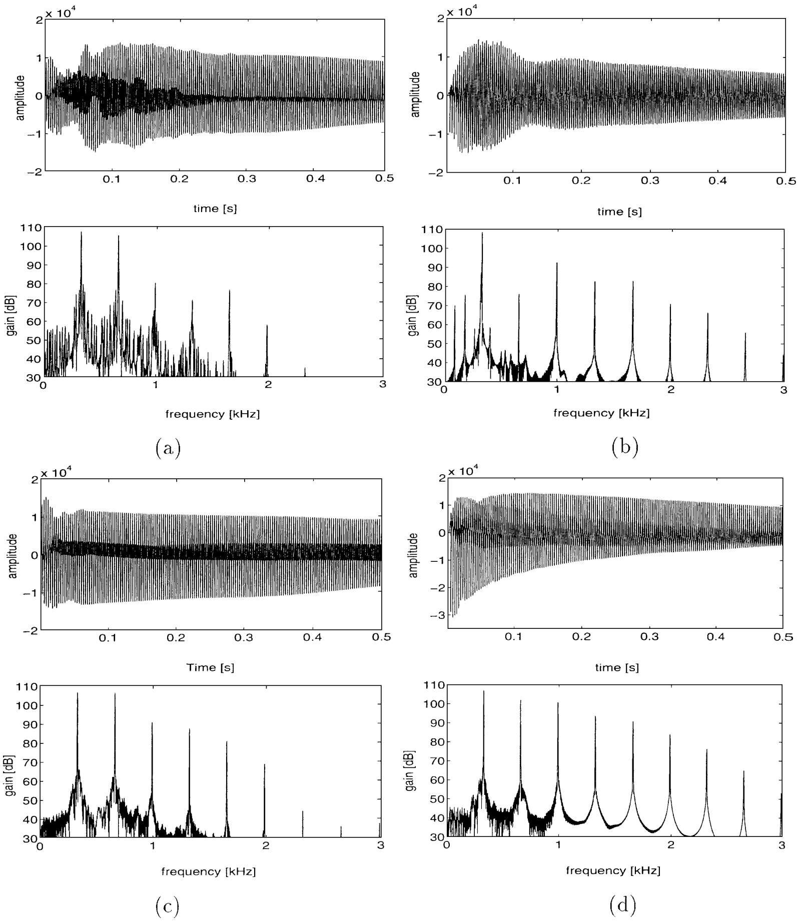


Fig. 5 Waveforms and FFT spectra of the synthesized sounds are shown. The length for analysis is 500 ms. (a) A piano tone. (b) A guitar tone. (c) A struck-string tone. (d) A plucked-string tone.

overall power of the stimuli was equalized. Ten subjects, from 18 to 27 years of ages, were employed for this experiment. In each trial, two tones randomly selected out of 10 synthesized tones were presented, preceded by 0.5-second-long white noise, and a decision interval of 2.5 s was given. Sounds were recorded on DAT, and subjects listen to the stimuli through headphones (STAX-Λ Pro). Subjects judged the timbral similarity of the pair of tones on a seven-point scale. The value 0 corresponds to the "same," and 6 corresponds to "totally different." There were 4 trials for each pair including both orders of the tones, and 180 trials in total.

For each pair, a mean score of the judgement across subjects and repetitions is calculated, and is regarded as a subjective distance. A multidimensional scaling (MDS) technique was adapted to the subject distance data. A two-dimensional solution modeled the responses with a stress of 12.2%. Figure 6 displays the solution.

