Phoneme-inspired playing technique representation and its alignment method for electric bass database

Junya Koguchi and Masanori Morise *

Meiji University
{korguchi, mmorise}@meiji.ac.jp

Abstract. In plucked string instruments such as electric bass, the attack phase is dominated by non-periodic components resulting from picking noise, while the sustain phase is dominated by periodic components resulting from string vibrations. This phenomenon is analogous to unvoiced consonants and voiced vowels in speech, suggesting the possibility of applying speech phoneme representations to plucked string instrument playing techniques. In this study, we design playing technique labels for an electric bass database by treating the attack phase as consonants and the sustain-to-decay phase as vowels. Furthermore, we employ a phoneme alignment algorithm to obtain the alignment between the playing technique labels and the acoustic signals of the electric bass. To conduct experiments, we construct a electric bass database and apply methods based on hidden Markov models and dynamic time warping. As a result, methods based on dynamic time warping, particularly those incorporating timbre transformations, provided the most accurate alignment.

Keywords: Electric bass, playing technique, phoneme aligment, hidden Markov model, dynamic time warping

1 Introduction

The advancement of musical information retrieval research is supported not only by machine learning and signal processing techniques, but also by open sound databases. Many of these databases include not only sound data but also annotation data. Sound databases with useful annotations accelerate research and enhance reproducibility. For instance, the presence of musical score information like MIDI can assist in automatic transcription and sound synthesis [1, 2], while attributes such as genre can aid in music information retrieval [3]. Furthermore, playing technique information plays a crucial role in accurately representing their timbre and articulations.

When considering applications for controllable instrument sound synthesis [4] and playing technique recognition [5], it is essential to include detailed information on

* This work was supported by JSPS KAKENHI Grant Number JP22J22158 and JP21H04900.

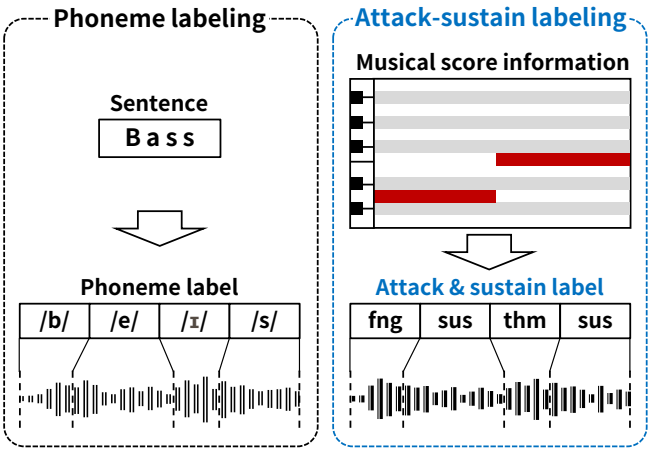


Fig. 1. Proposed attack-sustain label contrasted with phoneme label. For example, two notes played by finger picking and thumping are converted to the labels "fng-sus" and "thm-sus", respectively.

changes in playing techniques in addition to musical score information. However, there is no standard format for expressing playing techniques in MIDI. Some software synthesizers implement out-of-range notes as key switches for changing playing techniques [6, 7], but the types and assignments of these techniques vary among developers. Many databases provide only note information but playing technique information. This is due to the time-consuming annotation process. In addition, since performance techniques may change independently for each note of a multipitch instrument, it is difficult to track them on a single time axis.

This problem might be solvable, at least for electric bass signals, by applying insights from speech processing. Firstly, it is reasonable to assume a monophonic melody in normal performances. Although electric basses with multiple strings can play chords, their role within an ensemble is to provide a monophonic bass and rhythm part. Moreover, in the attack phase of electric bass, non-periodic components dominate due to picking noise, while periodic components dominate during the sustain phase due to string vibrations. Electric bass playing techniques can be broadly divided into those that change the attack phase, such as fingerpicking and slapping, and those that change the sustain phase, like harmonics and muting [8]. This is similar to the relationship between consonants and vowels in speech. Furthermore, string vibrations result in integer harmonic components, which are then shaped through pickups. This suggests that the source-filter model [9], which approximates vocal fold vibrations as a periodic impulse train and filters the vocal tract characteristics, is also a valid approximation for electric bass. Promising acoustic features and analysis algorithms based on the source-filter model are expected to be applicable.

In this study, we propose the Attack-sustain label for annotating electric bass playing techniques (**Fig. 1**). The Attack-sustain label treats playing techniques that depend on changes in the attack phase as consonants and those that depend on the sustain phase as vowels. This label is provided as a temporally aligned sequence of playing technique

symbols, separate from MIDI, similar to phoneme labels in singing voice. This provides detailed annotation data on the temporal transitions of playing techniques, which can be useful for instrument sound synthesis and playing technique recognition. Additionally, by focusing on the acoustic similarity between electric bass and speech, it is possible to automate segmentation using high-precision phoneme alignment methods.

