ANALYSIS-AND-MANIPULATION APPROACH TO PITCH AND DURATION OF MUSICAL INSTRUMENT SOUNDS WITHOUT DISTORTING TIMBRAL CHARACTERISTICS

Takehiro Abe, Katsutoshi Itoyama, Kazuyoshi Yoshii,
Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno

Department of Intelligence Science and Technology,
Kyoto University, Japan
abe@kuis.kyoto-u.ac.jp

ABSTRACT

This paper presents an analysis-manipulation method that can generate musical instrument sounds with arbitrary pitches and durations from the sound of a given musical instrument (called *seed*) without distorting its timbral characteristics. Based on psychoacoustical knowledge of the auditory effects of timbres, we defined timbral features based on the spectrogram of the sound of a musical instrument as (i) the relative amplitudes of the harmonic peaks, (ii) the distribution of the inharmonic component, and (iii) temporal envelopes. First, to analyze the timbral features of a *seed*, it was separated into harmonic and inharmonic components using Itoyama's integrated model. For pitch manipulation, we took into account the pitch-dependency of features (i) and (ii). We predicted the values of each feature by using a cubic polynomial that approximated the distribution of these features over pitches. To manipulate duration, we focused on preserving feature (iii) in the attack and decay duration of a *seed*. Therefore, only steady durations were expanded or shrunk. In addition, we propose a method for reproducing the properties of vibrato. Experimental results demonstrated the quality of the synthesized sounds produced using our method. The spectral and MFCC distances between the synthesized sounds and actual sounds of 32 instruments were reduced by 64.70% and 32.31%, respectively.

1. INTRODUCTION

A traditional equalizer enables users to change the spectral characteristics of acoustic signals. New equalizers that recently been developed for musical sounds can manipulate the volume and replace the timbre of individual musical instrument part [1, 2, 3]. These techniques are called as musical instrument equalizers. While the equalizer provided in a typical audio player changes musical sounds by manipulating the frequency range, a musical instrument equalizer changes the sounds by manipulating musical instrument parts, which enhances the listening experience. Yoshii's musical instrument equalizer (called Drumix [2]) can adjust the volume and replace the timbre of only percussive instruments (snare and bass drums). However, Itoyama's musical instrument equalizer can adjust the volume of all musical instruments [3]. Unfortunately, the latter is so far limited to volume, and cannot replace the timbre of each musical instrument part.

Our ultimate goal is to develop a musical instrument equalizer that can replace arbitrary musical instrument parts with users' favorite timbres. For example, the equalizer we envisage would enable the musical instruments typically used to play rock music (electric guitar, electric bass, keyboard, etc.) to be replaced by instruments used to play classical music (violin, wood bass, piano, etc.) Users could thus enjoy a classical remix of the music. In addition, by extracting the guitar sounds from a tune played by a favorite guitarist (Eric Clapton, Yngwie J. Malmsteen, etc.) and replacing the guitar part of another tune with the extracted sounds, users could listen to their favorite guitarist playing various phrases virtually.

To achieve our goal, we need to tackle the following problems:

(1) **separating the monophonic sounds of a target musical instrument from a polyphonic audio signal** to extract the musical instrument sounds that users want to replace; and

(2) **synthesizing new sounds that have arbitrary pitch and duration** from the separated sounds to play arbitrary phrases.

Many researchers including Itoyama have studied the former problem and have reported their results for sound-source separation [4, 5, 6]. However, there have been few studies of the application of separated sounds. We have therefore focused on the latter problem, that is, analysis-manipulation of musical instrument sounds from separated sounds.

2. MANIPULATION OF PITCH AND DURATION WITH CONSIDERATION OF TIMBRAL CHARACTERISTICS

Our aim is, given some actual sounds of an individual musical instrument (called *seed*), to synthesize the sound of that instrument with arbitrary pitch and duration based on the original sounds. A key point of this synthesis is to avoid distorting the timbral characteristics[1]. For example, if we synthesize a D sound based on the C sound of a musical instrument, users should feel as if the D sound was generated by the same individual instrument, not from a different one.

To synthesize a musical instrument sound without distorting its timbral characteristics, we need to define the timbral features mathematically and analyze the characteristics of timbres. Studies in acoustic psychology have found that auditory differences between timbres tend to be caused by (i) spectral energy distribution, (ii) synchronicity in the transients of higher harmonics, and (iii) low-amplitude, high-frequency energy in the attack segment [7]. We consider the these factors correspond to the following three features:

(i) **the relative amplitudes of the harmonic peaks**,
(ii) **the inharmonic component**, and
(iii) **temporal envelopes**.

We took the analysis-manipulation approach shown in Figure 1. Features (i) and (iii) are related to the harmonic component, and Feature (ii) is related to the inharmonic component. First,

[1] In this paper, we define the distortion of timbral characteristics as the difference between the timbre of the synthesized sound and the timbre of the sound obtained by playing the real musical instrument

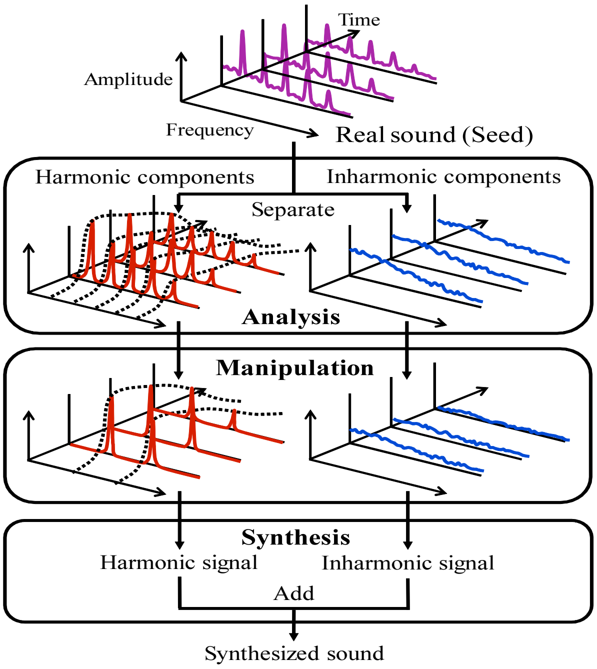


Figure 1: *Overview of our method.*

we analyzed each feature by separating the harmonic and inharmonic components of a *seed*. After the analysis, the pitch and duration were manipulated without distorting the timbral characteristics. Here, we note that it is not proper to manipulate only the pitch and duration without changing the timbral features. Finally, we synthesized the harmonic and inharmonic signals separately in adding the synthesized signals.