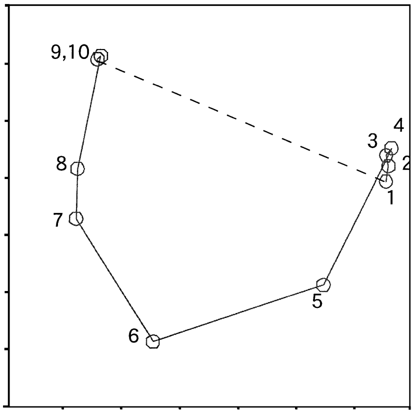


Fig. 6 Perceptual timbre space of the 10 synthesized tones (Series A). Two-dimensional space is generated by MDS (Kruskal's stress = 0.122). Index 1 means piano tone and 10 means guitar tone.

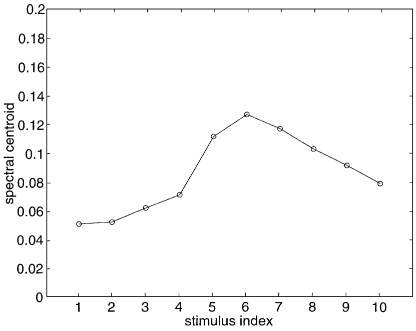


Fig. 7 The spectral centroid values versus the stimulus index is shown. When moving from piano to guitar, the centroid value once increases, and decreases.

The plot shows degeneration for the indices 1–4 and 9–10. The path makes a curve, and the index 6 is not located in between the two edges, although it has almost the same distances from the edges. This result does not satisfy intermediateness. This is because the subjective distances show saturation.

In order to interpret the axes in timbre space, synthesized signals are analyzed in terms of time and spectral characteristics. Here, the spectral centroid and the time envelope are calculated.

3.5. Signal Analyses

3.5.1. Spectral centroid

In the previous studies on timbral perception, the relationship between one axis in sound timbre space and the spectral energy distribution of a stimulus was often pointed out (*e.g.* [16, 18]). As a representative of the spectral distribution, the centroid frequency or the spectral centroid is calculated, and the relationship between the centroid and timbral perception is investigated.

The spectral centroid is calculated by the following procedure.

1. For a windowed (Hanning) signal with $2N$ length, FFT spectrum is calculated.
2. For the FFT spectrum $X(k),\ k = 0,\ \ldots,\ N-1$, the mean power for each subband $A(j)$ is calculated by

$$A(j) = \frac{1}{L}\sum_{i=0}^{L-1} |X(i+jL)|^2, \quad j = 0,\ \ldots,\ M-1,$$

   where M is the number of subbands and L is the number of samples included in each band.
3. Convert to log power (so that its minimum is zero), and normalize by the total power, and derive the relative power $B(j)$.

$$B(j) = \frac{10\log(A(j)+1)}{\sum_{j=0}^{M-1} 10\log(A(j)+1)}, \quad j = 0,\ \ldots,\ M-1$$

Spectral centroid C is calculated by

$$C = \sum_{j=0}^{M-1} B(j)\frac{2j+1}{2M}.$$

The centroid value is normalized by half the sampling frequency, so when the number of subbands is 1, $C = 0.5$. The main effect of using the subbands is smoothing, *i.e.*, detailed variations are neglected. The signal is truncated from the start with length $2N = 32,768$ samples ($\simeq 0.68$ s), and used for analysis.

The spectral centroids versus the stimulus index is shown in Fig. 7. The number of subbands = 24, 48, and 96 is investigated, and there was little difference between the trend of the centroids. Therefore, 24 (bandwidth = 1 kHz) is used. From stimuli 4 to 6, the centroid increases rapidly, and from 7 to 10, it decreases linearly. When moving from piano to guitar, the centroid value once increases, and decreases again.

In Fig. 6, the plot moves monotonically from right to left in terms of the axis parallel to the straight line which connects the two edges. For another axis which perpendicular to the first one, the plot of the tone first moves downward and reaches the bottom at the index 6, then moves upward. From Figs. 6 and 7, it is suggested that this second axis has a relationship with the spectral centroid of the stimuli.

