

# 特別研究論文

## (査読済み)

### 研究題目

深層学習に基づく音響特徴量からの振幅スペクトログラム予測とその評価

提出年月日	2024年1月31日
氏名	川口 翔也
主査	北村 大地 講師
副査	村上 幸一 准教授
副査	柿元 健 准教授

香川高等専門学校  
専攻科  
創造工学専攻







# Amplitude Spectrogram Prediction from Acoustic Features Based on Deep Learning and Its Evaluation

Shoya Kawaguchi

Department of Electrical and Computer Engineering  
National Institute of Technology, Kagawa College

## Abstract

The technique of extracting timbre features from observed instrumental signals can be applied to various problems such as instrument identification, music retrieval, timbre conversion, and source separation. The mel-frequency cepstrum coefficient (MFCC) is an analytical feature that well represents the timbre of instrumental sounds. In recent years, deep neural networks (DNNs) based on generative models have been widely studied for feature extraction, instrumental sound generation, and timbre conversion. Since DNN-based generative models are a relatively new technology, it is still desirable to propose a better technique for timbre conversion. In this thesis, I propose a new timbre conversion algorithm using a variational auto-encoder (VAE). VAE is a DNN that can learn latent features of input data as unsupervised learning, and its latent features have interpretability as the relationship between classes. Thus, by training the latent features of MFCCs of multiple instruments with VAE, we can generate a new acoustic signal with timbre from arbitrary latent features. The proposed system using VAE generates acoustic signals from three features, MFCC generated from VAE, pitch, and volume, through a decoder. However, the proposed system assumes that the amplitude spectrogram can be predicted from the three features, pitch, MFCC, and time-varying gains, and it is not clear whether such a generation is feasible. To cope with this problem, in this thesis, I investigate DNN-based spectrogram prediction with the three features. This is a subsystem necessary to realize the above-mentioned proposed system. In particular, three types of DNNs are investigated using two objective evaluation metrics to find the optimal network architecture. Since music is related to human sensibilities, the three types of DNNs are also evaluated in a subjective evaluation experiment. From the results, both objective and subjective evaluations showed that the bidirectional recurrent DNN achieves the highest prediction accuracy for general harmonic instruments. These results show that the DNN has satisfactory capability of amplitude spectrogram prediction. On the basis of this finding, I conduct to implement the above-mentioned proposed system. First, the latent features of various musical instruments are trained using VAE with MFCCs. Next, a new acoustic signal that has multiple instruments features is generated by the proposed system using the MFCCs. Finally, the accuracy of the generated acoustic signal is evaluated. From the evaluation result, it is confirmed that the timbre transition of output signals can be achieved by modifying the latent features in the pre-trained MFCC-based latent space of the VAE. These results suggest that the proposed system can generate the sound of musical instruments with new timbres.

**Keywords:** deep learning, mel-frequency cepstrum coefficient, timbre, conversion



## 概要

観測された楽器音信号に対して音色の特徴量を抽出する技術は、楽器識別、音楽検索、音色変換、音源分離等の様々な問題に応用できる。楽器音の音色をよく表現する解析的な特徴量にはメル周波数ケプストラム係数 (mel-frequency cepstrum coefficient: MFCC) があるが、近年は生成モデル系の深層ニューラルネットワーク (deep neural network: DNN) に基づく特徴量抽出、楽器音生成、音色変換等が広く研究されている。生成モデル DNN は比較的新しい技術であり、音色変換という目的においては現在もより良い技術の提案が望まれる状況にある。本論文では、生成モデル系 DNN の中でも変分自己符号化器 (variational auto-encoder: VAE) を用いた新しい音色変換アルゴリズムの構築を目指す。VAE は入力データの潜在的な特徴量を教師無しで学習できる DNN であり、潜在特徴量に複数クラスの相対関係を表す構造を導入することで、一定の解釈性を持たせることができる。従って、複数楽器音の MFCC の潜在特徴量を VAE で学習することで、任意の潜在特徴量から音色を持つ音響信号を生成できると考えられる。VAE を用いた提案システムでは、VAE から生成した MFCC、音高、及び音量の3つの特徴量から何らかのデコーダを通して音響信号を生成する。しかしながら、この提案システムは音高、MFCC、及び音量の3つの特徴量から音響信号の振幅スペクトログラムを生成できることを仮定しており、そのような生成が実現可能かどうかは不明である。そこで本論文では、前述の提案システムの実現に必要な部分システムとして、音高、MFCC、及び音量から振幅スペクトログラムを予測する手法について検討する。特に、3種類のネットワーク構造を比較し、どのようなネットワーク構造が高精度な振幅スペクトログラムの予測に効果的か、2種類の客観評価指標を用いて調査する。また、音楽は人間の感性と関連するものであるため、被験者を用いた主観評価実験による調査も行う。結果として、客観評価及び主観評価の両方において、一般的な調波構造を持つ楽器に対して長短期記憶ユニットを用いた双方向再帰型 DNN の予測精度が最も高いことを示した。これらの結果は、DNN が振幅スペクトログラムを予測する十分な能力を持っていることを示している。この結論に基づき、前述の提案システムの実装を行う。初めに、複数楽器音の MFCC の潜在特徴量を VAE で学習する。そして、提案システムを用いて複数楽器音の MFCC を併せ持つ音響信号の生成を行い、その性能を評価する。結果として、潜在空間上で推移させた音色の特徴量から時間波形を生成すると、音色が徐々に推移することを実験的に確認した。この結果は提案システムが新しい音色を持つ楽器音を生成する能力があることを示唆している。



# 目次

第 1 章	緒言	1
1.1	本論文の背景	1
1.2	本論文の目的	3
1.3	本論文の構成	5
第 2 章	要素技術及び類似研究	6
2.1	まえがき	6
2.2	STFT	6
2.3	MFCC	8
2.4	MLP 型 DNN	11
2.5	既存の類似研究	13
2.5.1	DDSP	13
2.5.2	知覚的メトリクスに基づく正規化付き VAE	16
2.5.3	VAE を用いた音色・音高の分離表現の学習	17
2.6	本章のまとめ	18
第 3 章	提案手法	19
3.1	まえがき	19
3.2	提案音生成システム全体の説明	19
3.3	提案音生成システムにおける問題	22
3.4	DNN に基づくデコーダ	23
3.4.1	MLP 型 DNN	23
3.4.2	BiLSTM 型 DNN	24
3.4.3	BiGRU 型 DNN	27
3.5	本章のまとめ	29
第 4 章	振幅スペクトログラム予測実験	30
4.1	まえがき	30
4.2	実験条件	30
4.3	実験結果	33
4.4	振幅相対誤差及び MFCC 相対誤差に基づく実験評価	35

4.5	予測楽器音の主観的評価実験 . . . . .	37
4.6	本章のまとめ . . . . .	40
第 5 章	MFCC 及び音響信号生成実験	43
5.1	まえがき . . . . .	43
5.2	実験条件 . . . . .	43
5.3	実験結果 . . . . .	44
5.4	本章のまとめ . . . . .	47
第 6 章	結言	49
	謝辞	50
	参考文献	51
付録 A	振幅スペクトラムの予測結果	55

# 第 1 章

## 緒言

### 1.1 本論文の背景

一般的な音響信号を構成する 3 要素は音高（ピッチ）、音色、及び音量である。これらは対象の音がどのような音であるかを表現するために用いられる。例えば、楽譜は楽器の種類（音色）と音高と音符（音量の時間的な変化）で音楽を表したものである。また、音声や音楽を対象とした様々な信号処理においても、これらの 3 要素は重要な特徴量である。音響信号処理において最も基本的な音響特徴量は、音響信号を離散 Fourier 変換（discrete Fourier transform: DFT）して得られる振幅（又はパワー）スペクトルである。また、スペクトルの時間的な変化を見るために、信号を短時間区間に区切って DFT を適用する短時間 Fourier 変換（short-time Fourier transform: STFT）も頻繁に適用され、音響信号処理全般で用いられる重要な時間周波数領域の特徴量といえる。振幅（又はパワー）スペクトルには音高、音色、及び音量の情報が強く含まれている。具体的には、振幅スペクトルの最も低い周波数にあるピークが基本周波数成分であり、一般的に音響信号の音高に対応する。また、スペクトルの概形や包絡（基本周波数やその倍音成分の強度バランス）が人間の認知する音色に対応し、スペクトル全体のエネルギーが音量に対応する。音高、音色、及び音量の中で、楽器音を対象とした信号処理分野では、音色を表す特徴量に関してこれまで様々なものが提案されてきた。例えば音声信号の音色を表す最も代表的な特徴量は線形予測符号（linear predictive coding: LPC）[1] であり、スペクトルの包絡を全極フィルタで近似した際の係数である。LPC は人間の音声の生成原理に基づき設計されたものであり、音声通信技術における基盤技術となっている。一方、音声信号だけでなく楽器音信号や音楽信号の音色を表す特徴量にメル周波数ケプストラム係数（mel-frequency cepstrum coefficient: MFCC）[2] がある。MFCC は、メル周波数と呼ばれる人間の聴覚特性を考慮した領域でのスペクトルの概形を表現した特徴量であり、通常のスペクトルよりも低次元な空間で音色をよく表現する性質を持つ。

音響信号の音色の特徴量は、様々な音響信号処理に活用することができる。音色を表す特徴量である MFCC を活用した手法の例として、音響信号を入力とした楽器の種類と同定 [3, 4, 5]、様々な楽器音が含まれる音楽信号中の音色抽出 [6, 7] と楽器構成推定 [8]、ユーザが好む音楽のレコメンデーション [9]、種別の異なる楽器音の類似性の評価 [10]、音色に基づくクラ

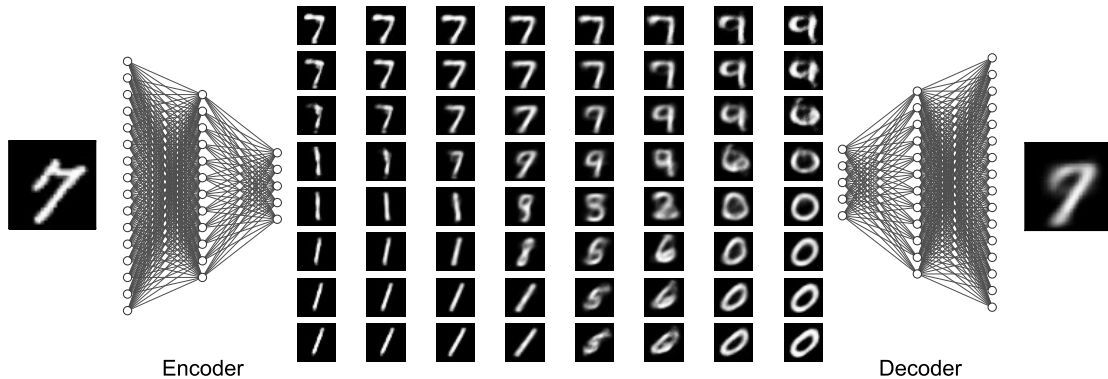


Fig.1.1. The latent space of VAE trained with images of handwritten numbers.

