EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

# Neural electric bass guitar synthesis framework enabling attack-sustain-representation-based technique control

Junya Koguchi[1*] and Masanori Morise[1]

**Abstract**

Musical instrument sound synthesis (MISS) often utilizes a text-to-speech framework because of its similarity to speech in terms of generating sounds from symbols. Moreover, a plucked string instrument, such as electric bass guitar (EBG), shares acoustical similarities with speech. We propose an attack-sustain (AS) representation of the playing technique to take advantage of this similarity. The AS representation treats the attack segment as an unvoiced consonant and the sustain segment as a voiced vowel. In addition, we propose a MISS framework for an EBG that can control its playing techniques: (1) we constructed a EBG sound database containing a rich set of playing techniques, (2) we developed a dynamic time warping and timbre conversion to align the sounds and AS labels, (3) we extend an existing MISS framework to control playing techniques using AS representation as control symbols. The experimental evaluation suggests that our AS representation effectively controls the playing techniques and improves the naturalness of the synthetic sound.

**Keywords** Musical instrument sound synthesis, Playing technique, Electric bass guitar, Phoneme, Deep neural networks

## 1 Introduction

Musical instrument sound synthesis (MISS) is a technique employed for artificially synthesizing performance audio from musical scores. MISS technology is pivotal for fully digital music production techniques known as in the box or desktop music. Traditional methods have utilized deterministic approaches such as concatenative [1] and physical modeling [2] synthesis. Their synthetic sounds follow numerical values of the score (such as MIDI) perfectly. Therefore, their users must program the fluctuations in order to synthesize a human-like performance. To synthesize realistic performance, data-driven approaches using deep neural networks (DNNs) have been investigated, thanks to large sound databases [3–5]. Since DNN-MISS learns the probability distribution of human performance, its synthetic performance is with fluctuations as a default output. The DNN enables MISS to simulate expressive components such as articulation, as seen in trills and staccatos. VirtuosoNet [6] accomplishes expressive piano sound synthesis by embedding the articulation symbols notated on score information. Other prior works have studied MISS via latent representations using encoder-decoder models to consider a musical context [7, 8]. Moreover, some DNN-based MISSs frequently utilize text-to-speech and singing voice synthesis techniques, anticipating that both speech synthesis and MISS share a common framework for producing acoustic signals from symbols. Tacotron2 [9],

*Correspondence:
Junya Koguchi
korguchi@meiji.ac.jp
[1] Graduate School of Advanced Mathematical Sciences, Meiji University, 4-21-1, Nakano, Tokyo 1648525, Japan

which has achieved human-comparable quality, is used as an acoustic model [10]. Deep Performer [11] is based on FastSpeech architecture [12], which is a controllable and fast TTS framework. These techniques, which successfully incorporate high-quality speech synthesis knowledge, hold promise for improving the quality and controllability of DNN-based MISS.

Nonetheless, developing a DNN-based MISS that can synthesize human-like expressive performances remains a challenging task. One of the reasons lies in a lack of timbral technique controllability. Compared with articulations, which are performance expressions controlling pitch, duration, and velocity, the timbral techniques such as pizzicato and bowing express timbre variations. Their notations in musical scores differ depending on the instruments. One approach is to tokenize tablature, which provides richer performance information than regular musical score [13–15]. Moreover, commercial MIDI-to-audio products use out-of-range notes as keyswitches and velocity ranges are classified according to the techniques [16, 17]. However, multiple streams must be prepared when dealing with a note which includes multiple techniques. This complicates the annotation process and hinders the development of the automatic playing technique recognition and controllable MISS. The efficient representation of playing techniques to annotate large sound datasets is limited.

These challenges, particularly for monophonic plucked instruments such as the electric bass guitar (EBG), could be addressed by further leveraging speech synthesis techniques. The EBG is used in various music, like the piano, which is the target of many MISS. It is a vital instrument in modern music. Its synthesis system is also essential for ITB. Some characteristics of EBG are suitable for MISS. Although an EBG with multiple strings can play chords, its primary role in an ensemble is to provide the bass and rhythm of a monophonic melody. The EBG can be recorded with a direct input, providing a clear performance signal free from background noise. We moreover focus on the acoustic similarities between EBG sound and speech. Transient acoustic changes from picking noise to string vibration occur in an EBG sound. A technique classification scheme focusing on these differences in plucking style and performance expression has also been proposed [18]. This phenomenon is similar to the relationship between consonants and vowels in speech. And this acoustic change is expected to be represented by a discrete symbol sequence of playing techniques, just as phonemes are used to represent phonological changes. On the other hand, the string vibrations, which constitute the integer harmonics, change in their timbre through the pickups. This suggests the potential applicability of the source-filter model [19], a model that approximates glottal vibration and filters vocal tract characteristics in speech. Actually, a digital waveguide model synthesizes an EBG sound by exciting frequency loss filters using physically modeled string vibrations. Acoustic features that are effective for speech, such as mel cepstrum, can be used for EBG as well.

In this paper, we propose an attack-sustain (AS) representation of the timbral techniques of plucked string instruments (Fig. 1). We categorize the techniques into two types based on whether they primarily affect the attack or sustain segment. The corresponding techniques then label each attack and sustain segment of the note. The AS representation allows multiple technique labels on a single note, enhancing versatility. Also, it can be used for percussive performances with only picking noise without sustain segments like a consonant cluster in speech. We also developed playing-technique-controllable MISS by extending the existing MISS framework. We assign the AS labels to our EBG sound database [20] and align the label and recorded sound temporally. This alignment was obtained using iteration of dynamic time warping (DTW) [21] and timbre conversion [22, 23]. We utilize Deep Performer as the base MISS framework and enable the control of playing techniques through the AS label input. This is because Deep Performer has high controllability and extensibility, which offers accurate duration control and allows polyphonic notes.