3.5.2. Time envelope

Time envelopes of the synthesized sounds are calculated and compared. Here, segmental power is calculated every 20 ms. Tones of both end-points have different characteristics, and intermediate tones show almost the same curves. Synthesized piano tone exhibits amplitude modulation, which is due to the slight detuning of the strings. Simulated guitar tone also has envelope variations because of the resonance between the string and the body. However, no correlation was found between time envelopes and the plot in Fig. 6.

3.6. Consideration of the Spectral Feature

From the experimental results obtained above, it was suggested that the spectral centroid was related to one axis in the perceptual space. In this section, a modified synthesis algorithm is proposed using the spectral centroid value. The basic idea is that the timbres of the sounds may be changed linearly by keeping their centroid values changing linearly.

As an adjusting parameter of the centroid, frequency damping coefficient of a string b_3 is used. This parameter has a relationship with damping for high-frequency band in spectral envelopes, and the value is in proportion to damping, *i.e.*, when b_3 value is high, the high-frequency band is also highly damped. Furthermore, when the high-frequency band of the spectrum is damped more, the spectral centroid becomes smaller. Therefore, the algorithm for calculating a series of interpolated sounds which uses the spectral centroid can be considered in the following.

1. The spectral centroids for the both endpoints' sounds are calculated.
2. For each interpolation rate, the initial b_3 value of the string is calculated by linear interpolation on a log scale. The plate's b_3 value is also calculated by linear interpolation on a log scale.
3. The target value for the spectral centroid is calculated by interpolating the endpoints' centroids linearly.
4. The synthesized tone is created, and the centroid is calculated. If the difference between this value and the target value is smaller than the threshold, calculation is stopped. If not, the damping coefficient b_3 of the string is modified, and go to Step 4 again.
5. The process above is done for all interpolation rates.

Using this method, two series of tones are synthesized. As a difference threshold for the centroid, 0.005 is used, and increment/decrement step size for the b_3 value is set to 2.0×10^{-10}, around 5–10% of the b_3 values used. The first series (referred to as Series B) uses the same endpoints as used for Series A (as described in section 3.3). The second series (referred to as Series C) does not use the simulated guitar tone, and uses another parameter set for the plucked-string endpoint, which is referred to as "plucked 2" in Table 1. This tone has higher energy in the high-frequency band than the one used in Series B. The spectral centroid of this sound is expected to have a much higher value in comparison with that of the piano tone. For Series C, the centroids of some samples did not reach their target values even when $b_3 = 0.0$. So, the struck-string point (index 4) was also used as an endpoint and two straight lines are drawn between three endpoints for the target centroid trend.

Using the synthesized tones, subjective tests are performed, and the subjective distances are measured. Experimental procedure is the same as in section 3.4.

Two-dimensional solutions of the MDS calculated from the similarity data for Series 1 and 2 are shown in Figs. 8 and 9, respectively. Each solution modeled the response with a stress of 7.1 and 4.5%, respectively.

In Fig. 8, the plot shows degeneration for the indices 1–3. The path again makes a curve, and all stimuli are located near the straight line which connects both edges, especially when excluding the stimulus 9. Therefore, the intermediateness is relatively confirmed.

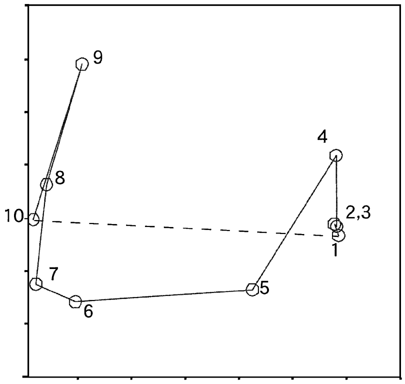


Fig. 8 Perceptual timbre space of the 10 synthesized tones (Series B). Two-dimensional space is generated by MDS (Kruskal's stress = 0.071). Index 1 means piano tone and 10 means guitar tone.