スタリングによる音源分離 [11, 12] 等が挙げられる。近年では、深層ニューラルネットワーク (deep neural network: DNN) に基づく非線形かつ強力な特徴量抽出及び楽器音生成が広く研究されている。例えば、Differentiable Digital Signal Processing (DDSP) [13] では、正弦波とノイズの合成パラメータを学習して楽器音信号を作り出す DNN が提案されており、DNN に基づく楽器音信号生成の成功例の一つといえる。また、敵対的生成ネットワーク [14, 15]、フローベース生成モデル [16, 17]、及び変分自己符号化器 (variational auto-encoder: VAE) [18, 19] 等の生成モデル系 DNN を用いて楽器音の音色特徴量を抽出する手法もいくつか提案されている。この中で VAE は他の生成モデルに比べて、安定した学習が可能であり、潜在空間を低次元で表現することができる [20]。さらに、潜在空間の解釈性の高く、比較的扱いやすいモデルであることから、VAE を用いた楽器音の解析や生成 [21, 22, 23, 24] は盛んに研究されており、より良い技術を模索しているのが現状である。本論文でも、より高精度・高品質に楽器音の音色特徴量の抽出及び音色特徴量から音響信号の生成を行う手法を目指し、VAE に基づく楽器音の生成及び変換について取り組む。

VAE とは、入力データの潜在的な特徴量を教師無しで学習できる DNN である。通常の自己符号化器のように、入力と同じデータを出力するような低次元の潜在空間を学習する DNN であるが、潜在空間に複数クラスの相対関係を表す確率分布の構造を導入している。これによって、潜在空間や潜在特徴量そのものに対して一定の解釈性を持たせることができる。Fig. 1.1 は手書き数字画像を VAE で学習した例を表している。入力の手書き数字画像を潜在空間に次元圧縮するエンコーダと、潜在空間の特徴量から入力を近似した手書き数字画像を出力するデコーダから成る。この時、学習済みの VAE の潜在空間上の特徴量を変化させて (あるいは乱数から潜在変数を生成して) デコーダに通すことで、複数クラス (この例では手書き数字の種類) の間を補完するような画像 (例えば「7」と「9」の間の画像等) が生成可能である。従って、複数楽器音の MFCC の潜在空間を VAE で学習することができれば、Fig. 1.1 で生成される新しい手書き数字画像と同じように、各楽器音の特徴量を併せ持つ新しい楽器音信号を生成できると考えられる。このような新しい楽器音の生成及び解析は、現代の音楽文化の更なる発展や興隆につながることを期待される。例えば、既存の楽器音や音楽の音色を一般ユーザが自由に編集できる高度なリミックスが実現できるならば、これまでには無い能動的

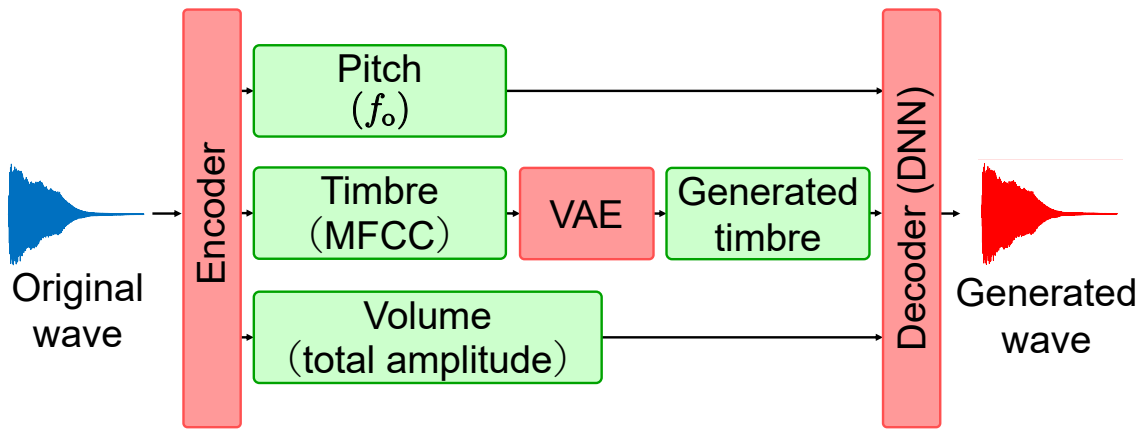


Fig.1.2. Overview of the proposed sound generation system.

な音楽鑑賞が新しい文化として定着する可能性がある。また、VAE で得られる音色特徴量の潜在空間上で楽器音の音色を解析できれば、より良い楽器の設計製作や高度な演奏技術習得のためのツールとしても役立てられる可能性がある。さらに、音響信号と空間的な伝搬と音色変化の関係性を潜在空間に同時に埋め込めることができれば、三次元ヴァーチャル空間上での音源及びユーザ位置に依存する音色変化を自動的に推定・生成できる等の発展的な用途も考えられる。なお、VAE を用いた楽器音生成の類似研究としていくつかの手法が提案されている [21, 22, 23, 24] が、2章で紹介する通り、本論文ではこれらの理論の良い点を融合した新しい音生成システムの提案を目指している。

## 1.2 本論文の目的

前節で述べた本論文で提案する音生成システムの概要を Fig. 1.2 に示す。以後、このシステムを提案音生成システムと呼ぶ。入力となる音響信号の時間波形 (Fig. 1.2 における Original wave) から、音高 (Fig. 1.2 における Pitch)、音色 (Fig. 1.2 における Timbre)、及び音量 (Fig. 1.2 における Volume) の3つの特徴量をエンコーダで抽出する。音高は音響信号の基本周波数  $f_0$ 、音色は MFCC、及び音量は total amplitude (音量の時間変化) を用いる。さらに、抽出された音色 (MFCC) のみを VAE に入力し、VAE の出力として新たに生成された音色 (Fig. 1.2 における Generated timbre) を得る。このようにして得られた音高、VAE で生成された音色、及び音量の3つの特徴量をデコーダに入力し、新しい音響信号を生成・出力する。提案音生成システムでは、音響信号の MFCC のみが VAE でモデル化されているため、音色のみを Fig. 1.1 に示す手書き数字画像のように新しいものに変換し、その MFCC を持つ音響信号を生成することを目的としている。例えば、Fig. 1.3 に示すように乱数により生成された音色の潜在変数ベクトル (Fig. 1.3 における Random latent vector) を潜在変数とみなして学習済み VAE のデコーダに入力することで新しい音色 (Fig. 1.3 における Generated timbre) を生成できる。その後は同様に、音高、新しい音色、及び音量変化の3つの特徴量を

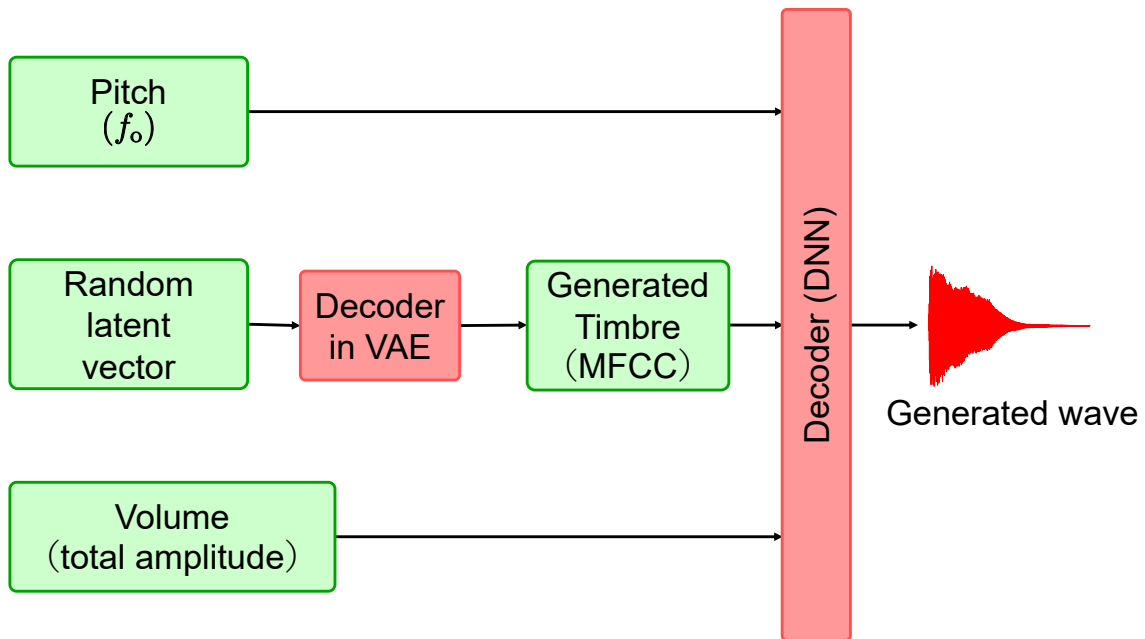


Fig.1.3. New sound generation using the proposed system with a random vector input.