In our experimental evaluation, we objectively assessed the effectiveness of our AS labels in controlling the playing techniques by calculating the DTW alignment scores to the concatenative synthetic sound. In addition, the quality of the synthesized sound using the AS representation was subjectively evaluated through listening experiments. The results suggest that our AS representation is not only effective for controlling the playing techniques but also enhances the naturalness of the synthetic sound.

The rest of this paper is organized as follows. In Section 2, we briefly review Deep Performer as the base MISS framework. In Section 3, we explain the details of the proposed AS label and the playing technique-controllable MISS. In Section 4, we discuss the evaluation of the controllability and the synthetic sound quality of our framework. In Section 5, we conclude the paper.

## 2 Base framework: deep performer

In this section, we describe the structure of the base framework, deep performer [11]. The acoustic model of the Deep Performer is composed of two transformer encoder-decoder [24] models called the alignment model for embedding musical score information and the synthesis model for mel spectrogram generation, respectively.

The alignment model's input of musical information is embedded in parallel as multiple vectors of the same

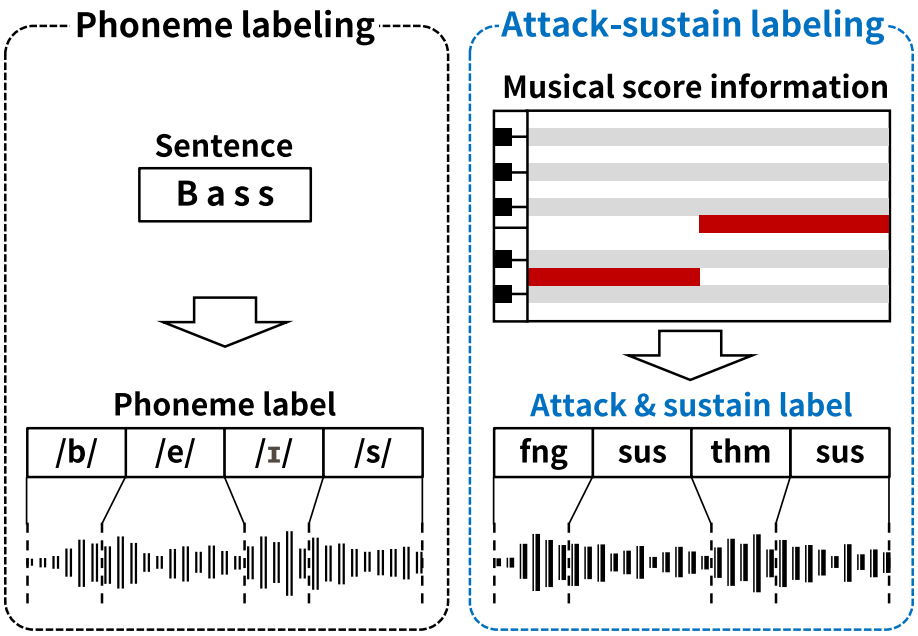


**Fig. 1** Proposed attack-sustain label contrasted with phoneme label. For example, two notes played by finger picking (fng) and thumping (thm) are converted to the labels "fng-sus" and "thm-sus," respectively

dimension for each note as tokens of pitch, onset, duration, and velocity. These are added together with the positional encoding and passed to the transformer encoder. The output vector from the encoder is added to the performer identification (ID) and tempo embedding and passed to the linear layer as a note embedding. The linear layer outputs the onset and duration as frame length and index. The loss function for training is the mean squared error of the onset and duration with the ground truth.

The synthesis model, on the other hand, infers a synthetic mel-spectrogram based on the onset and duration from the alignment model and the score information. Similar to the alignment model, it first encodes tokens of score information for each note and obtains a note embedding by adding the performer identification (ID) and tempo embedding. The note embeddings are then input to the polyphonic mixer and passed to the transformer decoder to process the polyphony. The polyphonic mixer duplicates the note embeddings by the number of frames in each duration, shifts them according to the onset, and adds all vectors in parallel. In addition, note-wise position encoding (NWPE) is applied when the note embedding vectors are converted to frame-wise vectors. The NWPE is applied to a note embedding $\boldsymbol{v}_{\text{note}}$ with $p \in [0, 1]$ as its relative position in the note as follows:

$$\boldsymbol{v}_{\text{frame}} = (1 + p\boldsymbol{w}) \odot \boldsymbol{v}_{\text{note}}. \tag{1}$$

where $\odot$ denotes the Hadamard product and $w$ is a learnable vector, initialized with a small random number so that $\boldsymbol{v}_{\text{frame}} \approx \boldsymbol{v}_{\text{note}}$. The NWPE is expected to condition the temporal timbre change of the note to the decoder. Finally, the decoder converts the frame embedding sequence into a mel spectrogram. The loss function for training is the mean squared error of the mel spectrogram with the ground truth.

## 3 Attack-sustain representation

In this section, we describe the details of the proposed AS label and the playing technique-controllable MISS.