2.1. Analysis of musical instrument sounds

To analyze timbral features, it is necessary to deal with harmonic and inharmonic components explicitly and to define the various mathematical features. To solve this problem, we used the integrated model of harmonic and inharmonic structures presented by Itoyama. We attempted to express musical instrument sounds using an integrated model, i.e., we adapted a mixed model weighted by w_H and w_I , which is a combination of a parametric model corresponding to the harmonic component $M_H(f,r)$ and a nonparametric model corresponding to an inharmonic component $M_I(f,r)$ to the spectrogram $M(f,r)$ of a *seed* as follows:

$$M(f,r) = w_H M_H(f,r) + w_I M_I(f,r), \tag{1}$$

where f and r represent the frequency and time, respectively. The following constraint applies: in $\sum_{f,r} M_I(f,r) = 1$, the weight w_I represents the energy of an inharmonic component, and $w_I M_I(f,r)$ is the spectrogram of an inharmonic component. $M_H(f,r)$ is expressed as a weighted mixture model, which is parametric to nth peaks as follows:

$$M_H(f,r) = \sum_n F_n(f,r) E_n(r), \tag{2}$$

where $\sum_n F_n(f,r)$ and $E_n(r)$ respectively correspond to the spectral and temporal envelopes of the harmonic component, as shown in Figures 2, 3.

$\sum_n F_n(f,r)$ is expressed as a Gaussian Mixture Model as follows:

$$F_n(f,r) = v_n \mathcal{N}(f - n\mu(r), \sigma^2), \tag{3}$$

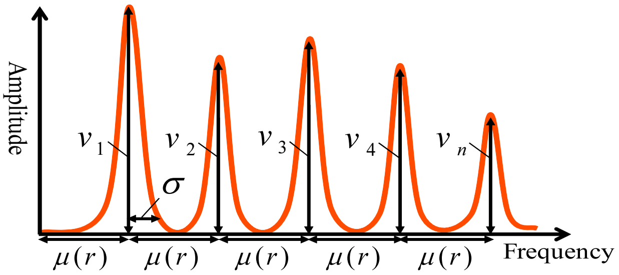


Figure 2: *Spectral envelope.*

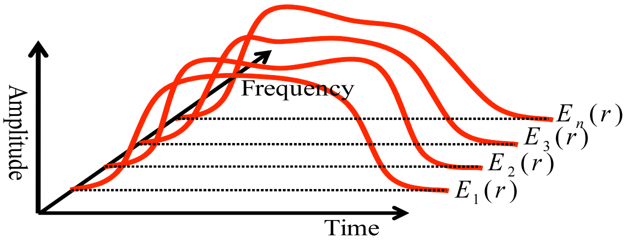


Figure 3: *Temporal envelopes.*

where $\mathcal{N}(x, y^2)$ represents the Gaussian, the mean and variance of which are x and y^2, respectively. σ therefore represents the extent on the frequency domain and $\mu(r)$ is the pitch trajectory on the time domain. v_n is the relative weight, where $\sum_n v_n = 1$.

$E_n(r)$is the nonparametric function, where $\sum_r E_n(r) = 1$. While Itoyama constructed $E_n(r)$by using a parametric function such as $F_n(f,r)$, we use the nonparametric function to express $E_n(r)$for a more detailed analysis. In the integrated model, the timbral features (i), (ii) and (iii) correspond to v_n, w_I, $M_I(f,r)$, and $E_n(r)$, respectively. We describe the analysis of these features in section 3.

2.2. Pitch manipulation

Pitch manipulation was achieved by multiplying the pitch trajectory $\mu(r)$ by a desired ratio. However, the values of the timbral features should not be held when manipulating pitch because timbres are pitch-dependent [8]; thus, the larger the ratio of pitch manipulation, the larger the distortion of timbral features. When we shift from $\mu(r)$ to $\mu'(r)$, we must also shift from v_n to v'_n properly, as shown in Figure 4.

In solving this problem, we noted the musical instrument identification method proposed by Kitahara *et al*, which considered the pitch-dependency of timbres [9]. They reported that the performance of the identification method was improved by learning the distribution of the acoustic features after removing the pitch-dependency of timbres by approximating the feature distribution as a cubic polynomial. In our study, except for feature (iii), which depends on articulation style rather than on pitch, we approximated the distribution of features (i) and (ii) over pitches as a cubic polynomial (called **pitch-dependent feature function**). Specifically, we dealt with the following parameters:

(1) the relative amplitudes of the harmonic peaks $v_n$and
(2) the ratio of harmonic energy to inharmonic energy w_H/w_I.