デコーダに入力し、新しい音色情報を持つ音響信号を出力する。

しかしながら、この提案音生成システムの実装には大きな問題がある。具体的には、音高、音色 (MFCC)、及び音量の3つの特徴量から音響信号を生成するようなデコーダが存在せず、学習データから新たに構築する必要があるという点である。この原因は、MFCCがスペクトルよりも低次元な特徴量であることに起因しており、圧縮された次元を復元する非線形な処理が必要となるためである。そこで本論文では、提案音生成システムを構築する予備段階として、音高、音色 (MFCC)、及び音量の3つの特徴量から振幅スペクトログラムを出力するデコーダをDNNで学習する。このデコーダには、最も基本的なDNNである多層パーセプトロン (multilayer perceptron: MLP)、再帰型ニューラルネットワーク (recurrent neural network: RNN) の1つである長・短期記憶 (long-short term memory: LSTM) ユニット [26] を用いた双方向再帰型ニューラルネットワーク (bidirectional RNN using LSTM: BiLSTM)、及びRNNの1つであるゲート付き回帰型ユニット (gated recurrent unit: GRU) [27] を用いた双方向再帰型ニューラルネットワーク (bidirectional RNN using GRU: BiGRU) の3種類のDNNを用い、どのアーキテクチャが高い精度で振幅スペクトログラムの予測ができるか、2種類の客観的評価指標を用いて実験的に調査する。また、音楽は人間の感性と関連するものであるため、被験者を用いた主観的評価実験による調査も行う。主観的評価実験では、GUIアプリケーション及び実験用音源を作成し、被験者15名を対象に実験を行う。被験者には、入力音響信号に対し、予測された音響信号がどの程度劣化して聞こえるか、5段階評価を用いて評価する。

さらに、作成したDNNデコーダ及びVAEを用いた提案音生成システムを作成し、提案音生成システムの音響信号の生成精度を調査する。調査方法として、入力の音響信号を提案音生



成システムを通して再生成し、音響信号及び振幅スペクトログラムの再現精度を評価する。また、潜在空間上での2つの楽器音の音色推移を出力し、複数楽器音間の音色を併せ持つ音響信号を生成可能かを調査する。

### 1.3 本論文の構成

まず2章では、提案音生成システムの詳細を説明するにあたって必要となる基礎技術のSTFT, MFCC, 及びMLP型DNNについて説明する。また、提案音生成システムと類似している既存研究を3種類紹介し、提案音生成システムとの類似点, 相違点, 及び目的の違いについて述べる。3章では、提案音生成システムのFig. 1.2におけるエンコーダ及びデコーダの詳細と音高, 音色, 及び音量の3つの特徴量の抽出方法を説明する。さらに、提案音生成システムを実装する上での問題の説明とその解決に用いるMLP型DNN, BiLSTM型DNN, 及びBiGRU型DNNを説明する。また、MFCCをVAEで学習し、新たな音色を生成する方法を説明する。4章では、3章で説明した3種類のアーキテクチャを用いて、音高, 音色, 及び音量の3つの特徴量から振幅スペクトログラムを予測する実験を行い、予測精度を客観的評価法及び主観的評価法を用いて評価し、考察を行う。5章では、MFCCをVAEで学習し、入力した音響信号の再現及び各楽器音の特徴量を併せ持つ新しい音響信号の生成を行い、評価及び考察を行う。最後に6章では、本論文全体の結論を述べる。

## 第 2 章

# 要素技術及び類似研究

### 2.1 まえがき

本章では、前章で述べた提案音生成システムを構成する要素技術として、音響信号の時間周波数領域への変換である STFT、楽器音の音色特徴量を示す MFCC [2]、及び深層学習における最も基本的な非線形変換である MLP について、それぞれ 2.2 節、2.3 節、及び 2.4 節で詳細を述べる。また、2.5 節で、本論文が最終的に目指す提案音生成システムに類似した研究を 3 つ紹介し、類似点や違い等を述べる。2.6 節で本章をまとめる。

### 2.2 STFT

STFT は、音響信号の時間的に変化するスペクトルを、時間周波数領域と呼ばれる二次元の特徴量空間で表現するための変換手法である。この変換を Fig. 2.1 に示す。STFT では、音響信号の時間波形を短時間区間に分割し、窓関数を乗じたうえで周波数領域へと変換する。この時の短時間区間長 (分析窓関数長) 及び短時間区間のシフト長をそれぞれ  $Q$  及び  $\tau$  としたとき、時間領域の信号  $(z(l))_{l=1}^L$  の  $j$  番目の短時間区間 (時間フレーム) の信号  $z^{(j)}$  は次式で表される。

$$\begin{aligned} z^{(j)} &= [z((j-1)\tau+1), z((j-1)\tau+2), \dots, z((j-1)\tau+Q)]^T \\ &= [z^{(j)}(1), z^{(j)}(2), \dots, z^{(j)}(q), \dots, z^{(j)}(Q)]^T \in \mathbb{R}^Q \end{aligned} \quad (2.1)$$

ここで、 $\cdot^T$  は転置を表し、 $l = 1, 2, \dots, L$ 、 $j = 1, 2, \dots, J$ 、及び  $q = 1, 2, \dots, Q$  は、それぞれ離散時間サンプル、時間フレーム、及び時間フレーム内の離散時間サンプルのインデックスを示す。また、セグメント数  $J$  は次式によって与えられる。

$$J = \frac{L}{\tau} \quad (2.2)$$

ただし、信号長  $L$  はセグメント数  $J$  が整数となるように各時間フレームの信号の両端にゼロを挿入する処理 (ゼロパディング) が施されている。式 (2.1) で定義される時間フレームの信号

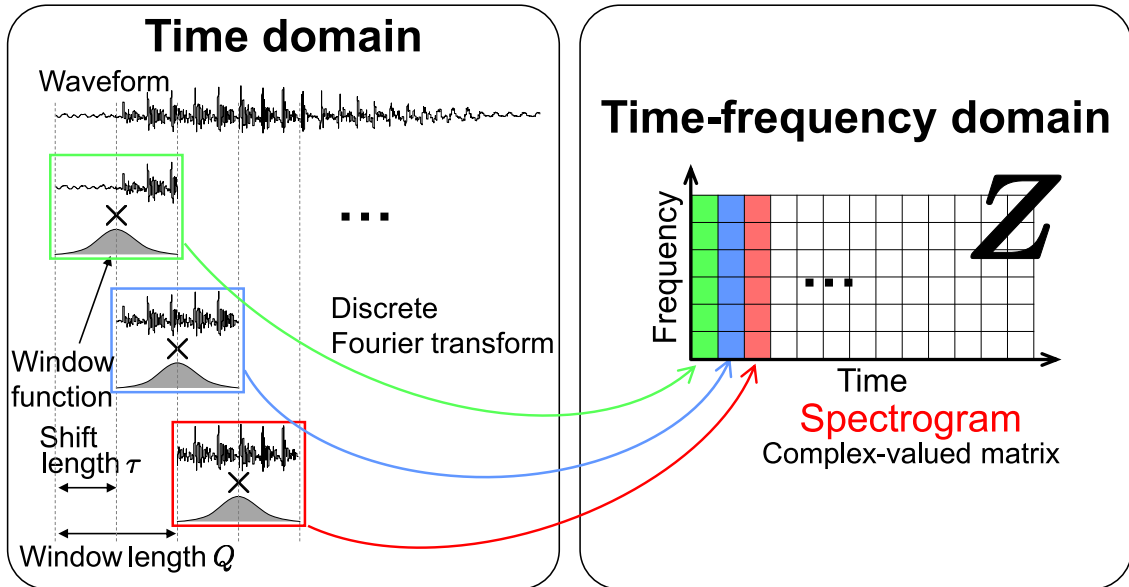


Fig.2.1. Mechanism of STFT.

を全ての  $j$  についてまとめた全時間フレームの信号を  $\mathbf{z} = [z^{(1)} z^{(2)} \dots z^{(J)}] \in \mathbb{R}^{Q \times J}$  と表記すると、STFT の処理は次式のように表される。

$$\mathbf{Z} = \text{STFT}_{\omega}(\mathbf{z}) \in \mathbb{C}^{I \times J} \quad (2.3)$$

ここで、 $\omega = [\omega(1), \omega(2), \dots, \omega(Q)]^T \in \mathbb{R}^Q$  は STFT で用いる窓関数であり、 $\mathbf{Z}$  は (複素) スペクトログラムと呼ばれる時間周波数行列である。スペクトログラム  $\mathbf{Z}$  の  $(i, j)$  番目の要素は次式で表される。

$$z_{ij} = \sum_{q=1}^Q \omega(q) z^{(j)}(q) \exp \left\{ \frac{-i2\pi(q-1)(i-1)}{F} \right\} \quad (2.4)$$

ここで、 $z_{ij}$  は  $\mathbf{Z}$  の要素、 $F$  は  $\lfloor \frac{F}{2} \rfloor + 1 = I$  を満たす整数 ( $\lfloor \cdot \rfloor$  は床関数)、 $i = 1, 2, \dots, I$  は周波数ビンのインデックス、 $i$  は虚数単位を示している。このように、時間領域の信号は一定幅の短時間毎に分析窓関数を乗じて DFT を行うことで、時間と周波数の 2 次元複素行列であるスペクトログラム  $\mathbf{Z}$  で表すことができる。また、音響信号処理では各時間周波数成分の大きさのみを取り扱うことも多い。その場合は、複素スペクトログラム  $\mathbf{Z}$  の各要素に関して絶対値を取った振幅スペクトログラム  $|\mathbf{Z}| \in \mathbb{R}_{\geq 0}^{I \times J}$  や、絶対値の 2 乗を取ったパワースペクトログラム  $|\mathbf{Z}|^2 \in \mathbb{R}_{\geq 0}^{I \times J}$  を処理の対象とする。ここで、行列に対する絶対値記号及びドット付き指数乗はそれぞれ要素毎の絶対値及び要素毎の指数乗を表す。

例として、Fig. 2.2(a) 及び (b) にそれぞれ音楽信号及び音声信号のパワースペクトログラムを示す。Fig. 2.2(a) 及び (b) における色の変化は、黒色に近づくほどパワーが小さく、黄色に近づくほど大きいことを表している。

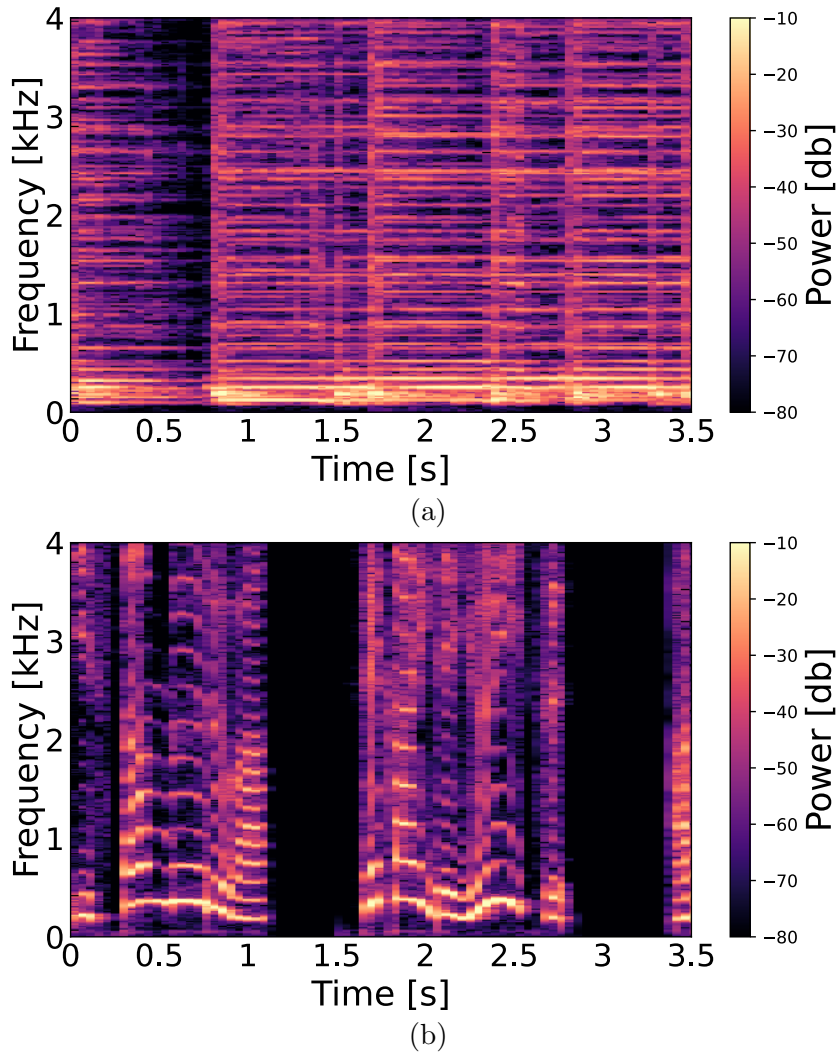


Fig.2.2. Power spectrograms of (a) music and (b) speech signals.