### 3.1 AS representation and label design

The EBG signals are generated by hitting the strings with a finger or pick. The strings collide with the pick/finger/fret, generating aperiodic noise. Then, depending on the playing technique (mute/harmonics/etc.), periodic string vibrations are generated and slowly decay. Focusing on this generation process, it is suggested that the acoustic differences in playing techniques can be broadly classified into those that appear in the attack segment and those that appear in the sustain segment [18]. For speech, phonemes are given as linguistic discrete symbols representing phonological changes, while for EBG sounds, playing techniques correspond as musical discrete symbols representing transient acoustic changes.

Table 1 lists techniques corresponding to attack and sustain (hereinafter, they are called "attack technique" and "sustain technique," respectively). Techniques that affect string vibration, such as mute and harmonics

**Table 1** The list of playing techniques corresponding attack and sustain labels

| Attack | Sustain |
|---|---|
| Finger, pick, thump, thumb up, pluck, hammer on, pull off | Sustain, mute, harmonics |

techniques, are distinguished. We assign "pause" to a silent segment, such as a rest, following the phoneme dictionary [25].

### 3.2 Automatic alignment method

Controllable systems typically use explicit temporal segmented data [10, 11]. However, manual annotation requires well-experienced annotators to detect segment boundaries. A common automatic method in speech processing is a Viterbi alignment based on hidden Markov models (HMMs) [26]. HMMs are trained using pairs of label sequences and acoustic features, and the Viterbi path is a temporal alignment of the technique label sequences. However, because the HMM is based on switching stationary signal sources, it is hard to model slowly decaying string vibration. Another method is an alignment to synthetic EBG sounds from the musical scores using existing concatenative synthesizer [11]. By annotating all sample units with AS labels in advance, the synthesizer outputs the sounds strictly to the labels. The AS label boundaries for the target sound can be obtained by calculating the temporal deviation between this synthetic and the target sound using the DTW. The DTW stretches one sequence and matches the other. It firstly calculates the distance between all points and finds the temporal correspondence that minimizes the distance between the two sequences based on dynamic programming.

Since the timbres differ between synthetic and recorded sounds, this deteriorates the alignment accuracy of the DTW. To reduce this problem, we utilize timbre conversion during the DTW using a voice conversion (VC) technique (Fig. 2). It has shown efficacy alignment of singing voice [27] and is also effective for EBG due to its acoustic similarity to speech [20]. First, the alignment of synthetic and recorded sound is obtained as described above. Next, using the aligned sound, a Gaussian mixture model (GMM) [22, 28] is trained to transform the synthetic acoustic features into the recorded ones. The DTW is again applied between the converted and the recorded sound. This method is expected to be more accurate in alignment because the distribution of acoustic features is closer to the recorded sound. The DTW is a dynamic programming based algorithm, and the GMM parameters are updated only with the given single pair data.

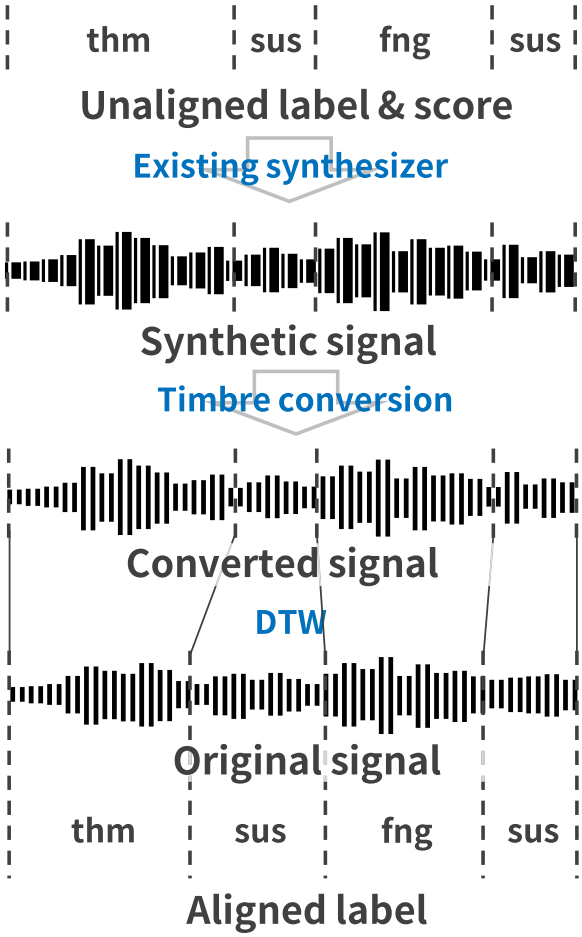


**Fig. 2** The overview of alignment algorithm based on the DTW and timbre conversion

Therefore, the alignment accuracy is independent of the amount of data. In addition, it is known that the DTW and timbre conversion can be sufficiently accurate in a single iteration [23].

### 3.3 Playing technique embedding of the AS label

We extend the deep performer framework using AS representations to control the playing technique. The musical context input to the encoder is embedded with not only notes but also the AS labels converted from playing technique information (Fig. 3).