Given that some *seed*s have various pitches, we analyzed their timbral features, so that we could obtain the pitch-dependent feature function using the least squares method. By using the obtained pitch-dependent feature function, the timbral features were determined for the desired pitch. For example, the relative amplitudes

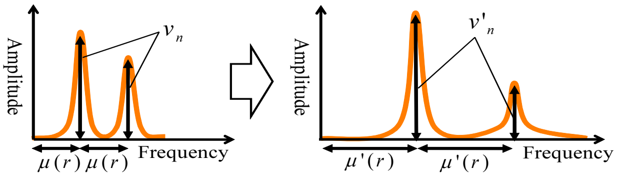


Figure 4: *Manipulation of spectral envelope.*

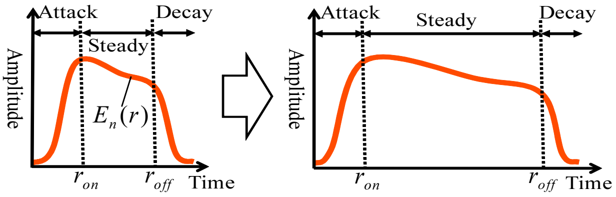


Figure 6: *Manipulation of temporal envelope.*

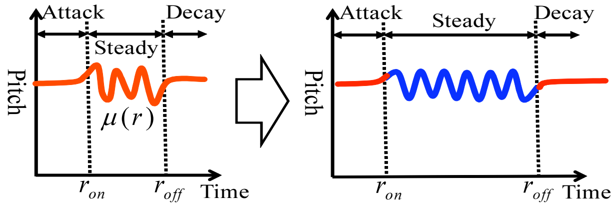


Figure 7: *Generation of pitch trajectory.*

of the 1st, 4th, and 10th peaks, and the ratio of harmonic energy to inharmonic energy for trumpet sounds, are shown in Figure 5.

2.3. Duration manipulation

Duration should be manipulated by expanding or shrinking the whole temporal envelope $E_n(r)$ by the desired ratio of duration because the excitation in the attack and decay segments and the properties of the pitch trajectory are similar in the same individual musical instrument regardless of the duration; therefore, the larger the ratio of duration manipulation, the larger the amount of distortion. In particular, in the attack and decay segments of a musical instrument sound, the level of energy changes the perception of loudness, which gives the impression of timbres. Similarly, the pitch trajectory affects auditory impressions, especially for musical instruments that are offen played using vibrato articulation (electric guitar, violin, etc.).

To solve this problem, we propose a method that preserves temporal envelopes in attack and decay segments, and a method that reproduces the properties of the pitch trajectory. In feature (iii), we define the end of the sharp emission of energy as onset R_{on}, and the start of the sharp decline in energy as offset R_{off}. When manipulating duration, only the temporal envelopes between onset and offset were expanded and shrunk, as shown in Figure 6. Moreover, we expressed the pitch trajectory between onset and offset by using a sinusoidal model, as shown in Figure 7, which reproduces the pitch trajectory that has the same spectral characteristic. The pitch trajectories before onset and after offset are the same as for *seed*.

2.4. Synthesis of musical instrument sounds

To synthesize a harmonic signal $s_H(t)$, we used a sinusoidal model, using the features (i) and (iii). To synthesize an inharmonic signal $s_I(t)$, we used overlap-add synthesis, using the feature (ii). Finally, the output sound was synthesized by adding the synthesized harmonic sound to the synthesized inharmonic sound.

3. IMPLEMENTATION OF OUR METHOD

In this section, we explain the specifics of the method described in section 2.

3.1. Analysis of musical instrument sounds

Here, the problem is the estimation of the unknown parameters of the integrated model: w_H, w_I, $F_n(f,r)$, $E_n(r)$, v_n, $\mu(r)$, σ, and $M_I(f,r)$. Itoyama proposed a method that renews these parameters by reducing the Kullback-Leibler Divergence (KLD) iteratively. This iterative calculation can be regarded as an Expectation-and-Maximization (EM) algorithm, which estimates these parameters efficiently. The unknown parameters were estimated by minimizing the following cost function:

$$\begin{aligned} J = {} & \\ & \sum_n \iint \Bigg(G_n^H(f,r) \log \frac{G_n^H(f,r)}{w_H E_n(r) F_n(f,r)} \\ & \qquad - G_n^H(f,r) + w_H E_n(r) F_n(f,r) \Bigg) df\,dr \\ & + \iint \Bigg(G^I(f,r) \log \frac{G^I(f,r)}{w_I M_I(f,r)} \\ & \qquad - G^I(f,r) + w_I M_I(f,r) \Bigg) df\,dr \\ & + \lambda_v \Bigg(\sum_n v_n - 1 \Bigg) + \sum_n \Bigg(\lambda_{E_n} \Bigg(\int E_n(r) dr - 1 \Bigg) \Bigg) \\ & + \beta_I \iint \Bigg(\bar{M}_I(f,r) \log \frac{\bar{M}_I(f,r)}{M_I(f,r)} \\ & \qquad - \bar{M}_I(f,r) + M_I(f,r) \Bigg) df\,dr, \end{aligned} \tag{4}$$

where β_I is the constraint weight, $\bar{M}_I(f,r)$ is obtained by smoothing the inharmonic model $M_I(f,r)$ with a Gaussian filter on the frequency domain, and λ_v and λ_{E_n} are Lagrange multipliers. $G_n^H(f,r)$ and $G^I(f,r)$ are the divided harmonic and inharmonic components, respectively.

3.2. Pitch manipulation

Pitch manipulation was achieved by multiplying $\mu(r)$ by a real number α (to low pitch: $0 \leq \alpha < 1$, to high pitch: $1 < \alpha$) as follows:

$$\mu'(r) = \alpha \mu(r), \tag{5}$$

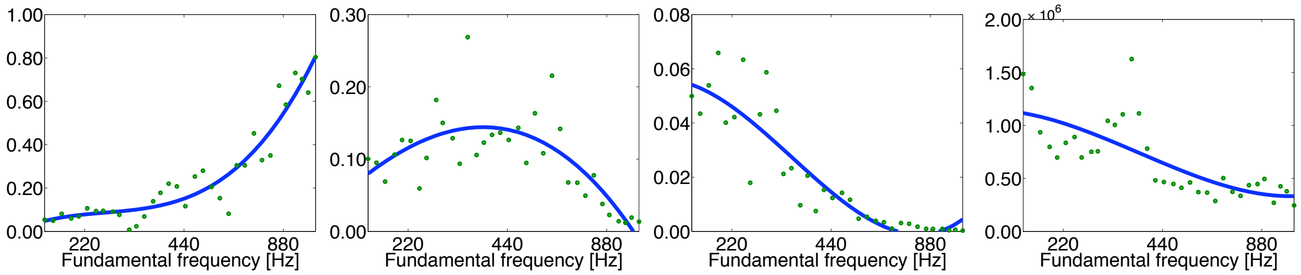
where $\mu'(r)$ is the desired pitch. For example, a musical instrument sound with a pitch that is one octave higher was synthesized by substituting two for α. The relative amplitudes of the harmonic peaks after pitch manipulation v'_n were obtained by calculating each relative amplitude of the harmonic peaks from pitch-dependent feature functions, which were normalized with the constraint $\sum_n v_n = 1$. The inharmonic energy w'_I was obtained by dividing the harmonic energy by the expected ratio w_H/w_I.

3.3. Duration manipulation

The duration was manipulated by manipulating the temporal envelopes $E_n(r)$ between onset and offset and generating the pitch trajectory $\mu(r)$.

3.3.1. Onset and offset detection

In our study, onset was defined as the moment at which the vibration energy of a musical instrument reached a sufficient level, and the variation in the energy was low. Offset was defined as the moment when this energy (with low variation) dropped suddenly.



(a) *Relative amplitude of 1st harmonic component.*

(b) *Relative amplitude of 4th harmonic component.*

(c) *Relative amplitude of 10th harmonic component.*

(d) *Ratio of harmonic energy to inharmonic energy.*

Figure 5: *Pitch-dependent feature functions for trumpet (circles and line are the analyzed timbral features and approximated pitch-dependent feature function in each figure).*

Specifically, we defined onset R_{on} and offset R_{off} as the start and end of r that satisfied the following condition, respectively:

$$\left|\frac{dE_n(r)}{dr}\right| \leq \epsilon, \ \ E_n(r) \geq Th, \tag{6}$$

where Th was the threshold for judging the vibration energy of the sound of musical instrument. While this detection method can be applied to wind and bowed string instruments, it cannot be applied to string instruments that are plucked or struck because the onset and offset occur at the same time in these instruments, so the temporal envelopes between onset and offset cannot be expanded or shrunk. When manipulating these string instruments, we regard the end of the temporal envelopes as the offset, and these are manipulated after the onset.

3.3.2. *Modeling of pitch trajectory*

To construct a model of the pitch trajectory $\mu(r)$, we propose a pitch trajectory model $M^{(\mu)}(r)$ based on a sinusoidal model, assuming that the variation in frequency and amplitude are stable, as follows:

$$M^{(\mu)}(r) = \sum_k A_k^{(\mu)} \exp[j\zeta_k r] + \mu_{ave}, \tag{7}$$

$$\mu_{ave} = \int \mu(r) dr / R, \tag{8}$$

where R is the time length of a musical instrument sound. The unknown parameters of this model are the amplitude $A_k^{(\mu)}$, frequency ζ_k and phase ψ that make up the pitch trajectory. These parameters were estimated by adapting this model to the pitch trajectory iteratively using the following algorithm:

Step 1: The signal, which is obtained by subtracting the average pitch trajectory μ_{ave} from the pitch trajectory $\mu(r)$, is transformed to the spectrum. The spectrum thus obtained is analyzed in the next step.

Step 2: $A_k^{(\mu)}$, ζ_k, and ψ of the largest peak are estimated by using the harmonic model of the integrated model (number of time frames: 1, number of peaks: 1). Simultaneously, the rest spectrum that is the result of the separation of the largest peak is estimated by using the inharmonic model of the integrated model.

step 3: The rest spectrum is regarded as the analyzed spectrum in step 2, which continues until the rest spectrum becomes small enough.

step 4: The estimated parameters of the pitch trajectory model are $A_k^{(\mu)}$, ζ_k, and ψ that are obtained using the above steps.

This algorithm is similar to the McAulay-Quatieri (MQ) algorithm [10] as a method of estimating the parameters of a sinusoidal model. In the MQ algorithm, the rest spectrum is obtained by subtracting the estimated peak in step 2 from the analyzed spectrum. However, we applied the integrated model to the subtraction by regarding the inharmonic component as the rest spectrum, so that the estimated peak was separated from the analyzed spectrum.

3.4. Synthesis of musical instrument sounds

The harmonic signal $s_H(t)$ and inharmonic signal $s_I(t)$ were synthesized from the harmonic and inharmonic models, respectively. Finally, the output sound $s(t)$ is synthesized by adding these signals as follows:

$$s(t) = s_H(t) + s_I(t) \tag{9}$$

3.4.1. *Synthesis of harmonic signal*

The synthesis of the harmonic signal $s_H(t)$ was achieved by using the sinusoidal model [10] as follows:

$$s_H(t) = \sum_n A_n(t) \exp[j\phi_n(t)] \tag{10}$$

$$\phi_n(t) = \phi_n(0) + \int_0^t \omega_n(\tau) d\tau \tag{11}$$

where $A_n(t)$, $\phi_n(t)$, and $\omega_n(t)$ are the amplitude, instantaneous phase, and instantaneous frequency of the nth sinusoid respectively. The instantaneous frequency was obtained from the pitch trajectory $\mu(r)$ by using the spline interpolation method. The amplitudes were calculated from the parameters of the harmonic model as follows:

$$A_n(t) = \frac{w_H E_n(r) v_n}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} w(\tau) d\tau \tag{12}$$

where $w(t)$ is the window used in analyzing the spectrogram of a *seed*.