### 2.3 MFCC

MFCC [2] とは，時間周波数領域で表現された音声及び楽器音等の音色の特徴量である．音色のみの特徴量を可能な限り抽出しているため，音高や音量はほとんど反映されない特徴量である．

MFCC を説明する前に，まずメル周波数について説明する．メル周波数とは，1000 Hz の純音に対応する音高尺度を 1000 mel（メル周波数と呼ぶ）と定義し，これを基準として人間が知覚する音高を線形な一次元軸に対応付けた尺度である．例えば，人間は 2000 mel のメル周波数の音を「1000 mel の 2 倍の高さ」と知覚するが，これは周波数上では 2000 Hz ではなくおよそ 3500 Hz 程度となることが知られている．周波数  $f$  Hz とメル周波数  $m$  mel の対応

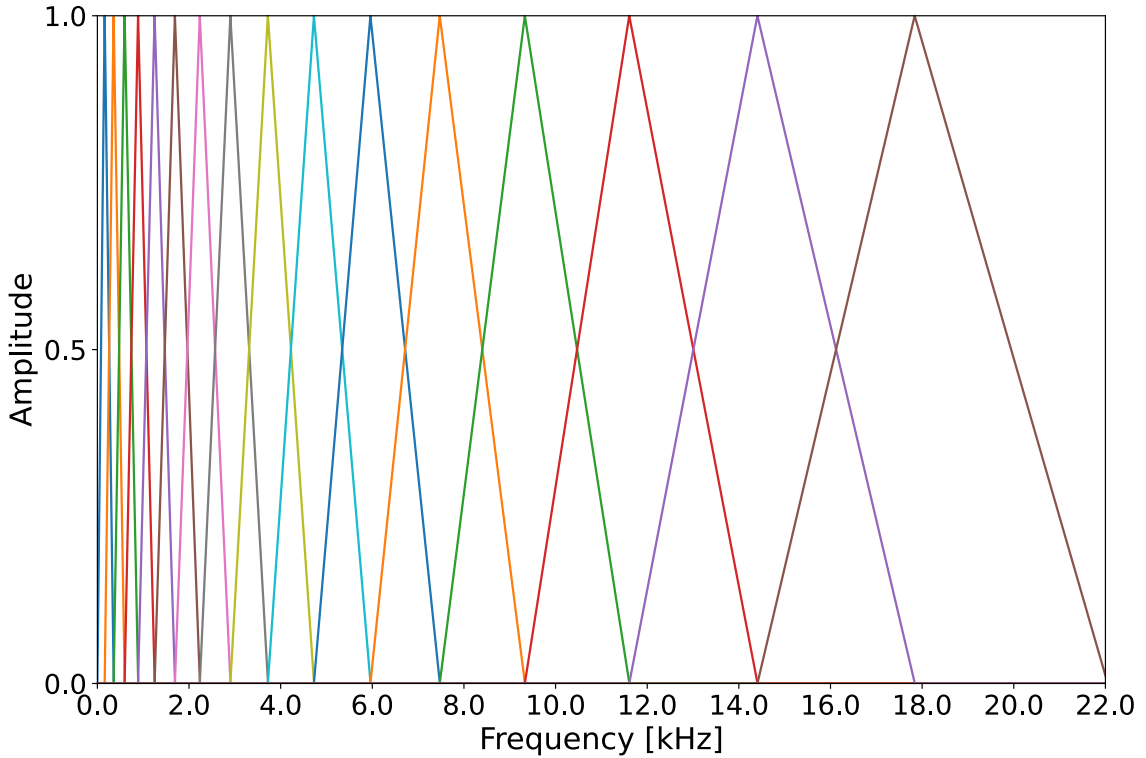


Fig.2.3. Mel-filter bank ( $f_{\min} = 0$  Hz,  $f_{\max} = 22050$  Hz,  $K = 16$ ).

関係は次式で定義される [28].

$$m = 1127 \ln \left( 1 + \frac{f}{700} \right) \quad (2.5)$$

$$= 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \quad (2.6)$$

このメル周波数軸上で等間隔に三角状のバンドパスフィルタを複数個 ( $K$  個) 構成したものをメルフィルタバンクと呼ぶ. Fig. 2.3 にバンドパスフィルタ数  $K = 16$ , 下限周波数  $f_{\min} = 0$  Hz, 及び上限周波数  $f_{\max} = 22050$  Hz の場合のメルフィルタバンクを示している.

STFT で得られる短時間区間のパワースペクトル (パワースペクトログラム  $|\mathbf{Z}|^2$  の各列ベクトル) に対して,  $K$  個のメルフィルタをそれぞれ畳み込むことでメルスペクトルと呼ばれる  $K$  次元の音色の特徴量に変換される. パワースペクトログラム  $|\mathbf{Z}|^2$  の全ての列ベクトルにメルフィルタバンクを畳み込みメルスペクトルに変換したものをメルスペクトログラム  $\mathbf{P}$  と呼び, この変換を次式で表す.

$$\mathbf{P} = \text{MelFiltering}(|\mathbf{Z}|^2) \in \mathbb{R}^{K \times J} \quad (2.7)$$

ここで, 周波数軸の次元 (周波数ビン数) が  $I$  から  $K$  に圧縮されている点に注意する.

MFCC は, メルスペクトログラム  $\mathbf{P}$  の各列ベクトル (各メルスペクトル) に離散コサイン変換 (discrete cosine transform: DCT) を適用して得られる短時間区間毎の実数係数である.

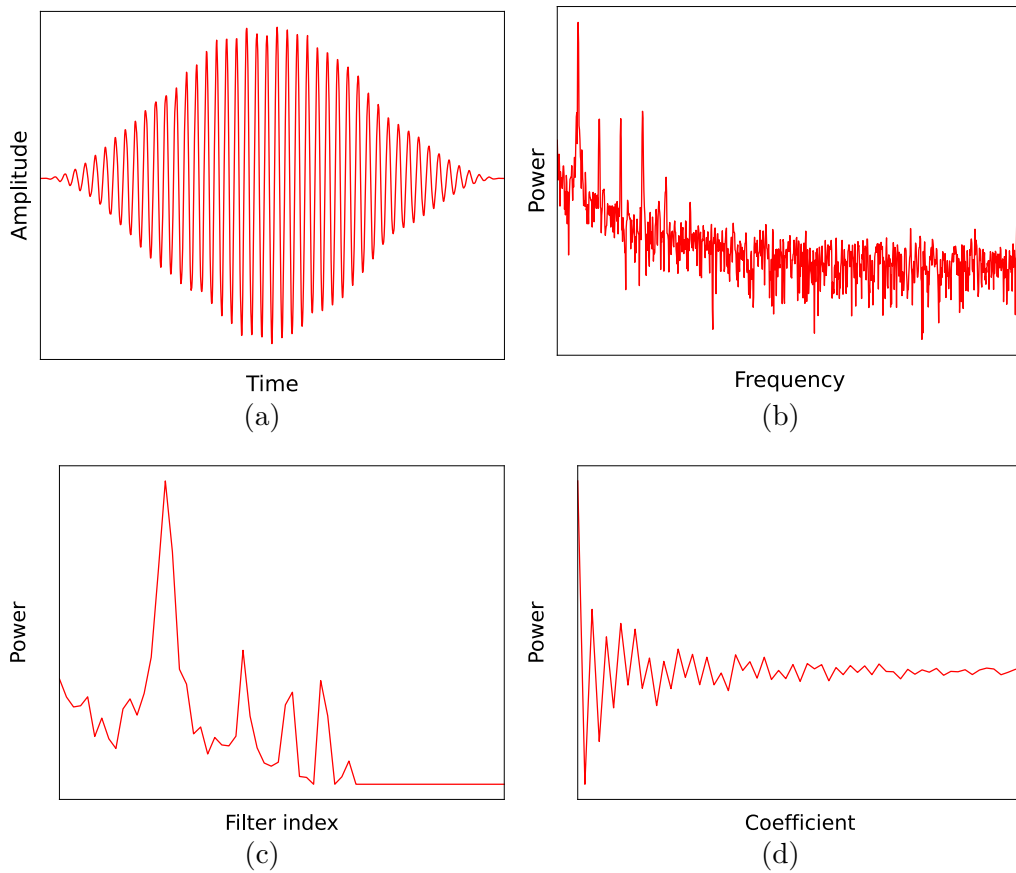


Fig.2.4. (a) windowed short-time signal of piano B5 sound, (b) power spectrum, (c) mel-frequency spectrum, and (d) MFCC.

MFCC を  $C$  と表記し、この処理を次式で表す。

$$C = \text{DCT}(P) \in \mathbb{R}^{K \times J} \quad (2.8)$$

以上の変換により得られる MFCC は、音高や音量の影響を可能な限り排除した音色に関する  $K$  次元の特徴量ベクトルとなる。短時間区間信号から MFCC までの変換の過程を Fig. 2.4 に示す。この図からも分かるように、MFCC はパワースペクトルの包絡を良く表現した特徴量であり、音色に関する情報を強く保持している。さらに、Figs. 2.5 及び 2.6 にピアノ及びギターの上昇音階のパワースペクトログラム  $|Z|^2$  と MFCC  $C$  をそれぞれ示す。ピアノとギターは音色が異なるため MFCC が大きく異なるが、音量の減衰に応じた MFCC の変化はほとんど無く、音高に対する MFCC の変化についても、低次 MFCC (Figs. 2.5 及び 2.6 におけるグラフ下部の濃淡) だけを見ればほとんど一定となっている。

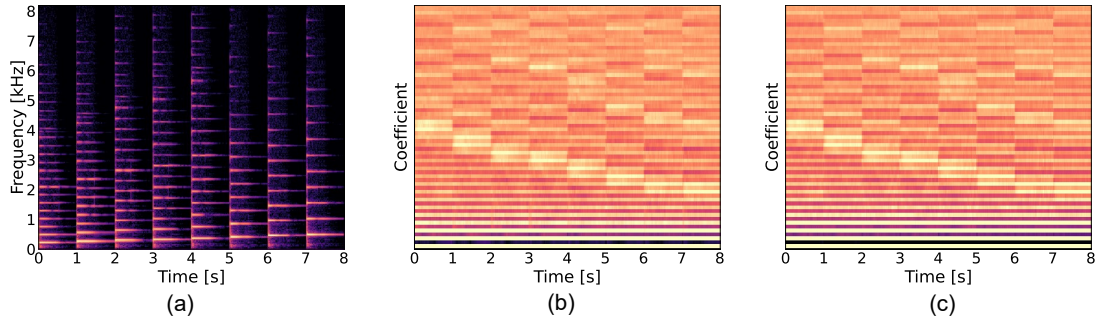


Fig.2.5. (a) power spectrogram of ascending notes played by piano sound, (b) MFCC, and (c) normalized MFCC.