For naive labels that assign one playing technique to one note (hereinafter referred to as note-wise labels), note and playing technique correspond one-to-one. On the other hand, there can be more than one AS label corresponding to a single note. We solve this using the same technique as in singing voice synthesis [29]. The note information is duplicated and embedded to correspond to the number of AS labels (Fig. 4). Since the deep

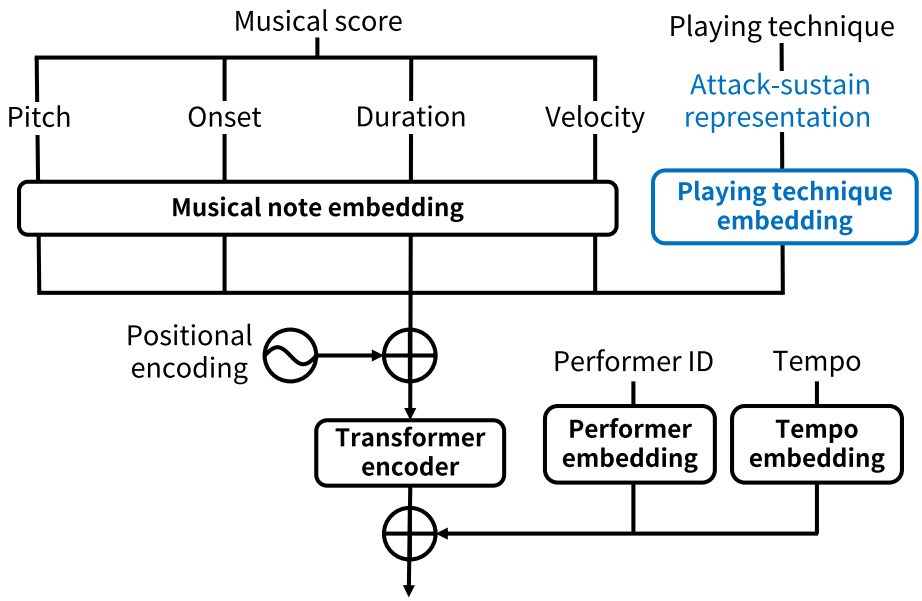


**Fig. 3** The encoder part of proposed MISS framework using attack–sustain representation. The playing technique information obtained from the score is converted to the attack–sustain labels and embedded

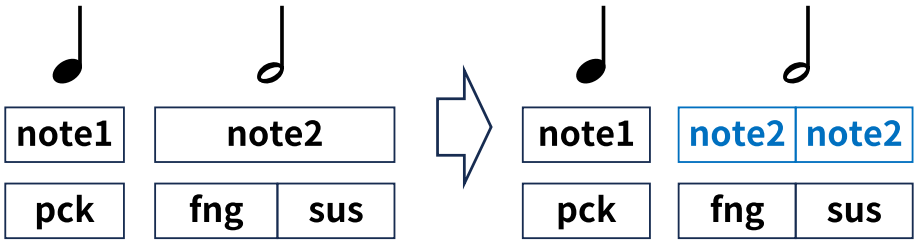


**Fig. 4** Note division technique based on singing voice synthesis

performer is based on a speech synthesis framework, it can be extended to a performance-controllable framework by treating AS labels in the same way as phoneme labels.

### 3.4 Waveform generation from mel spectrogram

The mel spectrogram output from the decoder is converted into a waveform by the waveform generation model. Unlike speech, electric bass sounds need to be synthesized robustly over a low, wide frequency range. Therefore, we use BigVGAN [30] of which a well-trained model performs robustly for instrumental sounds out-of-distribution of the training speech data. High-quality waveform generation is also expected for EBG sounds with various pitch and technique variations. We finetune the pre-trained model on our EBG sound database to polish the synthetic sound quality.

## 4 Experimental evaluation

This section describes configurations and discusses the results of the experimental evaluation.

### 4.1 Condition

#### *4.1.1 Dataset*

The EBG sound database was constructed to evaluate the accuracy with respect to actual acoustic signals for training the DNN [20]. The sounds used were 180 phrases of four bars of the monophonic bass line (approximately 112 min), containing all techniques in the list (Table 1), and each with a various tempo between 60 to 200 beats per minute (BPM). The AS label sequences were annotated manually before the alignment. The note-wise label was also annotated as a combined label of the attack and sustain labels. A finger-picked note, for example, was annotated as “fng” and “sus” for the AS label, “fng-sus” for the NW label.

**Table 2** The configuration of the DTW and timbre conversion

| | |
|---|---|
| Sampling rate | 24 kHz |
| Hop length | 240 |
| Window length | 1024 |
| Dimension of mel cepstrum | 24 |
| GMM mixture | 16 |

**Table 3** The configuration of the alignment model

| | |
|---|---|
| Encoder layer | 3 |
| Multi-head attention heads | 2 |
| Multi-head attention hidden units | 64 |
| Feed-forward network hidden units | 256 |
| Feed-forward network kernel sizes | 9, 1 |
| Max sequence | 1000 |
| Max time | 96 |
| Max duration | 96 |

The electric bass used was a Fender custom shop 1962 Jazz Bass [31], the audio interface was an RME ADI-2 Pro FS R [32], and an experienced player recorded the performance.

For the alignment between the labels and sounds, the audio was downsampled to 24 kHz. 24-dimensional mel-cepstral coefficients were used as acoustic features for the GMM-based timbre conversion and DTW. The number of Gaussian mixtures was 16 for the GMM-based timbre conversion. For the DTW, we used Standard Bass V2 [33] as the concatenative synthesizer. Each sample was manually labeled and aligned in advance. The alignment score of the DTW was calculated based on the mean squared error with the mel cepstrum. Table 2 compiles these conditions in a table.

#### *4.1.2 Condition of the DNNs*

The basic conditions were the same as the Deep Performer [11]. The pitch, onset, duration, and velocity used for note embedding were extracted from the note information in MIDI at a time resolution of 24 quarter-note divisions. The pitch was used in MIDI note number. The onset and duration are used in frame. The velocity was used on a linear scale in decibels, with the smallest note in the dataset set as 10 and the loudest as 127. The performer ID embedding was omitted because all data used are by the single performer. The audio setting was the same as the label alignment (Table 2).