3.4.2. *Synthesis of inharmonic signal*

The synthesis of the inharmonic signal $s_I(t)$ was achieved by using the overlap-add method [11]. This algorithm is commonly used to transform a spectrogram to a signal. Here, the spectrogram of the inharmonic component is $w_I M_I(f, r)$, and the phase was obtained directly from a *seed*.

4. EVALUATION

To evaluate our method for pitch manipulation, we carried out an experiment in which we compared the results obtained using our method and a baseline method. The baseline method was simply a version of our method with no consideration of the pitch-dependency of timbres.

Table 1: *Types and number of musical instrument sounds used in the experiment*

Instrument names	Piano (PF), Electric Piano (EP), Harpsichord (HC), Vibraphone (VI), Marimba (MB), Organ (OR), Accordion (AC), Harmonica (HM), Classic Guitar (HC), Ukulele (UK), Acoustic Guitar (AG), Mandolin (MD), Electric Guitar (EG), Electric Bass (EB), Violin (VN), Viola (VL), Cello (VC), Contrabass (CB), Harp (HP), Trumpet (TR), Trombone (TB), Tuba (TU), Soprano Sax (SS), Alto Sax (AS), Tenor Sax (TS), Baritone Sax (BS), Oboe (OB), Fagot (FG), Clarinet (CL), Piccolo (PC), Flute (FL), Recorder (RC)
Individuals	3 individuals.
Intensity	Forte only.
Articulation	Normal articulation style only.
Number of tones	PF: 264, EP: 206, HC: 178, VI: 104, MB: 145, OR: 178, AC: 141, HM: 101, HC: 111, UK: 71, AG: 111, MD: 123, EG: 111, EB: 88, VN: 138, VL: 126, VC: 134, CB: 111, HP: 241, TR: 103, TB: 96, TU: 90, SS: 99, AS: 99, TS: 98, BS: 98, OB: 96, FG: 120, CL: 120, PC: 98, FL: 111, RC: 75

4.1. Experimental conditions

To evaluate the quality of the synthesized musical instrument sounds, we calculated the distances between a synthesized sound and the sound of a real musical instrument using the following criteria:

1. Spectral distance
$$D_S \;=\; \sum_{f,r}(S_{syn} - S_{real})^2/R, and \qquad (13)$$
2. Mel-Frequency Cepstrum Coefficient (MFCC) distance
$$D_M \;=\; \sum_{d,r}(M_{syn} - M_{real})^2/R, \qquad (14)$$

where S_i and M_i are the spectrogram and MFCC respectively, and these indexes, *syn* and *real*, indicate the synthesized sound and the real sound: the smaller these distances, the more similar the synthesized sound to the real sound. The spectral distance is mainly a measure of the difference between each peak of the harmonic component because the frequency domain is on a linear-scale. The MFCC distance is commonly used as a criterion for quantitative auditory measurement. It can be used to evaluate the difference in both harmonic and inharmonic components. The energy of an inharmonic component is smaller than the energy of the peaks of a harmonic component because the frequency domain is on a log-scale. The number of MFCC dimensions was 12.

The actual sounds used for the experiment were extracted from the RWC Music Database, RWC-MDB-I-2001, developed by Goto *et al* [12]. In this database, the solo tones of musical instruments are recorded with each semitone, which are sampled by 44.1 kHz with 16 bits, monaurally. We selected three individuals instruments from 32 instruments, and extracted the sounds of the selected musical instruments played using forte and normal articulation[2]. Details of the experimental data are shown in Table 1.

The evaluation was carried out using 10-fold cross validation within each individual musical instrument to enable us to calculate the distances between a synthesized sound and the sound of a real musical instrument. First, we divided 10% and 90% of the solo tones of an individual instrument into learning data and evaluation data respectively. The learning data were used to learn pitch-dependent feature functions. These data were also regarded as *seed*, so that we synthesized the sounds using the same pitch as for the evaluation data. Finally, we calculated the distances between the synthesized sounds and real musical instrument sounds. For example, in the case of piano which has 88 keys, the number of pieces of learning data is 9 (or 8) and the number of pieces of evaluation data is 79 (or 80), so that there were 8 [cross] $\times$ 79 $\times$ $9 + 2$ [cross] $\times$ $80 \times 8 = 6,968$ trials. We carried out the above evaluation for all the data i.e., we conducted a total of 447,772 trials.

Here, to reflect the quality of a synthesized sound in relation to distances, we canceled the variation in a musical performance in both the temporal envelopes $E_n(r)$ and pitch trajectory $\mu(r)$. We extracted $E_n(r)$ and $\mu(r)$ from the evaluation data, and extracted other parameters from the learning data (i.e., *seed*), so that we synthesized sounds using the extracted parameters. In this experiment, we only evaluated pitch manipulation.