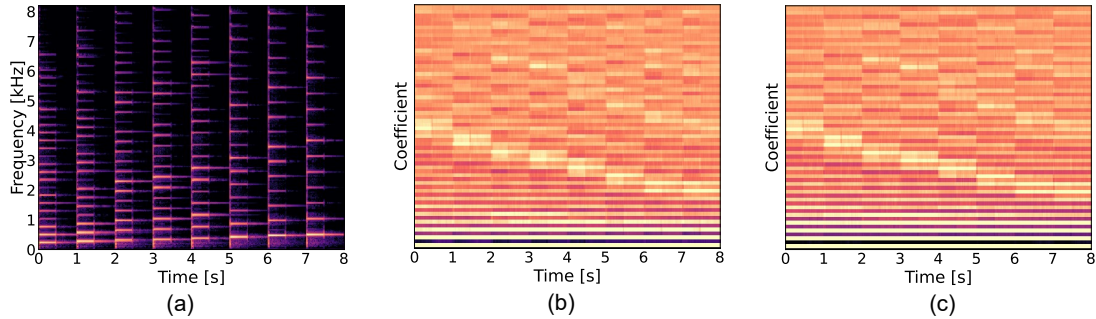


Fig.2.6. (a) power spectrogram of ascending notes played by guitar sound, (b) MFCC, and (c) normalized MFCC.

## 2.4 MLP 型 DNN

DNN の基本形である MLP 型 DNN は、Fig. 2.7 のようにニューラルネットワークを多層に重ねたネットワーク構造で構成されており、ニューラルネットワークの基本単位である「層」が重要となる。まず、ニューラルネットワークが MLP 型 DNN においてどのような役割を果たしているかを説明する。

ニューラルネットワークとは、出力データが入力データの重みづけと非線形写像で表現できるようなネットワーク構造である。中でも、入力データを構成する個々の要素が全ての出力データの要素と結びついているものを全結合層 (dense layer) と呼び、本論文で用いている MLP 型 DNN はこの全結合層を用いて構成している。全結合層の構造は Fig. 2.8 のようになっており、全結合層の各出力  $y_h$  は次式のように表すことができる。

$$y_h = \phi \left( \sum_{n=1}^N w_{nh} a_n + b_h \right) \quad (2.9)$$

ここで、 $n = 1, 2, \dots, N$  は入力データの要素インデックス、 $h = 1, 2, \dots, H$  は出力データの要素インデックス、 $w_{nh}$  は重み係数を表す。また  $\phi$  は活性化関数と呼ばれ、入力データの総和をどのように出力に反映させるかを決定する役割を持つ。これは次の層に渡す値を整えるようなも

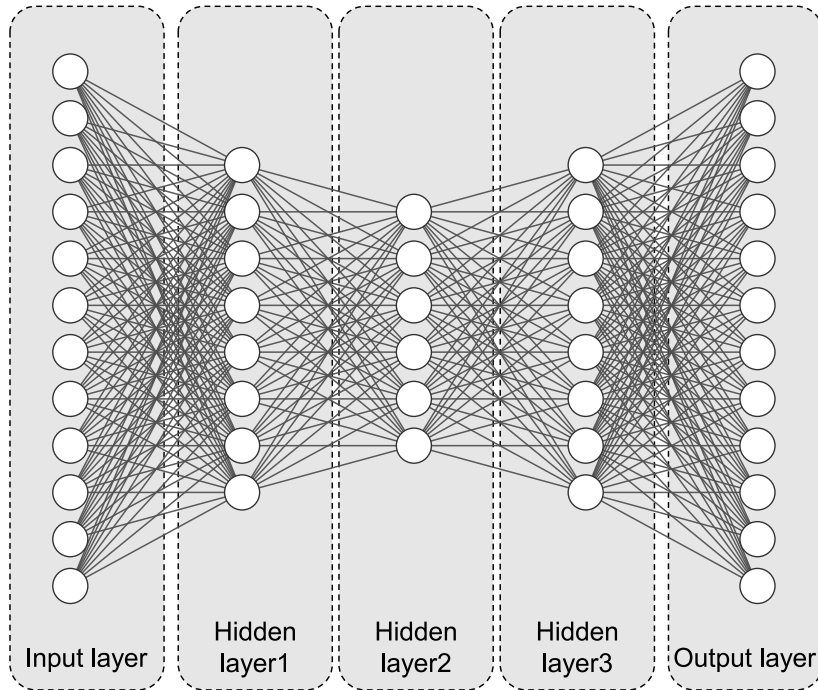


Fig.2.7. Multilayer neural network.

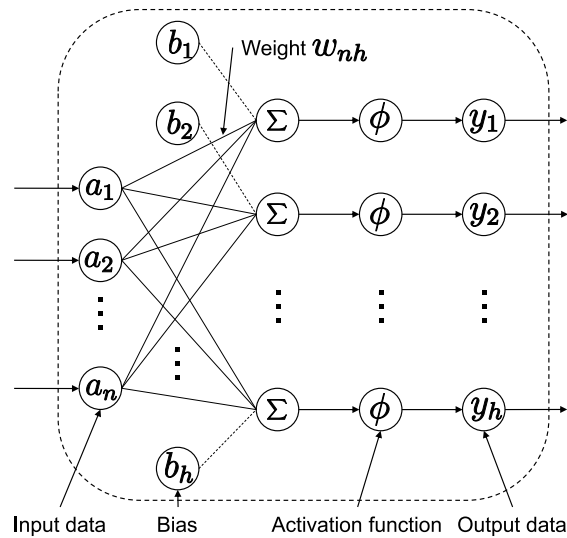


Fig.2.8. Structure of fully connected layer.

のであり、Fig. 2.9 のように様々な活性化関数が定義されている。

式 (2.9) のニューラルネットワーク 1 層を基本単位として多層に重ねたものを DNN と呼ぶ。本節では、例として隠れ層が 3 層で構成される DNN を取り扱う。Fig. 2.7 に示すように、データが最初に入力される層を入力層、入力されたデータが伝達されていく層を隠れ層（中間層）、最終的に値が出力される層を出力層と呼ぶ。また、本研究では、本節で説明した全結合層を用いた MLP 型 DNN と、主に時系列データのモデル化に用いられる、ニューラルネット



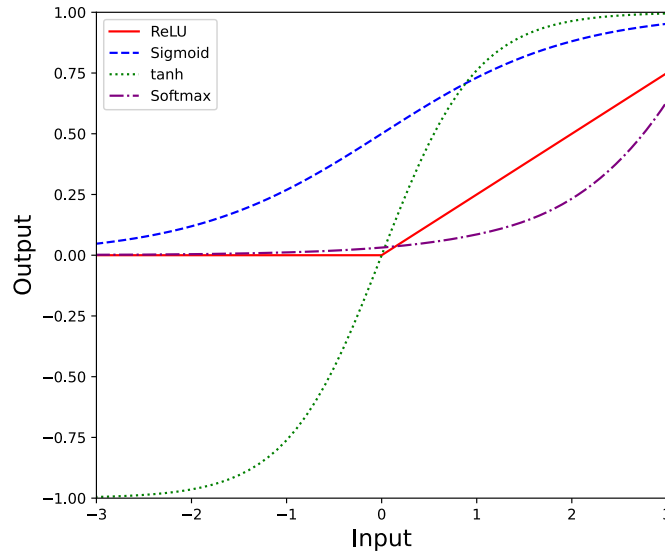


Fig.2.9. Popular activation functions used in DNN.

ワーク内でフィードバックを行う RNN を扱う。RNN については次章で説明する。

## 2.5 既存の類似研究

本論文で取り扱う深層学習に基づく楽器音の生成の類似研究として、DDSP [13], 知覚的メトリクスに基づく正則化付き VAE [21], 及び VAE を用いた音色・音高の分離表現の学習 [22, 23, 24] の 3 手法を紹介する。いずれも、深層学習に基づく生成モデルを活用した新しい音響信号の生成方法を提案したものである。各手法の概要を説明し、本論文で取り扱う提案音生成システムとの類似点や相違点、目的の違いについて述べる。

### 2.5.1 DDSP

類似研究である DDSP について説明する。Fig. 2.10 に DDSP の全体の概要図を示す。DDSP では、入力された音響信号から、音高 (Fig. 2.10 における  $F_0$ ), 音色 (Fig. 2.12 における  $Z$ ), 及び音量 (Fig. 2.10 における Loudness) の 3 つの特徴量をエンコーダで抽出する。この 3 つの特徴量はデコーダに入力され、事前に抽出済みの  $F_0$  及びその整数倍の周波数からなる複数正弦波 (Fig. 2.10 における Harmonic audio) を駆動する変数 (各周波数の振幅値) と、白色雑音に音色を与える FIR フィルタの伝達関数 (Fig. 2.10 における Filtered noise の生成) を出力する。このようにして推論された Harmonic audio と Filtered noise を合成し、最後に必要に応じて残響を付与して合成音を出力する。すべての学習可能なパラメータは、出力の合成音と入力の音響信号間で計算される損失関数が小さくなるように最適化される。DDSP の損失関数には、次式で定義される multi-scale spectral (MSS) ロスが用いられている。

$$L_{\text{MSS}}^{(F)}(\mathbf{A}^{(F)}, \hat{\mathbf{A}}^{(F)}) = \|\mathbf{A}^{(F)} - \hat{\mathbf{A}}^{(F)}\|_1 + \|\log \mathbf{A}^{(F)} - \log \hat{\mathbf{A}}^{(F)}\|_1 \quad (2.10)$$