**Table 4** The configuration of the synthesis model

| | |
|---|---|
| Encoder layer | 3 |
| Decoder layer | 6 |
| Multi-head attention heads | 2 |
| Multi-head attention hidden units | 128 |
| Feed-forward network | 512 |
| Feed-forward networks | 9, 1 |
| Max sequence | 1000 |
| Max time | 96 |
| Max duration | 96 |
| Mel spectrum bands | 100 |

**Table 5** The hyperparameters of the alignment and synthesis model

| | |
|---|---|
| Batch size | 16 |
| Dropout | 0.2 |
| Adam optimizer $\beta_1$ | 0.9 |
| Adam optimizer $\beta_2$ | 0.98 |
| Adam optimizer $\epsilon$ | $10^{-9}$ |
| Learning rate annealing steps (alignment) | 1000 |
| Learning rate annealing rate (alignment) | 0.5 |
| Gradient clipping threshold | 1.0 |
| Warm up steps (alignment) | 1000 |
| Warm up steps (synthesis) | 4000 |
| Training steps (alignment) | 10,000 |
| Training steps (synthesis) | 100,000 |

The model and training configurations are compiled in Tables 3, 4, and 5. For the alignment model, all embedding dimensions were 128-dimensional, and the encoder and decoder consisted of feed-forward transformer blocks. Each block consisted of a Multi-head attention and a position-wise feed-forward network sub-layer. Each MHA layer has 64 hidden units and 2 attention heads. Each feed-forward network layer had 256 hidden units with kernel sizes of 9 and 1 for the two convolutional layers. The maximum length of the sequences was 1000, and the maximum duration per label was 96 frames. The synthesis model was almost the same as the alignment model but differs in 6 decoder layers, and in that the hidden layers of multi-head attention and feed-forward transformer were each double the size. The decoder outputted 100-dimensional mel spectrograms. We also followed the Deep Performer for training the models. The batch size was 16, and the dropout rate was 0.2 after each sub-layer. We trained the alignment and synthesis models separately, and the learning rate annealing schedule was used for the alignment model. We trained the alignment model for 10,000 and the synthesis model for 100,000, respectively. Adam [34] was used for the optimizer.

**Table 6** The acoustic configuration of the BigVGAN

| | |
|---|---|
| Sampling rate | 24 kHz |
| Mel bands | 100 |
| Hop length | 256 |
| Window length | 1024 |
| Segment size | 8192 |

**Table 7** The training configuration of the BigVGAN

| | |
|---|---|
| Adam optimizer $\beta_1$ | 0.9 |
| Adam optimizer $\beta_2$ | 0.98 |
| Adam optimizer $\epsilon$ | $10^{-8}$ |
| Weight decay | 0.01 |
| Batch size | 32 |
| Training steps | 100,000 |

The BigVGAN, for waveform generation, was trained by fine-tuning a pre-trained model[1] provided in the author's GitHub repository [35]. The mel spectrogram calculation and optimization algorithms were the same as for the acoustic model. The batch size was 32, and segment size was 8192, and training was performed in 100,000 steps. AdamW [36] was used for the optimizer. These conditions are compiled in Tables 6 and 7

#### *4.1.3 Evaluation of technique controllability*

Experiments were conducted to evaluate the playing technique controllability of the proposed MISS. First, pairs of synthetic sounds with various playing techniques controlled by the existing concatenative and proposed MISS were prepared. The alignment score obtained by DTW between them was then evaluated. The proposed MISS synthesizes using onsets and durations in the score rather than the outputs from the alignment model. The alignment score of the DTW between the proposed MISS sound without technique control (TC) and concatenative sound with TC is high. On the other hand, we expect lower alignment scores for the proposed MISS sounds with TC. This gap allows us to evaluate the accuracy of the playing technique control. We evaluate the contollability of a attack technique and a sustain technique. The attack technique control is a replacement from plectrum picking to fingerpicking, and the sustain technique control is a replacement from regular sustain to mute.

We compared the synthetic sounds generated from following methods.

[1] bigvgan_base_24khz_100band

**Table 8** The alignment score of synthetic sound with or without playing technique control. If the score is lowered by technique control, it indicates it is accurate control

| | Alignment score with CS w/ TC | |
|---|---|---|
| **Method** | **Finger to plectrum** | **Sustain to mute** |
| CS w/o TC | 7.36 | 5.10 |
| AS w/o TC | 10.46 | 18.79 |
| AS w/ TC | 5.35 | 6.14 |
| NW w/o TC | 9.46 | 15.11 |
| NW w/ TC | 7.35 | 8.98 |

- CS w/o TC: Concatenative synthetic sound with technique control.
- AS w/o TC: DNN-MISS synthetic sound without the proposed AS-based technique control.
- AS w/ TC: DNN-MISS synthetic sound with the proposed AS-based technique control.
- NW w/o TC: DNN-MISS synthetic sound without the note-wise technique control.
- NW w/ TC: DNN-MISS synthetic sound with the note-wise technique control.

#### *4.1.4 Evaluation of synthetic sound quality*

To evaluated the quality of the synthetic sound by using our framework, we conducted five-scale mean opinion score (MOS) tests on naturalness. Forty-eight listeners participated in the experiment on a crowdsourcing platform. Each listener answered to 100 EBG sound samples. We compared the synthetic sounds generated from following methods.