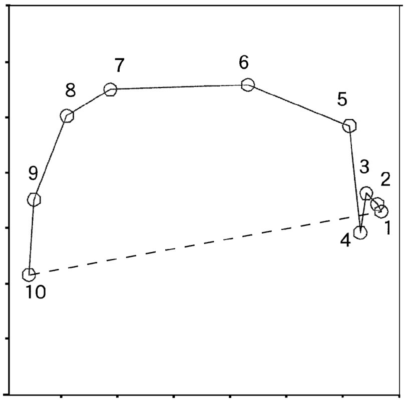


Fig. 9 Perceptual timbre space of the 10 synthesized tones (Series C). Two-dimensional space is generated by MDS (Kruskal's stress = 0.045). Index 1 means piano tone and 10 means plucked-string tone.

By comparing Figs. 6 with 8, the effect of varying damping coefficient b_3 on the plot is clear. That is, when a different b_3 value is used, the plot moves upward. The stimulus 9 moved too much to locate near the stimulus 10 by considering the centroid. The reason for this is that the simulated guitar tone is heard differently from the other plucked-string sounds because of the resonance, and that the similarity of the centroid values does not necessarily indicate increasing perceptual similarity.

Figure 9 shows degeneration for 1–4. The path looks smooth, and it is shown that smooth interpolation is completed. The plot of the samples is more equally spaced than that in Figs. 6 and 8.

4. DISCUSSION

By considering the spectral centroid, the tones approach the straight line which connects both endpoints.

That is, the intermediateness is improved. However, the distances between the adjacent tones have variance, as depicted in Fig. 8. This depends on how to determine the interpolation rates for intermediate sounds. In the present situation, perceptual feedback is needed on this point. When spectral centroid values of the endpoints are fairly different (Series C), smoother interpolation is achieved. In this case, contribution of the centroid to the timbral similarity judgement is considered to increase.

Next, two criteria are discussed. First, continuity criterion is considered. When a transition is continuous, it seems that subjective distances between the adjacent tones are closely located, and that they are equally spaced. So, the mean and the standard deviation values of the distances between adjacent pairs are regarded as the continuity measures.

$$C_\mathrm{m} = \frac{1}{N-1}\sum_{i=1}^{N-1} d_{i,i+1},$$

$$C_\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1} d_{i,i+1}^2 - C_\mathrm{m}^2},$$

where $d_{i,i+1}$ denotes the distance between stimuli i and $i+1$, N is the number of stimuli. When all the tones are equally spaced, $C_\sigma = 0$. Furthermore, when they are located in line, $C_\mathrm{m} = 1/(N-1)$.

For intermediateness, the mean distance between the interpolated tones and the straight line connecting both endpoints is calculated.

$$I = \frac{1}{N-2}\sum_{i=2}^{N-1} \hat{d_i},$$

where $\hat{d_i}$ denotes the distance between the interpolated tone i and the straight line connecting both endpoints. When all the tones are located in line, $I = 0$.

All values are normalized by the distance between the two edges, and shown in Table 2. When comparing Series A with B, all criteria decrease in value by considering the centroid. Series C has the most continuity of the three series, and its values are less than half the values for Series A. In terms of intermediateness, on the other hand, Series B has the smallest value. Although these criteria seem to be acceptable, they depend on stimuli used, and the number of stimuli, *etc*. Much investigation will be needed for various tones.

5. SUMMARY AND CONCLUSIONS

Our goal is to develop sound synthesis technology that can be used to synthesize arbitrary sound timbre, including musical instrument sounds, natural sounds, and their interpolation/extrapolation on demand. Such a technology will enrich expression, and make a breakthrough in music and contents creation. For this purpose, we investigated sound interpolation based on physical modeling. An interpolation algorithm using a sound-synthesis model composed of a one-dimensional string and excitation by striking and plucking was proposed. Global characteristics of a piano and a guitar were simulated, and the interpolation between piano and guitar tones was investigated. The strategy for parameter control was proposed, and subjective tests were performed to evaluate the algorithm. A multidimensional scaling technique was applied to evaluate our algorithm, and perceptual characteristics were discussed. One axis of the timbre space was interpreted as spectral energy distribution, so the spectral centroid was used as a reference to adjust parameters for synthesized tones. The results showed that both considering continuity of parameters and the centroids, smooth interpolation was achieved. The result of this paper suggests the possibility of developing a morphing system by using a physical model.