4.2. Results and discussion

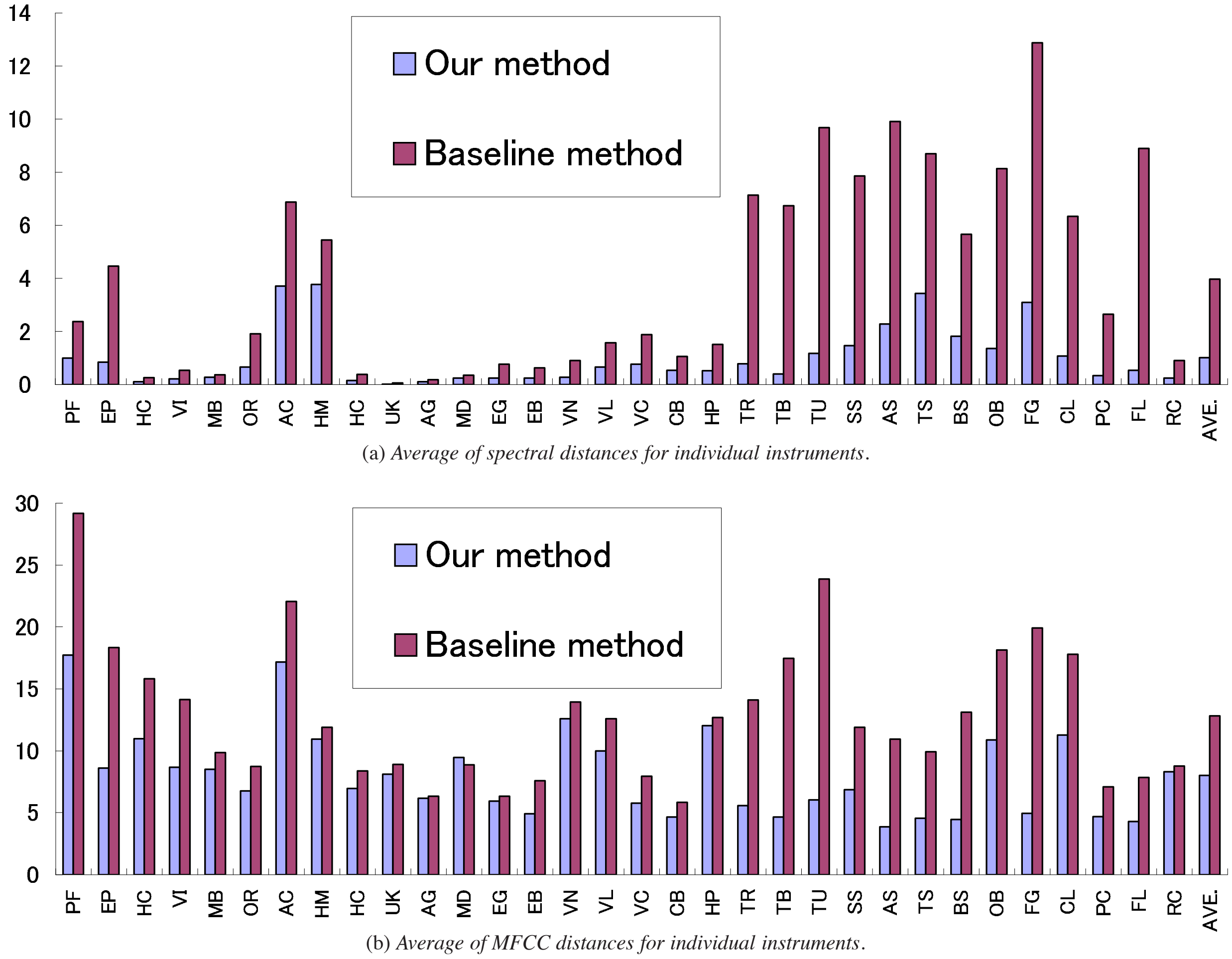
Figure 8 summarizes the spectral distance and MFCC difference for both methods respectively. The values were averaged for each musical instrument. Our method improved the spectral distances for all musical instruments and also improved the MFCC distances for all musical instruments except the mandolin. Our method reduced the average spectral difference and the average MFCC difference by 64.70% and 32.31%, respectively. The experimental results demonstrated the validity of our method, which uses pitch-dependent feature functions.

The distances for fagot (average reduction for the spectral distance: 76.02 %, and for the MFCC distance: 75.17 %) are shown in Figure 9 (a), (b) as an example of a much improved result. In the baseline method, both distances increased with an increase in the absolute value of manipulated halftones. However, in our method, both distances were stable in spite of an increase in the absolute value of manipulated halftones. When the value of manipulated halftones was small, the baseline method performed a little better than our method in terms of the MFCC distance because of an error in approximating the timbral features using pitch-dependent feature functions in our method.

The larger the improved value becomes, the stronger the pitch-dependency of the timbre. In addition to the distances for fagot, there were also good improvements in the distances for the piano and for brass instruments such as the trumpet, trombone, tuba, etc. We believe that the timbre of the piano has strong pitch-dependency because of the complex structure of this instrument. On the other hand, we consider that the strong pitch-dependencies of the trumpet, trombone, and tuba are due to the qualities of the materials used. In many musical instruments, except the above-mentioned instruments in which the MFCC distances were improved, we found that both distances tended to be stable in our method, in spite of increases in the absolute value of manipulated halftones.

The MFCC distances did not show any improvement for some

[2] Normal articulation is a common style of articulation in contrast to vibrato and staccato articulations. However, for the violin, because vibrato sounds in the RWC Music Database are registered as normal articulation, we selected sounds registered as non-vibrato



(a) *Average of spectral distances for individual instruments.*

(b) *Average of MFCC distances for individual instruments.*

Figure 8: *Differences in spectral distance and MFCC distance with consideration (our method) and no consideration (baseline method) of pitch-dependence. Spectral distances were normalized with distance of piano in our method.*