- Ground truth: Natural EBG sound performed by human.
- Analysis-synthesis: BigVGAN synthetic sound of the ground truth.
- AS w/ NWPE: our AS-label-based DNN-MISS with the NWPE.
- NW w/ NWPE: NW-label-based DNN-MISS with the NWPE.
- AS w/o NWPE: our AS-label-based DNN-MISS without the NWPE.
- NW w/o NWPE: NW-label-based DNN-MISS without the NWPE.

### 4.2 Result and discussion

The results of the objective evaluation in playing technique control are shown in Table 8. The alignment score of the sound with AS playing technique control is smaller

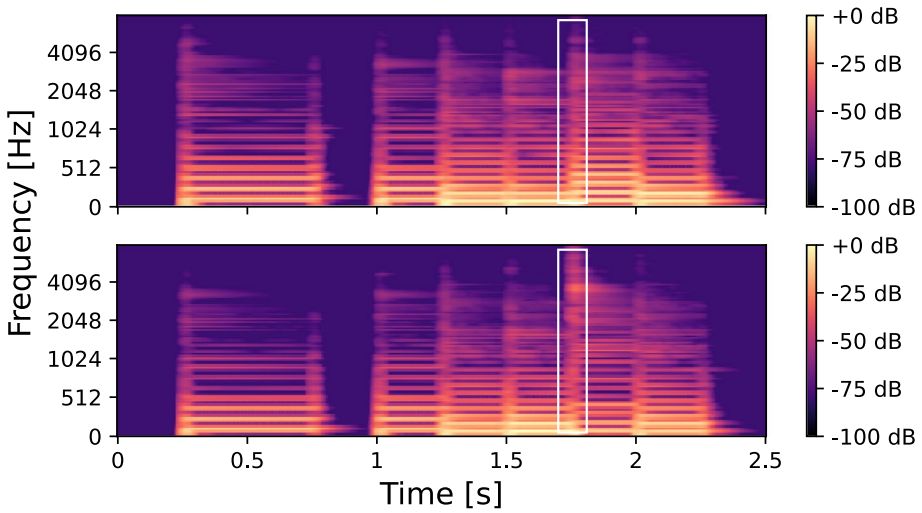


**Fig. 5** The mel spectrograms of synthetic sounds. The upper one consists entirely of finger picking, while in the lower one only the attack technique on the sixth note was replaced by plectrum picking

than the one without the control. This result indicates that our AS label is more effective in controlling playing technique than the naive note-wise label. In addition, in technique control from sustain to mute, the score of CS are lower even without TC. The muting technique is played by softly touching the strings with the fingers or palms, resulting in a quick decay not only a timbre change. Thus, the result suggests that each duration of the MISS methods is different from the duration on the score.

The mel spectrograms of the actual synthetic instrumental sounds are shown in Fig. 5. The upper row consists entirely of finger picking, while in the lower row, only the attack technique on the sixth note was replaced by plectrum picking. Focusing on that attack section surrounded by a white square shows a prominet

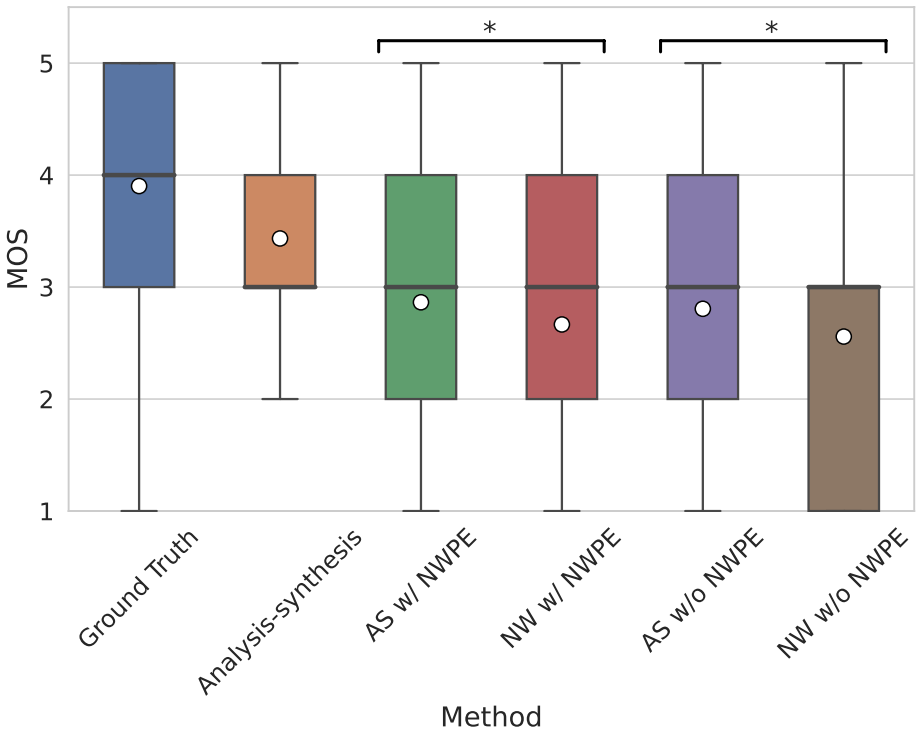


**Fig. 6** MOS scores for naturalness of ground truth and synthetic sounds. The white circles denote the means

non-periodic spectrum. Plectrum picking is a playing technique in which the strings are plucked with a pick. Compared to finger picking, the strings collide the frets more strongly. Therefore, this acoustic variation is reasonable as a result of the playing technique control.