In our experiments, we aligned our Attack-sustain labels with acoustic signals. We constructed a new electric bass database and applied conventional alignment methods which are based on viterbi algorithm of hidden Markov model [10] and dynamic-time-warping (DTW) [4], DTW with timbre conversion based on a voice conversion (VC) [11]. Our results demonstrate that our Attack-sustain labels provide temporally accurate annotations of playing techniques.

2 Attack-sustain label

2.1 Label design

A naive annotation method of a technique to a note is an assignment of a single technique to a single note (hereinafter referred to as "note-wise"). For example, for a note played by plectrum picking, "plectrum" is assigned to that note. However, annotating a performance that combines multiple techniques, such as a muted string played with plectrum picking, requires multiple symbol sequences, complicating the annotation process.

We focus on the acoustic properties of the electric bass signal. Electric bass signals are generated by plucking the strings with a finger or pick. The strings collide with the pick/finger/fret depending on the playing technique, generating aperiodic noise. Then, depending on the playing technique (mute/harmonics/etc.), periodic string vibrations are generated and slowly decay. Focusing on this generative process suggests that the acoustic differences in playing techniques can be broadly classified into those that appear in the attack phase and those that appear in the sustain phase [8].

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 harmonic techniques, are distinguished. We assign "pause" to a silent segment such as a rest.

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, slide up, slide down

2.2 Automatic alignment method

Viterbi alignment of HMM Because controllable systems typically uses explicit temporal segmented data [4,12]. However, the manual annotation requires well-experienced

annotators in detecting segment boundaries. A common automatic method in speech processing is a Viterbi alignment based on hidden Markov models (HMMs) [10]. HMMs are trained using pairs of label sequences and acoustic features, and the Viterbi path is a temporal alignment of the technique label sequences [13]. The HMM perform robustly for performances that contain some disturbance such as noise and small fluctuation. However, because the HMM is based on switching stationary signal sources, it is difficult to model slowly decaying string vibration. The effects of its improvements such as hidden semi-Markov models [14] and trajectory HMMs [15] are also limited, because not only the playing technique, but also the pitch and duration of the notes vary depending on the musical context. In addition, the accuracy of data-driven approaches is highly dependent on the amount of data.

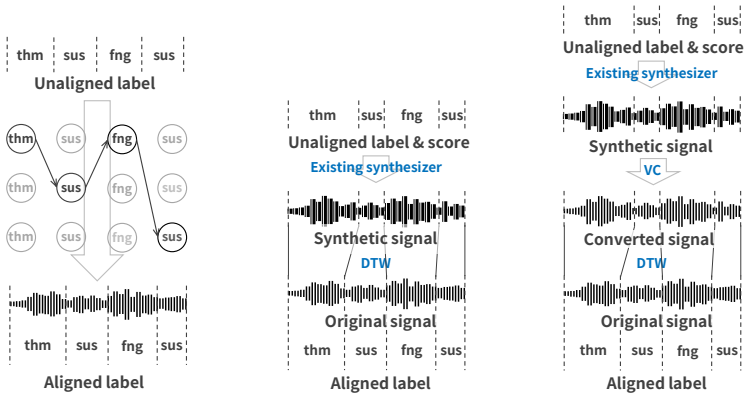
DTW Another method is synthesizing electric bass signals from the musical scores using existing synthesizers (e.g., sample concatenative synthesizer), and obtaining the alignment with the recorded signal by DTW [16]. Since the synthesizer generates a faithful performance to the musical score, the label's temporal offset can be obtained from the alignment of the synthetic and recorded sound.

DTW with timbre conversion Since the timbres differ between synthetic and recorded sound, this affects the alignment accuracy of the DTW. To reduce this problem, we utilize timbre conversion during the DTW using a VC technique. It has shown efficacy in singing voice alignment [11] and is also promising for electric bass with acoustic similarity to speech. First, the alignment of synthesized and recorded speech is obtained as described above. Next, using the aligned sound, a VC model (e.g., affine transformation [17] or Gaussian mixture model (GMM) [18]) is trained to transform the synthetic sound's timbre into the recorded one's timbre. Finally, the DTW takes 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. In addition, it is known that the DTW and timbre conversion can be sufficiently accurate in a single iteration [17].