Much future work remains to be done. First, the relationship between the interpolation rate α for physical parameters and perceptual scale is not known. In this report, linear interpolation is basically used, but the result suggests the need for the consideration of some nonlinear functions. Other future work includes extending the timbral range. One way to do this is to build models to treat bowing, wind instruments, percussion instruments, *etc*.

After the parameter estimation problem from sound signals is solved, comparison with an algorithm using a sinusoidal model will be needed to investigate various aspects of morphing, including continuity and intermediateness. Merits and demerits of both algorithms should be clarified.

This is a preliminary report for sound morphing by physical modeling. In this paper, the psychoacoustical research aspect was rather prominent, and the need for verification of the timbre of synthesized tones from a perceptual point of view was pointed out. Once the performance of this algorithm is evaluated in terms of interpolation between two natural sounds, a system can be made available to users for extrapolating and freely modifying sounds for their musical creations.

Table 2 Criteria for continuity and intermediateness.

Criteria	Series A	Series B	Series C
		(centroid is considered)	
Continuity C_m	0.41	0.32	0.17
Continuity C_σ	0.25	0.19	0.12
Intermediateness I	0.082	0.057	0.060

ACKNOWLEDGMENTS

The authors are grateful to Dr. Ken'ichiro Ishii, the director of the NTT Communication Science Laboratories, and Dr. Norihiro Hagita, the executive manager of the Media Information Laboratory, for their continuing support.

REFERENCES

[1] G. Borin, G. De Poli and A. Sarti, "Algorithms and structures for synthesis using physical models", *Comput. Music J.* **16**, 30–42 (1992).

[2] J. O. Smith, "Physical modeling synthesis update", *Comput. Music J.* **20**, 44–56 (1996).

[3] S. Van Duyne and J. O. Smith, "Physical modeling with a 2-D digital waveguide mesh", *Proc Int. Computer Music Conf.*, 40–47 (1993).

[4] O. Calvet, R. Laurens and J. M. Adrien, "Modal synthesis: Compilation of mechanical sub-structures and acoustical subsystems", *Proc. Int. Computer Music Conf.*, 57–59. (1990).

[5] J. M. Adrien, "The missing link: modal synthesis", in *Representations of Musical Signals*, G. De Poli, A. Piccialli and C. Roads, Eds. (MIT Press, Cambridge, Mass., 1991), Chap. 8, pp. 269–297.

[6] J. O. Smith, "Physical modeling using digital waveguides", *Comput. Music J.* **16**, 74–91 (1992).

[7] P. Cook, "TBone: An interactive waveguide brass instrument synthesis workbench for the NeXT Machine", *Proc. Int. Computer Music Conf.*, 297–299 (1991).

[8] E. Tellman, L. Haken and B. Holloway, "Timbre morphing of sounds with unequal numbers of features", *J. Audio Eng. Soc.* **43**, 678–689 (1995).

[9] N. Osaka, "Timbre interpolation of sounds using a sinusoidal model", *Proc. Int. Computer Music Conf.*, 408–411 (1995).

[10] K. Fitz, L. Haken and B. Holloway,"Lemur — A tool for timbre manipulation", *Proc. Int. Computer Music Conf.*, 158–161 (1995).

[11] N. Bernardini and A. Vidolin, "Real-time sound hybridization", *Proc. Int. Computer Music Conf.*, 179–182 (1995).

[12] M. Abe, "Speech morphing by gradually changing fundamental frequency and spectra", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 259–260 (1995) (in Japanese).