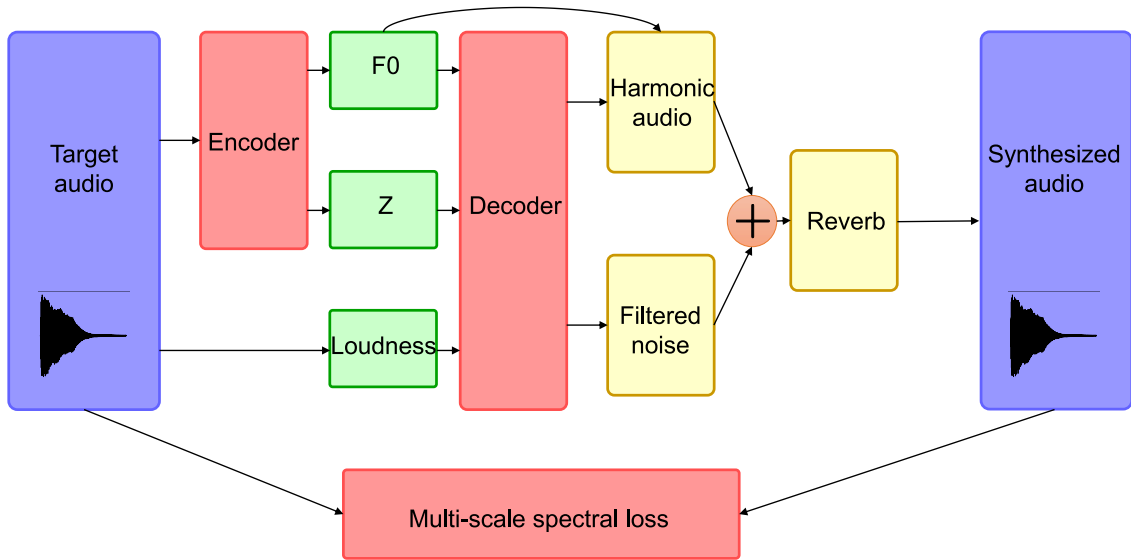


Fig.2.10. DDSP autoencoder architecture. Red components are part of the neural network architecture, green components are the latent representation, and yellow components are deterministic synthesizers and effects. For the detailed explanation of this figure, see [13].

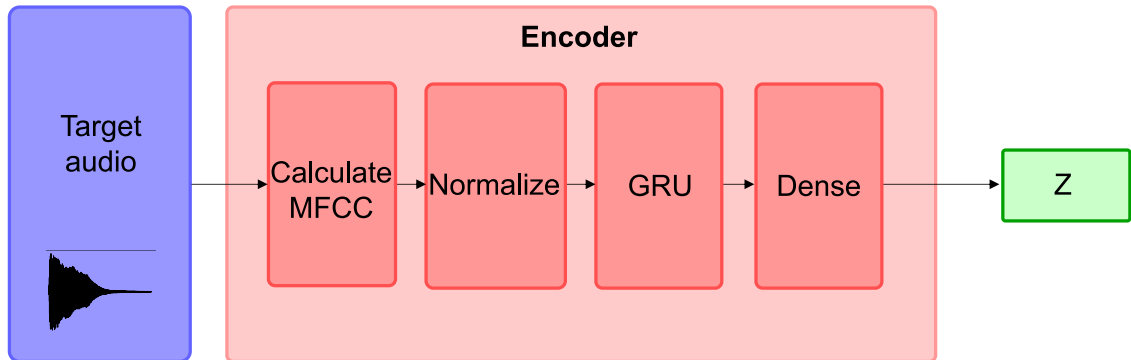


Fig.2.11. Diagram of the Z-encoder. For the detailed explanation of this figure, see [13].

ここで、 $\mathbf{A}^{(F)}$  及び  $\hat{\mathbf{A}}^{(F)}$  はそれぞれ DDSP の入力及び出力の音響信号を、 $F$  点の窓長で STFT して得られる振幅スペクトログラムである。また、 $\|\cdot\|_1$  は行列又はベクトルに対する  $L_1$  ノルム、行列又はベクトルに対する対数関数は要素毎の対数を表す。

DDSP において、入力の音響信号から音高 (F0) を抽出するエンコーダには、CREPE [29] と呼ばれる時間領域の畳み込みニューラルネットワークに基づく推定器が用いられている。また、音色 (Z) を抽出するエンコーダは、Fig. 2.11 に示すように MFCC を入力とするネットワークアーキテクチャが用いられている。具体的には、MFCC を計算し、学習可能な正規化係数を持つ正規化層、GRU、及び全結合層 1 層 (Fig. 2.11 における Dense) を通し、時間フレーム毎の 16 次元の音色特徴量 Z を計算している。最後に、入力の音響信号から音量 (Loudness) を抽出するエンコーダには、決定的な (学習可能なパラメータの無い) 抽出過程

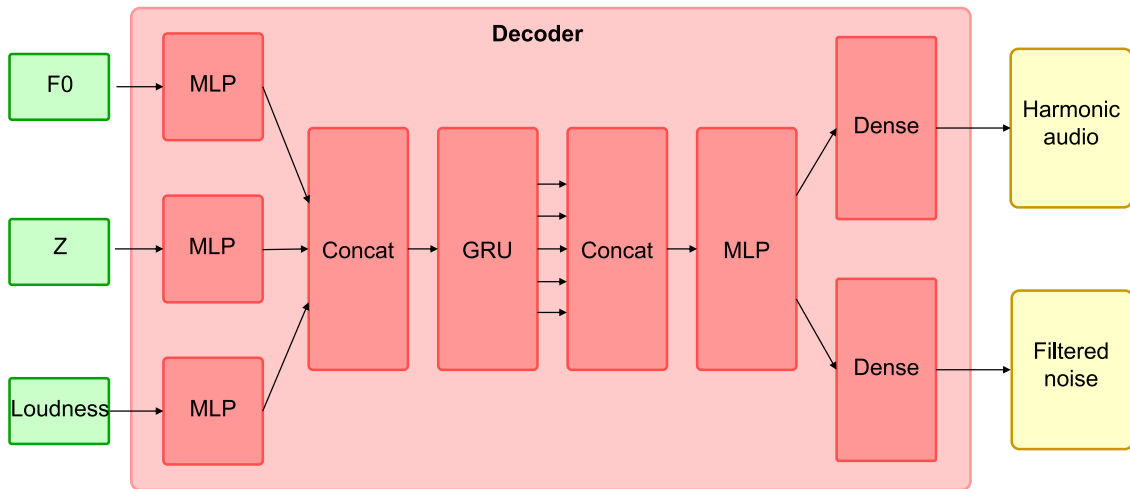


Fig.2.12. Diagram of the decoder for the harmonic synthesizer and the filtered noise synthesizer. For the detailed explanation of this figure, see [13].

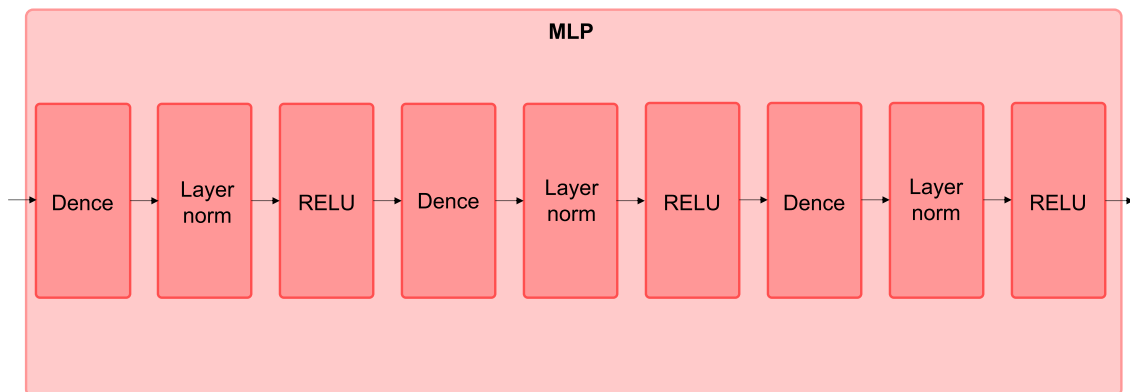


Fig.2.13. MLP in the decoder of DDSP [13].

として、A 特性と呼ばれる人間の聴覚を考慮した周波数特性で重み付けされたパワースペクトルの対数を取ったベクトルの平均と分散 [30] が用いられている。

DDSP において、エンコーダで抽出された各特徴量から、Harmonic audio 及び Filtered noise を駆動する出力を得る計算には、Fig. 2.12 に示すアーキテクチャを持つデコーダが用いられている。また、Fig. 2.12 中の MLP は Fig. 2.13 の構成となっている。

DDSP と提案音生成システムの類似点、手法の違い、及び目的をまとめる。類似点は、入力音響信号から音色、音高、及び音量という 3 つの特徴量に対してエンコーダ・デコーダを通して音響信号を生成している点である。一方で、DDSP はデコーダを通して得られるパラメータから Harmonic audio 及び Filtered noise を駆動しており、これは入力音響信号を正弦波とノイズで人工的に合成していることに対応する。このような音響信号の合成は比較的頑健な音響信号の生成が可能である反面、どれだけパラメータを正しく推定しても、非常にリアルな楽器音の再現などは難しく、人工的な音響信号しか生成できないデメリットがある。提案音生