3 Experimental evaluation

3.1 Dataset

A new electric bass sound database was constructed to evaluate the accuracy with respect to actual acoustic signals. The sounds used were 180 phrases of four bars of monophonic bass line (approximately 112 minutes), containing all techniques in the list (**Table 1**), and each with a various tempo between from 60 to 120 beat per minute (BPM). The label series before alignment was given manually. The note-wise label gave the attack label and sustain label pair as a single symbol. Finger picking, for example, is annotated as "fng-sus" for a single note. The electric bass used was a Fender custom shop 1962 Jazz Bass [19], the audio interface was an RME ADI-2 Pro FS R [20], and the performance was recorded by an experienced player.



(a) (b) (c)

Fig. 2. Overview of automatic alignment algorithms. Each figure shows (a) Viterbi alignment of HMM, (b) DTW and (c) DTW with timbre conversion.

3.2 Conditions

We apply the alignment algorithms to the proposed attack-sustain label and evaluate its accuracy. The most straightforward evaluation in comparing alignment methods is to calculate the error to ground truth. However, it is difficult to manually obtain ground truth for all the data. Therefore, we performed manual labeling on randomly selected pieces and calculated the mean absolute error (MAE) [21] on the rest of pieces for each attack and sustain technique.

In addition, for all data, we segmented acoustic features following the resulting alignment, and we calculated a separation metric (SM) R [11] defined as

$$R = \sum_{D} \frac{\sum_a \omega_a \left(\mu_a - \mu\right)^2}{\sum_a \omega_a {\sigma_a}^2}. \tag{1}$$

The subscript a indicates a technique label. μ_a and μ is the mean in the segment of technique label a and the global mean, respectively. σ_a is the standard deviation in the segment of technique label a. ω_a is the amount ratio of a: the number of frames in a segment divided by the total number of frames. These values are calculated from each dimension of D-dimensional acoustic features segmented following the resulting alignment. μ is the global mean calculated from the whole of database. When the resulting alignment can segment acoustic features for each label accurately, intra-technique standard deviation (i.e., σ_a) becomes smaller, and R becomes larger.

We first evaluated whether the Note-wise label or our attack-sustain label gives a more accurate alignment. The SM and MAE for the HMM-based alignment result were calculated for the two labels. To ensure fair conditions, Attack-sustain labels were compared to the start and end times of the Note-wise label, while Attack-sustain labels were compared to the start time of the Attack label and the end time of the Sustain label. The performance of the DTW-based method was omitted because it depends only on the acoustic signal.

We secondly compared alignment methods described in **Section 2.2** as follows.

- **HMM**: Viterbi alignment of the HMMs [10]
- **DTW**: DTW between synthetic and recorded sound [16].
- **DTW+AF**: DTW with Affin-transform-based timbre conversion [17].
- **DTW+GMM**: DTW with GMM-based timbre conversion [18].

10% of the dataset was manually annotated, and 20% was evaluated by SM and the remaining 70% was used to train the HMM. These subsets were randomly selected. The sound was recoreded at 48 kHz sampling/16-bit PCM and were downsampled to 16 kHz for acoustic feature extraction. Mel cepstrum was downsampled to 16 kHz with a window length of 1024 and a hop size of 5 ms. 24-dimensional mel-cepstral coefficients were used as acoustic features. The number of Gaussian mixtures was set to 4 for "HMM" and 16 for "DTW+GMM". For the DTW-based method, we used Standard Bass V2 [22] as a sample concatenative synthesizer. Each sample was manually labeled and aligned in advance. The cost of DTW was calculated as the mean squared error between the acoustic features.

3.3 Result and discussion

Table 2 lists the result of the label comparison. The alignment accuracy with our attack-sustain label became higher. This is because the note-wise HMM assumes a stationary signal for the steep acoustic change from attack to sustain. On the other hand, our attack-sustain label improved the accuracy by distinguishing between harmonic and non-harmonic states. However, there are still estimation errors in DTW-based methods. Focusing on the MAE, there were about 10 ms of errors in the time boundary of the technique. It is possible that noise from the release of the pressed strings interfered with the DTW path and was incorrectly estimated as the attack phase.

Table 3 lists the results. First, there are no large differences of SM and MAE in attack technique. This is considered that aperiodic components were dominant in the attack segment, and the acoustic features varied steeply. On the other hand, "HMM" scored the worst in sustain technique, and " DTW", "DTW+AF", and "DTW+GMM" scored better in that order. This indicated that the DTW-based method worked robustly because the synthesizer replaced the modeling of non-stationary decay of the string vibration. In addition, the affine-transformation-based conversion is equivalent to the single-component GMM. "DTW+GMM" therefore enhanced the performance because of the higher accuracy of the timbre conversion.