[13] M. Slaney, M. Covell and B. Lassiter, "Automatic audio morphing", *Proc. ICASSP* 96, Vol. 2, 1001–1004 (1996).

[14] H. Banno, K. Takeda, K. Shikano and F. Itakura, "Speech morphing by independent interpolation of spectral envelope and source excitation", *Trans. IEICE* **J81-A**, 261–268 (1998) (in Japanese).

[15] D. A. Jaffe, "Ten criteria for evaluating synthesis techniques", *Comput. Music J.* **19**, 76–87 (1995).

[16] J. M. Grey, "Multidimensional perceptual scaling of musical timbres", *J. Acoust. Soc. Am.* **61**, 1270–1277 (1977).

[17] J. R. Miller and E. C. Carterette, "Perceptual space for musical structures", *J. Acoust. Soc. Am.* **58**, 711–720 (1975).

[18] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres", *J. Acoust. Soc. Am.* **63**, 1493–1500 (1978).

[19] L. Hiller and P. Ruiz, "Synthesizing musical sounds by solving the wave equation for vibrating objects: Part I", *J. Audio Eng. Soc.* **19**, 462–470 (1971).

[20] L. Hiller and P. Ruiz, "Synthesizing musical sounds by solving the wave equation for vibrating objects: Part II", *J. Audio Eng. Soc.* **19**, 542–551 (1971).

[21] A. Chaigne and A. Askenfelt, "Numerical simulations of piano strings. I. A physical model for a struck string using finite difference methods", *J. Acoust. Soc. Am.* **95**, 1112–1118 (1994).

[22] A. Chaigne and A. Askenfelt, "Numerical simulations of piano strings. II. Comparisons with measurements and systematic exploration of some hammer-string parameters", *J. Acoust. Soc. Am.* **95**, 1631–1640 (1994).

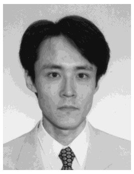
[23] A. Chaigne, "On the use of finite differences for musical synthesis. Application to plucked stringed instruments", *J'd Acoust.* **5**, 181–211 (1992).

[24] M. Pavlidou and B. Richardson, "The string-finger interaction on the classical guitar", *Int. Sym. Musical Acoustics*, 558–564 (1995).

[25] M. Kurz and B. Feiten, "Physical modelling of a stiff string by numerical integration", *Proc. Int. Computer Music Conf.*, 361–364 (1996).

[26] G. Cuzzucoli and V. Lombardo, "Physical model of the plucking process in the classical guitar", *Proc. Int. Computer Music Conf.*, 172–179 (1997).

[27] I. Nakamura, "Simulation of the sound production mechanism — Acoustical research on the piano Part 2 —", *J. Acoust. Soc. Jpn.* (*J*) **37**, 65–75 (1981) (in Japanese).



Takafumi Hikichi was born in Nagoya, in 1970. He received his Bachelor and Master of Electrical Engineering degrees from Nagoya University in 1993 and 1995, respectively. In 1995, he joined the Basic Research Laboratories of NTT, Atsugi. He is now working at the Communication Science Laboratories, NTT. His research interests include sound synthesis, sound timbre control (morphing), physical modeling of sound sources, and digital signal processing. He was awarded the Awaya Prize from the ASJ in 2000. He is a member of the IEICE and ASJ.

Naotoshi Osaka was born in Nagano prefecture, Japan in 1953. He received M.S. degrees in electrical engineering from Waseda University in 1978. His main research interests include telephone transmission performance and speech dialogue. He received a Dr. degrees for an objective model for telephone transmission performance. He is currently studying timbre synthesis for both sounds and speech. He is presently leading a computer music research group at NTT Communication Science Laboratories in Atsugi, Kanagawa, Japan. He is a senior member of IEEE and is also a member of the IEICE, IPSJ and ASJ.