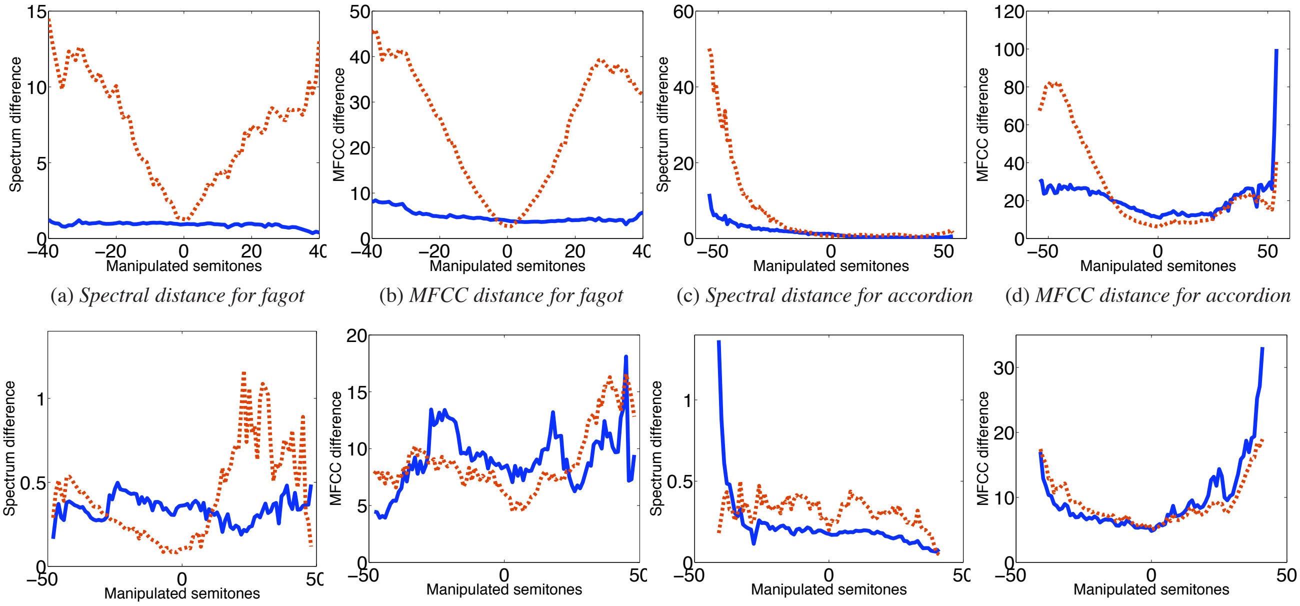
musical instruments. We discuss the possible reasons for this below.

(1) **Little pitch-dependency**
The distances for the accordion (average reduction for the spectral distance: 46.19 %, and for the MFCC distance: 22.08 %) are shown in Figure 9 (c), (d). When manipulating the pitch to low, both distances were improved by using our method. However, when manipulating the pitch to high, there was little improvement in both distances. This is because the accordion has little pitch-dependency at high pitch.

(2) **Complex pitch-dependency**
The distances for the marimba (average reduction for the spectral distance: 24.49 %, and for the MFCC distance: 13.99 %) are shown in Figure 9 (e), (f). There were large changes in these distances using both methods. This result may be due to the difficulty of learning the pitch-dependency of this musical instrument. The sound of the marimba includes percussive elements with an independent structure, which is similar to that of a piano. Approximating pitch-dependent feature functions as a cubic polynomial is not sufficient to represent the complex pitch-dependency of an instrument like the marimba. We could suggest increasing the polynomial number as a simple solution to the problem, but it is not possible to execute learning accurately by increasing the degree of the polynomial functions.

(3) **Pitch-dependency of an inharmonic component**
The distances for the mandolin (average reduction for the spectral distance: 31.21 %, and for the MFCC distance: -6.64 %) are shown in Figure 9 (g), (h). There was an improvement in the spectral distance. This result indicates that the relative amplitudes of the harmonic peaks of a synthesized sound are similar to those of a real sound in terms of pitch-dependency. However, there was no improvement in the MFCC distance. This is because the distribution of the inharmonic component of a synthesized sound differs from that of a real sound. Our method deals with the pitch-dependency of an inharmonic component according to the ratio of harmonic energy to inharmonic energy w_H/w_I, but not according to the distribution of the inharmonic component $M_I(f, r)$.

The distances for other struck and plucked string instruments such as the mandolin were improved a little in the MFCC distance. It is known that these sounds include a large inharmonic component at high frequencies during the attack segment [7]. The pitch-dependency of the inharmonic component of struck and plucked string instruments is therefore strong. When we tried listening to



(a) *Spectral distance for fagot* (b) *MFCC distance for fagot* (c) *Spectral distance for accordion* (d) *MFCC distance for accordion*

(e) *Spectral distance for marimba* (f) *MFCC distance for marimba* (g) *Spectral distance for mandolin* (h) *MFCC distance for mandolin*

Figure 9: *Spectral distances and MFCC distances between a synthesized sound and a real sound against the value of pitch manipulation (solid and dashed lines indicate the distances in our method and the baseline method, respectively).*

the synthesized sounds of these musical instruments, the harmonic signal seemed to be synthesized well, but the inharmonic signal did not sound natural. In contrast, there was good improvement for the piano, in spite of the results for other string instruments. This is because the pitch-dependency of the relative amplitudes of the harmonic peaks is stronger for the piano than for other string instruments.

In addition, the real sounds of struck and plucked string instruments include high overtones that do not exist strictly at integral multiples of the pitch; this is called *inharmonicity* [13]. Our harmonic model assumes that all harmonic peaks are strictly at integral multiples of the pitch, so it dose not perform well in analyzing the high harmonic peaks of these instruments. The results for struck and plucked string instruments were due to this assumption.

The musical instrument sounds synthesized using our method are available at:
http://winnie.kuis.kyoto-u.ac.jp/members/abe/DAFx-08/.

5. RELATED WORKS

In this section, we explain use of a phase vocoder and sinusoidal model as a representative method of analysis-manipulation.

5.1. Phase vocoder

The phase vocoder technique has a long and well-established history of use in synthesizing musical instrument sounds. There are many variations of the phase vocoder [14, 15, 16]. Sound synthesis is achieved by overlap-add synthesis. Duration is manipulated by expanding or shrinking the spectrogram on a time-scale and calculating the phase that matches neighboring frames. Pitch manipulation is achieved by re-sampling sounds after duration manipulation with the sampling rate multiplied by the reciprocal number of the ratio of pitch manipulation. One method of pitch manipulation expands or shrinks the spectrogram as well as manipulating the duration on a frequency-scale [17].

The phase vocoder distorts the inharmonic component of a musical instrument sound, which is a timbral feature because this technique manipulates the sound without dividing the harmonic component and inharmonic components. In addition, because this technique is a non-parametric method, it is difficult to analyze the timbral features as explicit parameters.