The results of the subjective experimental evaluation are shown in Fig. 6. Results of Welch's $t$-test for AS w/ NWPE and NW w/ NWPE, AS w/o NWPE and NW w/o NWPE, $p$-values corrected by Bonferroni were below the 5% significance level. This result indicates that our AS label improves the MISS sound quality more than the NW label. This improvement is by using explicit discrete symbols to represent acoustic changes within the note. In Japanese text-to-speech synthesis, it is known that the quality of synthetic speech is improved when phoneme-level representation is used rather than syllable-level [37]. Our result is consistent with this because the AS and NW labels correspond to the phoneme and syllable levels, respectively. However, gaps still exist in the scores compared to the ground truth. This is due to a lack of data and insufficient alignment accuracy.

## 5 Conclusion

This paper proposed the AS label inspired by phoneme representation. By labeling the playing technique changes separately into attack and sustain techniques, as in the case of vowels and consonants, the method in speech processing can also be applied to EBG signals.

We propose the MISS framework for the EBG that can control the playing technique: (1) we constructed a sound database containing a rich set of playing techniques for electric bass guitar, (2) we developed a dynamic time-stretching and timbre-translation system to temporally correspond sounds to AS labels, (3) a controllable speech synthesis framework was applied to MISS. The experimental evaluation suggests that the proposed AS representation improves naturalness in addition to being effective for playing technique control. Future work includes extensions to polyphony and other musical instruments.

### Abbreviations

| | |
|---|---|
| MISS | Music instrument sound synthesis |
| DNN | Deep neural network |
| MIDI | Musical instrument digital interface |
| EBG | Electric bass guitar |
| AS | Attack-sustain |
| DTW | Dynamic time warping |
| ID | Identification |
| NWPE | Note-wise positional encoding |
| HMM | Hidden Markov model |
| VC | Voice conversion |
| GMM | Gaussian mixture model |
| NW | Note-wise |
| BPM | Beat per minute |
| TC | Technique control |
| CS | Concatenatve synthesis |
| MOS | Mean opinion score |

### Acknowledgements

The authors would like to thank Prof. Hideki Kawahara for helpful reviews and discussions.

### Authors' contributions

JK implemented the framework, conducted the experimental evaluations, and wrote the original draft. MM supervised this study and edited the manuscript. All authors read and approved the manuscript.

### Funding

This work was supported by JSPS KAKENHI Grant Number JP22KJ2855 and JP21H04900.

### Availability of data and materials

The datasets and implementations used during the current study are not publicly available due to copyright and abuse prevention but are available from the corresponding author on reasonable request. Please contact the corresponding author's e-mail address (korguchi@meiji.ac.jp) with the information of the user, affiliation, and purpose.

## Declarations

### Competing interests

The authors declare that they have no competing interests.

Received: 31 July 2023 Accepted: 4 January 2024
Published online: 11 January 2024

### References

1. D. Schwarz, Concatenative sound synthesis: the early years. J. New Music Res. **35**(1), 3–22 (2006). https://doi.org/10.1080/09298210600696857
2. S. Bilbao, C. Desvages, M. Ducceschi, B. Hamilton, R. Harrison-Harsley, A. Torin, C. Webb, Physical modeling, algorithms, and sound synthesis: the NESS project. Comput. Music. J. **43**(2–3), 15–30 (2019). https://doi.org/10.1162/comj_a_00516
3. M. Goto, H. Hashiguchi, T. Nishimura, R. Oka, in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, RWC music database: Music Genre Database and Musical Instrument Sound Database (2003), pp. 229–230. https://staff.aist.go.jp/m.goto/RWC-MDB/. Accessed 10 Jan 2024
4. C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.Z.A. Huang, S. Dieleman, E. Elsen, J. Engel, D. Eck, in *Proc. International Conference on Learning Representations (ICLR)*, Enabling factorized piano music modeling and generation with the MAESTRO dataset (2019). https://openreview.net/forum?id=r1lYRjC9F7. Accessed 10 Jan 2024
5. J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, K. Simonyan, in *Proc. International Conference on Machine Learning (ICML)*, Neural audio synthesis of musical notes with wavenet autoencoders (2017), pp. 1068–1077. https://proceedings.mlr.press/v70/engel17a.html. Accessed 10 Jan 2024
6. D. Jeong, T. Kwon, Y. Kim, K. Lee, J. Nam, in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, VirtuosoNet: a hierarchical RNN-based system for modeling expressive piano performance (2019), pp. 908–915. https://doi.org/10.5281/zenodo.3527962
7. Y. Wu, E. Manilow, Y. Deng, R. Swavely, K. Kastner, T. Cooijmans, A. Courville, C.Z.A. Huang, J. Engel, in *Proc. International Conference on Learning Representations (ICLR)*, MIDI-DDSP: Detailed control of musical performance via hierarchical modeling (2022). https://openreview.net/forum?id=UseMOjWENv. Accessed 10 Jan 2024
8. B. Wang, Y.H. Yang, PerformanceNet: Score-to-Audio music generation with Multi-Band convolutional residual network. Proc. Am. Assoc. Artif. Intell. (AAAI) Symp. Ser. **33**(01), 1174–1181 (2019). https://doi.org/10.1609/aaai.v33i01.33011174