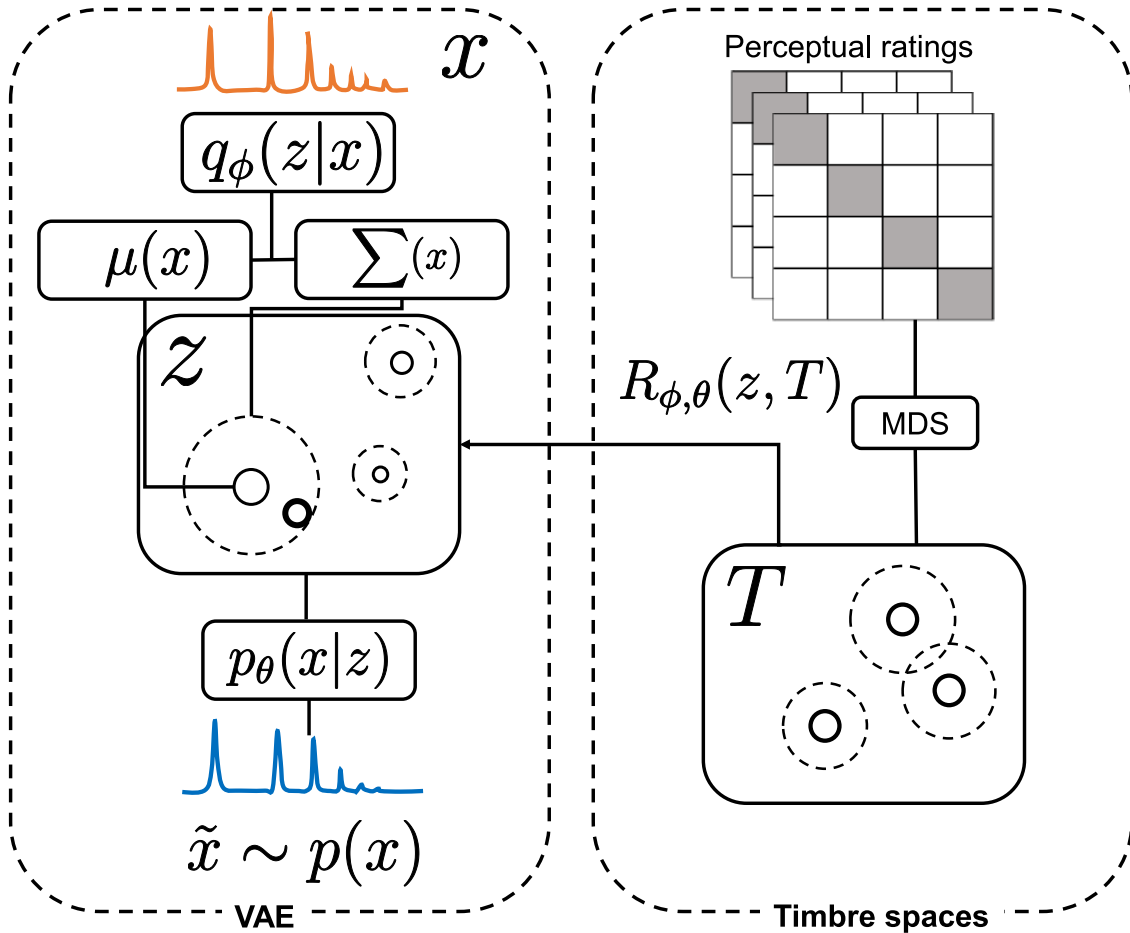


Fig.2.14. Concept of regularizing VAE with perceptual metrics. Latent spaces and variables in VAE are forced to be matched to the perceptual timbre spaces structured by perceptual ratings. For the details of this figure, see [21].

成システムは、デコーダから直接振幅スペクトログラムを生成することを目指しており、最適な学習ができるならば DDSP よりも自由度が高く自然な音響信号が生成できると考えられる。また、DDSP と提案音生成システムは目的も異なる。DDSP は入力音響信号の合成器（シンセサイザー）を構成することが目的であり、結果的に音色変換等へ応用することは可能であるが、音色変換が主目的ではない。MFCC も GRU 等で非線形に変換されているため、音色に対応する潜在特徴量の制御等は難しいと考えられる。提案音生成システムは、VAE に基づく新しい音色の生成を目的としており、解釈性のある程度担保した潜在空間で音色を制御することができる。

## 2.5.2 知覚的メトリクスに基づく正規化付き VAE

VAE は学習データに潜む構造を低次元の多次元正規分布に従う特徴量として潜在空間に埋め込むことができる。この潜在空間は多次元正規分布に従うことから、Fig. 1.1 に示すように

一定の解釈性や乱数からのデータ生成が可能であるが、潜在空間は物理的又は知覚的な構造と結びついていないという問題がある。この問題を解決するために、楽器音の生成を目的として、知覚的メトリクスに基づく正則化付き VAE が提案されている [21]。Fig. 2.14 に、この手法の概要図を示している。Fig. 2.14 の右上に示す Perceptual ratings は、過去の研究 [31, 32] で蓄積された楽器音の音色に関する知覚メトリクスである。即ち、複数の楽器音間の音色の類似・相違を数値的にレーティングした表形式のデータである。この知覚的メトリクスを、多変量解析手法の 1 つである多次元尺度構成法 (multi-dimensional scaling) で低次元空間に変換することで、知覚的メトリクスに基づく音色空間  $T$  を構成している。さらに、音響信号のスペクトルを入出力とする VAE において、潜在空間の分布構造が先の音色空間  $T$  の構造と類似するように、VAE の学習に正則化項 (Fig. 2.14 中の  $R_{\phi, \theta}(z, T)$ ) を付与している。このような正則化を導入することで、VAE の潜在空間が知覚的メトリクスと対応付けられた形で学習され、音色という知覚的情報と深く結びついた解釈性の高い生成モデルを得ることができる。そのため、例えば VAE の潜在空間上で各楽器音の相関を解析することや、ある楽器から別の楽器へ連続的に音響信号を変化させることなどが実現でき、1 章で述べた本論文の目的が達成できる可能性がある。

知覚的メトリクスに基づく正則化付き VAE と提案音生成システムの類似点、手法の違い、及び目的をまとめる。類似点は、VAE の潜在空間を音色空間に近づけ、潜在空間上で乱数を与えて様々な楽器音の生成を可能とすることである。しかしながら、文献 [21] の手法は主観的な音色の尺度を持つ知覚メトリクスに基づいており、十分な主観評価データを収集するコストは非常に大きくなるという困難性がある。さらに、主観的な音色の尺度は個人差を多く含んでいるため、正確な音色空間が構成できない可能性もある。提案音生成システムは客観的な音色の尺度である MFCC を用いているため、収集コストは低く、個人差によるデータの問題も発生しないという利点がある。両手法の目的は類似しており、いずれも VAE を用いて潜在空間上に解釈性の高い音色空間を構築することで、「複数楽器音の中間の音色」という通常では定義できない音響信号の生成を実現することである。

### 2.5.3 VAE を用いた音色・音高の分離表現の学習

本論文の目的に合致した既存手法として、VAE を用いて音色と音高が分離された表現 (timbre-pitch-disentangled representations) の学習法が提案されている [22, 23, 24]。これらの手法ではいずれも、Fig. 2.15 に示すように、楽器音の音色と音高が分離された潜在空間及び生成モデルを学習することを目的としている。このような生成モデルの学習において VAE を用いることで、音色の潜在空間及び音高の潜在空間を独立に獲得でき、1 章で述べた本論文の目的に合致したモデルを構築できる可能性がある。分離された音色と音高の潜在変数はデコーダの入力で結合され、楽器音信号が再構成される仕組みである。

上記の手法は本論文の目的と合致したモデルが構築できる可能性がある。また、これらのモデルを用いることで、例えば「複数の楽器音の中間の音色を持つ音響信号」を生成することが

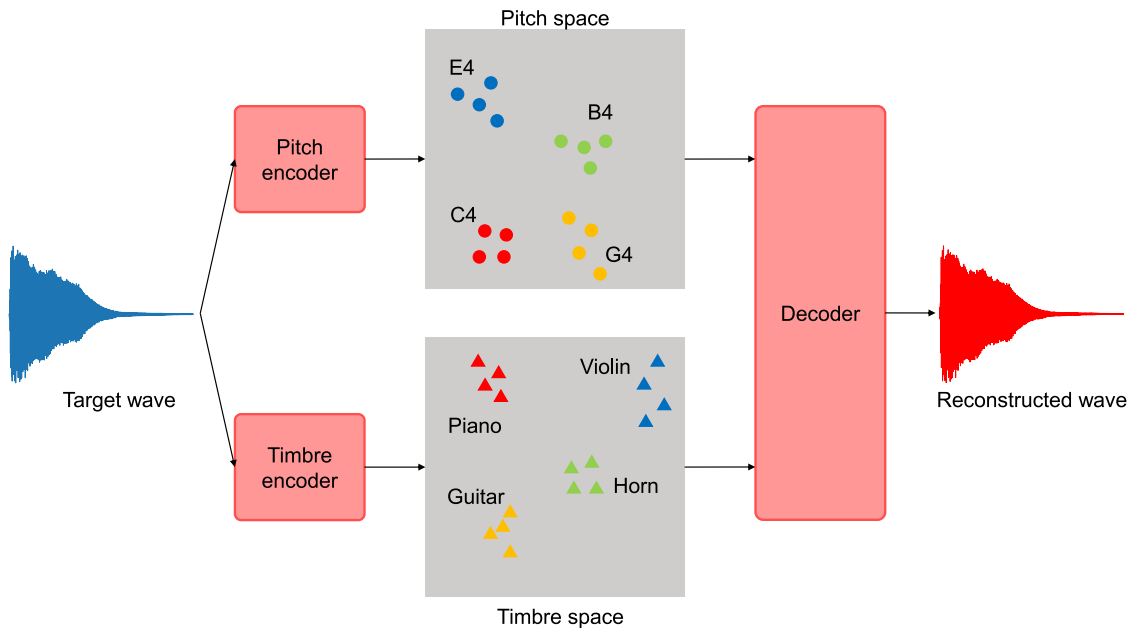


Fig.2.15. VAE for learning disentangled pitch and timbre representations of music instrument sounds. For the details of this figure, see [24].

可能と考えられる。しかしながら、前述のいずれの手法でも、時間的な音量変化（DDSPにおけるラウドネス）<sup>\*1</sup>という特徴量は分離されておらず、音色か音高のいずれか又は両方の潜在空間に押し込められていると考えられる。音を決定づける3要素は音色・音高・音量であるため、時間的な音量変化が人間の「その楽器音らしさ」という知覚に与える影響は無視できない可能性がある。本論文ではこの違いに焦点を当て、音色と音高の他にも時間的な音量変化を加味した生成モデルとして、1章で説明した提案音生成システムの構築を目指している。

## 2.6 本章のまとめ

本章では、1章で説明した提案音生成システムにおいて必要となる基礎技術と提案音生成システムと類似した研究について説明した。2.2節では、音響信号処理でよく用いられるSTFTについて説明した。2.3節では、本論文で主題となる音色特徴量のMFCCについて説明した。2.4節では、最も基本的なDNNであるMLP型DNNについて説明した。2.5節では、提案システムに類似した研究としてDDSP、知覚的メトリクスに基づく正規化付きVAE、及びVAEを用いた音色・音高の分離表現の学習について紹介し、提案音生成システムとの類似点、相違点、及び目的について述べた。次章では提案音生成システムの詳細と問題点について説明し、その問題の解決方法について説明する。

<sup>\*1</sup> ラウドネスとは、人間の聴覚特性を考慮した時の音の大小を示す心理量である [25]。本論文で用いる音量特徴量は人間の聴覚特性を考慮していないため、時間的な音量変化または total amplitude と呼ぶ。

## 第3章

# 提案手法

### 3.1 まえがき

本章では、1章で説明した提案音生成システムについての詳細を説明し、本論文で取り組むべき問題である音高、音色、及び音量の3つの特徴量から振幅スペクトログラムを予測する方法について説明する。3.2節では、提案音生成システムの詳細な処理の流れについて説明する。3.3節では、提案音生成システムの問題点を述べ、その問題を解決する方法について説明する。3.4節では、音高、音色、及び音量の3つの特徴量から振幅スペクトログラムの予測に用いるDNNモデルである、MLP型DNN、BiLSTM型DNN、及びBiGRU型DNNについて説明する。3.5節で本章をまとめる。