5.2. Sinusoidal model

The sinusoidal model is a well-known method of synthesizing the sounds of voices and musical instruments [10]. This technique tracks the peaks of a spectrogram and analyzes the frequencies of each peak and the amplitude of each peak on the time domain. Sound synthesis is achieved by adding the sinusoids, with the analyzed frequencies multiplied by the analyzed amplitudes. Duration is manipulated by expanding or shrinking the space of the analyzed peaks on the frequency domain and pitch is manipulated by expanding or shrinking the analyzed peaks on the time domain. Unlike the phase vocoder, the sinusoidal model does not require complex calculation of phases, so it can also be used to morph musical instrument sounds [18]. In addition, the sinusoidal model is applied to sound source-separation. Various methods for the parameter estimation have been reported [19, 20, 21].

The sinusoidal model deals with the inharmonic component as a timbral feature by using the rest spectrogram, which is the result of subtracting the tracked peaks from an analyzed sound. However, the timbral features are not defined as explicit parameters. The analysis of some musical instrument sounds has been dealt with only in morphing. The application of this technique has not included consideration of timbral characteristics (pitch-dependency of timbres).

6. CONCLUSION

We presented a method for manipulating the pitch and duration of musical instrument sounds that considers timbral features, which are defined as mathematical parameters. We defined three timbral features as (i) the relative amplitudes of the harmonic peaks, (iii) the inharmonic component, and (ii) temporal envelopes by referring to the spectrogram factors that correspond to difference in auditory effects as reported by Grey. When manipulating pitch, it is necessary to take into account the pitch-dependency of the features (i) and (iii). Therefore, we predicted the values of each feature by using a cubic polynomial that approximates the distribution of these features over pitches. In manipulating duration, it is necessary to preserve feature (iii) in the attack and decay segments of a *seed*. Therefore, only steady durations are expanded or shrunk. In addition, we proposed a method that can reproduce the properties of vibrato.

Future work will include applying our method to musical instrument parts separated from the polyphonic audio signals of commercial CD recordings. Because these separated sounds include various noises, it will be important that we select as much clean *seed* as we can. In addition, as analysis of the harmonic component of high tones and a consideration of the pitch-dependency of the duration of the inharmonic component were insufficient for synthesizing the sound of struck and plucked string instruments, we will try to improve our method for these instruments. We also plan to evaluate our method for duration manipulation.

7. ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Global Century of Excellence Program (GCOE), and the CrestMuse Project of the Japan Science and Technology Agency (JST).

8. REFERENCES

[1] O. Gillet and G. Richard, “Extraction and remixing of drum tracks from polyphonic music signals,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-05)*, New Paltz, New York, USA, 2005, pp. 315–318.

[2] K. Yoshii, M. Goto, and H. G. Okuno, “Drumix: An audio player with real-time drum-part rearrangement functions for active music listening,” *IPSJ Journal*, vol. 48, no. 3, pp. 1229–1239, 2007.

[3] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, “Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-07)*, Honolulu, Hawaii, USA, 2007, pp. 57–60.

[4] C. Uhle, C. Dittmar, and T. Sporer, “Extraction of drum tracks from polyphonic music using independent subspace analysis,” in *Proc. International Symposium on Independent Component Analysis and Blind Signal Separation (ICA-03)*, Nara, Japan, 2003, pp. 834–848.

[5] M. Helen and T. Virtanen, “Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine,” in *Proc. European Signal Processing Conference (EUSIPCO-05)*, Antalya, Turkey, 2005.

[6] D. Fitzgerald, M. Cranitch, and E. Coyle, “Sound source separation using shifted non-negative tensor factorization,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-06)*, Toulouse, France, 2006, pp. 653–656.

[7] J. M. Grey, “Multidimensional perceptual scaling of musical timbres,” *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1270–1277, 1977.

[8] J. Marozeau, A. Cheveigne, S. McAdams, and S. Winsberg, “The dependency of timbre on fundamental frequency,” *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2946–2957, 2003.

[9] T. Kitahara, M. Goto, and H. G. Okuno, “Musical instrument identification based on f0-dependent multivariate normal distribution,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, 2003, pp. 421–424.

[10] R. McAulay and T. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[11] M. Portnoff, “Implementation of the digital phase vocoder using the fast fourier transform,” *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, 1976.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Music genre database and musical instrument sound database,” in *Proc. International Symposium on Music Information Retrieval (ISMIR-03)*, Washington, DC, USA, 2003, pp. 229–230.

[13] H. Fletcher, E. Blackham, and R. Stratton, “Quality of piano. tones,” *J. Acoust. Soc. Am.*, vol. 34, no. 6, pp. 749–761, 1962.

[14] M. Dolson, “The phase vocoder: A tutorial,” *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[15] J. Laroche and M. Dolson, “Improved phase vocoder timescale modification of audio,” *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[16] A. Robel, “A new approach to transient processing in the phase vocoder,” in *Proc. Digital Audio Effect (DAFx-08)*, 2003, pp. 344–349.

[17] J. Laroche and M. Dolson, “New phase-vocoder techniques for real-time pitch-shifting, chorusing, harmonizing and other exotic audio modifications,” *J. Audio Eng. Soc*, vol. 47, no. 11, 1999.

[18] N. Osaka, “Concatenation and stretch/squeeze of musical instrumental sound using morphing,” in *Proc. International Computer Music Conference (ICMC-05)*, Barcelona, Spain, 2005.

[19] S. Godsill and M. Davy, “Bayesian harmonic models for musical pitch estimation and analysis,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-02)*, Orlando, Florida, USA, 2002, pp. 1769–1772.

[20] P. Jinachitra, “Constrained em estimates for harmonic source separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, 2003, pp. 609–612.

[21] N. Ono H. Kameoka and S. Sagayama, “Auxiliary function approach to parameter estimation of constrained sinusoidal model for monaural speech separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-08)*, Las Vegas, Nevada, USA, 2008, pp. 29–32.