9. J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R.A. Saurous, Y. Agiomvrgiannakis, Y. Wu, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions (2018), pp. 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368
10. E. Cooper, X. Wang, J. Yamagishi, in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, Text-to-Speech Synthesis Techniques for MIDI-to-Audio Synthesis (2021), pp. 130–135. https://doi.org/10.21437/ssw.2021-23
11. H.W. Dong, C. Zhou, T. Berg-Kirkpatrick, J. McAuley, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Deep performer: Score-to-audio music performance synthesis (2022), pp. 951–955. https://doi.org/10.1109/ICASSP43922.2022.9747217
12. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.Y. Liu, FastSpeech: Fast, robust and controllable text to speech. Proc. Adv. Neural Inf. Process. Syst. (NeurIPS). **32** (2019). https://proceedings.neurips.cc/paper_files/paper/2019/hash/f63f65b503e22cb970527f23c9ad7db1-Abstract.html. Accessed 10 Jan 2024
13. C. Goddard, *Virtuosity in computationally creative musical performance for bass guitar* (Ph.D. thesis, Queen Mary University of London, 2021)
14. P. Sarmento, A. Kumar, C. Carr, Z. Zukowski, M. Barthet, Y.H. Yang, in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, DadaGP: A Dataset of Tokenized GuitarPro Songs for Sequence Models (2021), pp. 610–617
15. J. Lotha, P. Sarmento, C. Carr, Z. Zukowski, M. Barthet, ProgGP: From GuitarPro Tablature Neural Generation To Progressive Metal Production. in Proc. 16th International Symposium on Computer Music Multidisciplinary Research (CMMR) (2023), pp. 122-133
16. Toontrack. Ezbass. https://www.toontrack.com/product/ezbass/. Accessed 10 Jan 2024
17. I. Multimedia. Modo bass 2. https://www.ikmultimedia.com/products/modobass2/. Accessed 10 Jan 2024
18. J. Abeßer, H. Lukashevich, G. Schuller, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Feature-based extraction of plucking and expression styles of the electric bass guitar (2010), pp. 2290–2293. https://doi.org/10.1109/ICASSP.2010.5495945
19. G. Fant, *Acoustic Theory of Speech Production With Calculations based on X-Ray Studies of Russian Articulations* (De Gruyter Mouton, 1971)
20. J. Koguchi, M. Morise, Phoneme-inspired playing technique representation and its alignment method for electric bass database. in Proc. 16th International Symposium on Computer Music Multidisciplinary Research (CMMR) (2023), pp. 170-177
21. N. Hu, R. Dannenberg, G. Tzanetakis, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* Polyphonic audio matching and alignment for music retrieval (2003), pp. 185–188. https://doi.org/10.1109/ASPAA.2003.1285862
22. T. Toda, A.W. Black, K. Tokuda, Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio Speech Lang. Process. **15**(8), 2222–2235 (2007). https://doi.org/10.1109/TASL.2007.907344
23. G. Kotani, H. Suda, D. Saito, N. Minematsu, in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Experimental investigation on the efficacy of Affine-DTW in the quality of voice conversion (2019), pp. 119–124. https://doi.org/10.1109/APSIPAASC47483.2019.9023107
24. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Proc. Adv. Neural Inf. Process. Syst. (NeurIPS). **30** (2017). https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed 10 Jan 2024
25. Weide, R. L., The Carnegie Mellon pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Accessed 10 Jan 2024
26. F. Brugnara, D. Falavigna, M. Omologo, Automatic segmentation and labeling of speech based on Hidden Markov Models. Speech Commun. **1**(4), 357–370 (1993). https://doi.org/10.1016/0167-6393(93)90083-W
27. J. Koguchi, S. Takamichi, M. Morise, in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, PJS: phoneme-balanced Japanese singing-voice corpus (2020), pp. 487–491. https://ieeexplore.ieee.org/document/9306238. Accessed 10 Jan 2024
28. K. Kobayashi, T. Toda, in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, sprocket: Open-Source Voice Conversion Software (2018), pp. 203–210. https://doi.org/10.21437/Odyssey.2018-29
29. P. Lu, J. Wu, J. Luan, X. Tan, L. Zhou, XiaoiceSing: A High-Quality and integrated singing voice synthesis system (2020). arXiv:2006.06261
30. S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, S. Yoon, in *Proc. International Conference on Learning Representations (ICLR)*, BigVGAN: A universal Neural Vocoder with Large-Scale Training (2023). https://openreview.net/forum?id=iTtGCMDEzS_. Accessed 10 Jan 2024
31. Fender Custom Shop. 1962 jazz bass. https://www.fendercustomshop.com/basses/jazz-bass/. Accessed 5 Sept 2022
32. RME. ADI-2 Pro FS R. https://www.rme-audio.de/adi-2-pro-fs-be.html. Accessed 31 July 2023
33. Purple_Shikibu_. Standard Bass V2. https://unreal-instruments.wixsite.com/unreal-instruments/standard-bass. Accessed 31 July 2023
34. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2014). arXiv:1412.6980
35. NVIDIA. BigVGAN [source code]. https://github.com/NVIDIA/BigVGAN. Accessed 31 July 2023
36. I. Loshchilov, F. Hutter, in *Proc. International Conference on Learning Representations (ICLR)*, Decoupled Weight Decay Regularization (2019). https://openreview.net/forum?id=Bkg6RiCqY7. Accessed 10 Jan 2024
37. T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, in *10th ISCA Speech Synthesis Workshop (SSW 10)*, Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis (2019). https://doi.org/10.21437/SSW.2019-30


## Publisher's Note