In this experiment, both DTW and HMM were performed on a single player's performance. Different players perform different types of electric bass and in different styles, resulting in different acoustic characteristics. Thus, the accuracy may vary depending on a performer. This difference correspond to speaker differences in speech. Parallel voice conversion also uses the DTW between different speakers and performs high quality conversion. Since the electric bass signal exhibits similar acoustic characteristics to speech, it is expected to produce similar results in the signals of different performers.

Table 2. Comparison of alignment accuracy between note-wise label (Note) and our attack-sustain labels (AS). Separation metric (SM) and mean absolute error (MAE) from the ground truth of alignment methods. Higher SM value and lower MAE indicate more accurate.

Method	SM		MAE [ms]	
	AS (ours)	Note	AS (ours)	Note
HMM	**35.73**	20.01	**24.15**	32.00

Table 3. Accuracy comparison of automatic alignment methods. Separation metric (SM) and mean absolute error (MAE) from the ground truth of alignment methods. Higher SM value and lower MAE indicate more accurate.

Method	SM		MAE [ms]	
	Attack	Sustain	Attack	Sustain
HMM	11.98	40.03	35.14	20.06
DTW	13.07	60.03	21.23	11.69
DTW+AF	12.31	67.79	19.45	**11.28**
DTW+GMM	**13.98**	**68.15**	**19.24**	12.42

4 Conclusion

This paper proposed the attack-sustain 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 electric bass signals.

We investigated automatic labeling method to align the label sequence to the acoustic signal. The experimental evaluation demonstrated that 1) our attack-sustain label is effective for accurate alignment 2) the method based on DTW with timbre conversion achieved better accuracy. In our future work, we will increase the data and train DNN-based synthesis models using our the label and acoustic signal pairs. Moreover, constructed sound database will be available in the public domain.

References

1. C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc ICLR*, 2019.
2. Valentin Emiya, Roland Badeau, and Bertrand David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.
3. M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, Paris, France, Oct. 2002, vol. 2, pp. 287–288.
4. Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep performer: Score-to-audio music performance synthesis," in *Proc. ICASSP*, 2022, pp. 951–955.
5. Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, and Elaine Chew, "Adaptive scattering transforms for playing technique recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1407–1421, 2022.
6. Toontrack, "Ezbass," `https://www.toontrack.com/product/ezbass/`.
7. IK Multimedia, "Modo bass 2," `https://www.ikmultimedia.com/products/modobass2/`.
8. Jakob Abeßer, Hanna Lukashevich, and Gerald Schuller, "Feature-based extraction of plucking and expression styles of the electric bass guitar," in *Proc. ICASSP*, 2010, pp. 2290–2293.
9. Gunnar Fant, *Acoustic Theory of Speech Production*, De Gruyter Mouton, Berlin, Boston, 1971.
10. F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 1, no. 4, pp. 357–370, 1993.
11. J. Koguchi, S. Takamichi, and M. Morise, "PJS: phoneme-balanced japanese singing-voice corpus," in *Proc. APSIPA ASC*, 2020, pp. 487–491.
12. E. Cooper, X. Wang, and J. Yamagishi, "Text-to-Speech Synthesis Techniques for MIDI-to-Audio Synthesis," in *Proc. SSW 11*, 2021, pp. 130–135.
13. K. Nakamura, K. Oura, Y. Nankaku, and K. Tokuda, "HMM-based singing voice synthesis and its application to Japanese and English," in *Proc. ICASSP*, 2014, pp. 265–269.
14. Shun-Zheng Yu, "Hidden semi-markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, 2010.
15. H. Zen, K. Tokuda, and T. Kitamura, "A viterbi algorithm for a trajectory model derived from hmm with explicit relationship between static and dynamic features," in *Proc. ICASSP*, 2004, vol. 1, pp. I–837.
16. N. Hu, R.B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. WASPAA*, 2003, pp. 185–188.
17. G. Kotani, H. Suda, D. Saito, and N. Minematsu, "Experimental investigation on the efficacy of affine-dtw in the quality of voice conversion," in *Proc. APSIPA ASC*, 2019, pp. 119–124.
18. T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.
19. Fender Custom Shop, "1962 jazz bass," https://www.fendercustomshop.com/basses/jazz-bass/.
20. RME, "ADI-2 Pro FS R," `https://www.rme-audio.de/adi-2-pro-fs-be.html`.
21. A. Cont, D. Schwarz, N. Schnell, and C. Raphael, "Evaluation of real-time audio-to-score alignment," in *Proc. ISMIR*, 2007, pp. 315–316.
22. Purple_Shikibu_, "Standard Bass V2," `https://unreal-instruments.wixsite.com/unreal-instruments/standard-bass`.