### 3.2 提案音生成システム全体の説明

Fig. 1.2をより詳細にした提案音生成システムの全体図をFig.3.1に示す。本節では、この図の左端から右端まで順に詳細を説明する。今、入力となる音響信号の時間波形 (Fig. 3.1における Original wave) を  $\mathbf{x} = [x(1), x(2), \dots, x(L)]^T \in \mathbb{R}^L$  と定義する。この信号の振幅スペクトログラムは次式で得られる。

$$\mathbf{X} = |\text{STFT}_\omega(\mathbf{x})| \in \mathbb{R}_{\geq 0}^{I \times J} \quad (3.1)$$

この振幅スペクトログラム  $\mathbf{X}$  をエンコーダに入力し、音高、MFCC、及び total amplitude の3つを抽出する。本論文では、音高は入力された音響信号に C4 や E5 等のラベルとして与えられている状況を想定し、エンコーダで音高を抽出する過程は省略している。将来的には、DDSPのように音響信号から音高を推定する手法を組み合わせることを想定している。

次に、total amplitude は DDSP よりも簡易的な抽出方法として、次式で計算する。

$$v_j = \sum_{i=1}^I x_{ij} \quad (3.2)$$

ここで、 $x_{ij}$  は振幅スペクトログラム  $\mathbf{X}$  の要素を表す。式 (3.2) は、振幅スペクトログラム  $|\mathbf{X}|$  の各列ベクトルの  $L_1$  ノルムに対応する。式 (3.2) は DDSP における Loudness と実質的

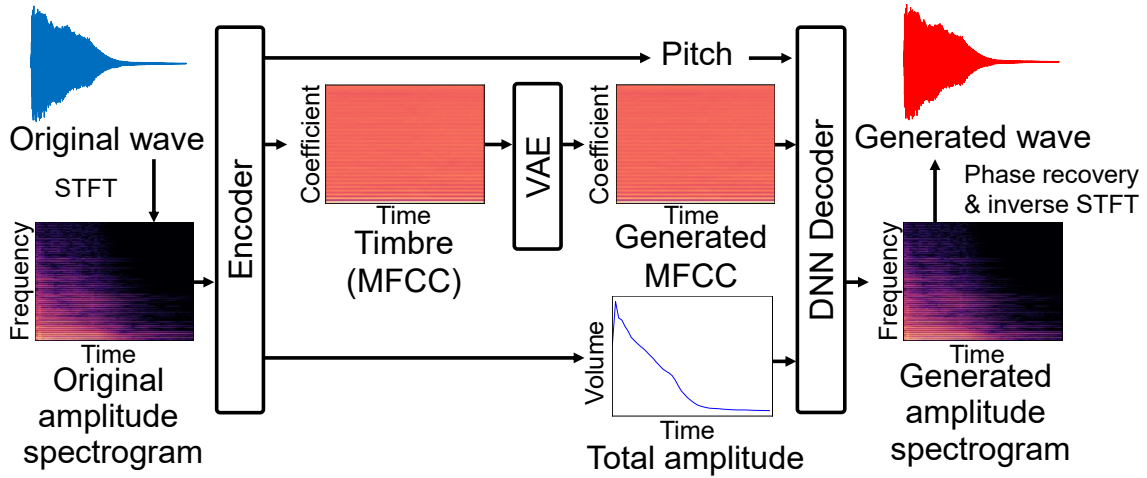


Fig.3.1. Detailed process flow of the proposed sound generation system. The system consists of inner VAE and outer encoder-decoder. The outer encoder extracts pitch, MFCC, and total-amplitude from an input amplitude spectrogram, Then, MFCC to the inner VAE. The generated MFCC obtained from the inner VAE is input to the outer decoder with the extracted pitch and total amplitude. Finally, the amplitude spectrogram is reconstructed from the outer decoder.

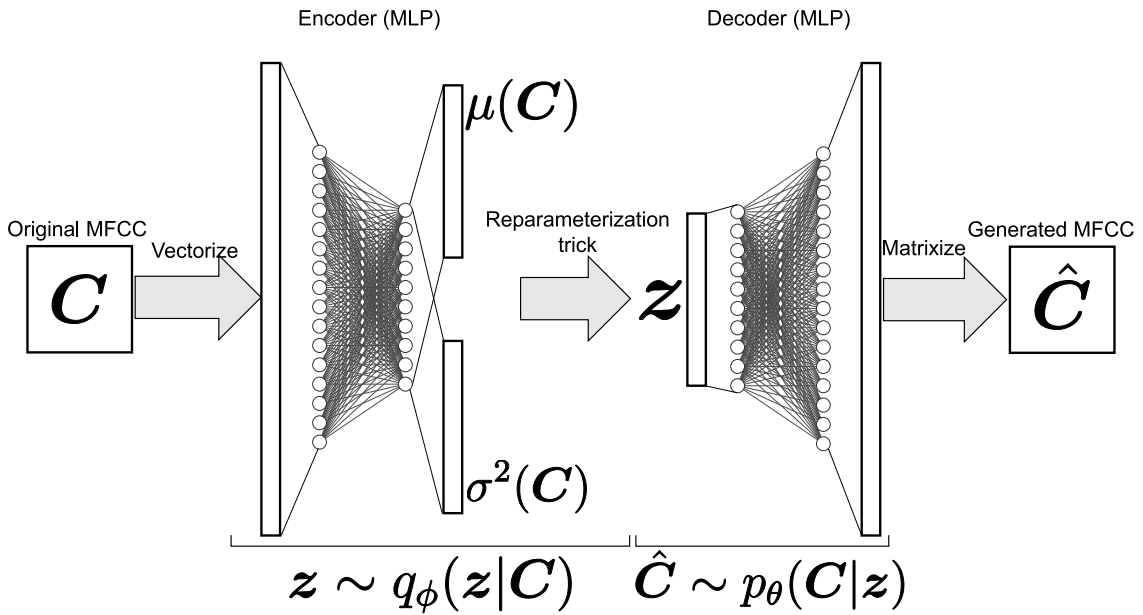


Fig.3.2. Architecture of vanilla VAE for inferring a generative model of MFCCs.

に同じであるが、音響分野で聴覚指標として用いられる専門用語の「ラウドネス」との混同を防ぐために、本論では式 (3.2) を以後 total amplitude と呼ぶ。

このようにして得られた total amplitude は、全時間フレームに関してまとめたベクトル  $v = [v_1, v_2, \dots, v_J]^T \in \mathbb{R}_{\geq 0}^J$  と定義する。最後に、音色特徴量である MFCC は 2.3 節で述べた方法で計算する。但し、入力の振幅スペクトログラム  $X$  を音量に依存しないものに変



換するため、MFCC の計算の前に  $\mathbf{X}$  を次式で正規化し、正規化後の振幅スペクトログラム  $\overline{\mathbf{X}} \in \mathbb{R}_{\geq 0}^{I \times J}$  に変換する。

$$\overline{x}_{ij} = \frac{x_{ij}}{v_j} \quad (3.3)$$

ここで、 $\overline{x}_{ij}$  は正規化後の振幅スペクトログラム  $\overline{\mathbf{X}}$  の要素である。従って、正規化後の振幅スペクトログラム  $\overline{\mathbf{X}}$  から 2.3 節の方法で MFCC  $\mathbf{C} \in \mathbb{R}^{K \times J}$  を得る。ここで、 $K$  は 2.3 節の定義と同じく MFCC の次元数（メルフィルタバンクのフィルタ数）である。

エンコーダによって音高、MFCC、及び total amplitude を抽出した後は、MFCC のみを VAE に入力する。この VAE には、最も基本的な構造（バニラ VAE）を用いる。バニラ VAE の生成モデルの概要図を Fig. 3.2 に示す。また、下記に Fig. 3.2 を用いたバニラ VAE の大まかな説明を示す。より具体的な定式化や学習方法に関しては文献 [19] を参照されたい。まず、潜在変数  $\mathbf{z} \in \mathbb{R}^D$  の事前分布  $p_\theta(\mathbf{z})$  を原点对称多次元 Gauss 分布

$$\mathbf{z} \sim p_\theta(\mathbf{z}) \quad (3.4)$$

$$= \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (3.5)$$

と仮定する。ここで、 $D$  は潜在変数の次元であり、 $0 < D \ll KJ$  を満たす。また、 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  は平均ベクトル  $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_D]^\top \in \mathbb{R}^D$  及び分散共分散行列  $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$  を持つ多次元 Gauss 分布、 $\mathbf{0}$  及び  $\mathbf{I}$  はそれぞれ適切なサイズのゼロベクトル及び単位行列を表す。バニラ VAE では、この事前分布仮定の下で、入力として与えられた MFCC  $\mathbf{C}$  から、 $\mathbf{z}$  の近似事後分布  $q_\phi(\mathbf{z}|\mathbf{C})$  (VAE のエンコーダ) 及び尤度関数  $p_\theta(\mathbf{C}|\mathbf{z})$  (VAE のデコーダ) を変分下限最大化に基づき学習する。このとき、エンコーダ及びデコーダにそれぞれ下記の分布を仮定する。

$$q_\phi(\mathbf{z}|\mathbf{C}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{C}), \text{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{C}))) \quad (3.6)$$

$$p_\theta(\mathbf{C}|\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{z}), \text{diag}(\boldsymbol{\sigma}_\theta^2(\mathbf{z}))) \quad (3.7)$$

ここで、 $\boldsymbol{\sigma}^2 = [\sigma_1^2, \sigma_2^2, \dots, \sigma_D^2]^\top \in \mathbb{R}^D$  は分散共分散行列の対角成分（分散）を並べたベクトルである。また、 $\text{diag}(\cdot)$  は入力ベクトルを対角要素に持つ対角行列を表す。実際には、これらの分布からのサンプリングは Fig. 3.2 に示すように MLP 型 DNN によるオートエンコーダで実装されるため、VAE はオートエンコーダ型変分推論（auto-encoding variational Bayes）と呼ばれる [18]。但し、Fig. 3.2 中の Reparameterization trick は、潜在変数  $\mathbf{z}$  の  $q_\phi(\mathbf{z}|\mathbf{C})$  からのサンプリングを微分可能な確定的関数に置き換える処理である [19]。これは VAE 全体を誤差逆伝播可能とするために必要となる。このような定式化で推定される VAE は、MFCC を生成する  $\mathbf{z}$  の潜在空間を多次元 Gauss 分布仮定の下で学習できる。この時、VAE における損失関数は以下の式を用いる。

$$\mathcal{L}(\mathbf{C}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{C})}[\log p_\theta(\mathbf{C}|\mathbf{z})] - D_{\text{KL}}[q_\phi(\mathbf{z}|\mathbf{C})||p_\theta(\mathbf{z})] \quad (3.8)$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{C})}[\log p_\theta(\mathbf{C}|\mathbf{z})] + \frac{1}{2} \sum_{d=1}^D (\log(\sigma_d^2) - \sigma_d^2 - \mu_d^2 + 1) \quad (3.9)$$

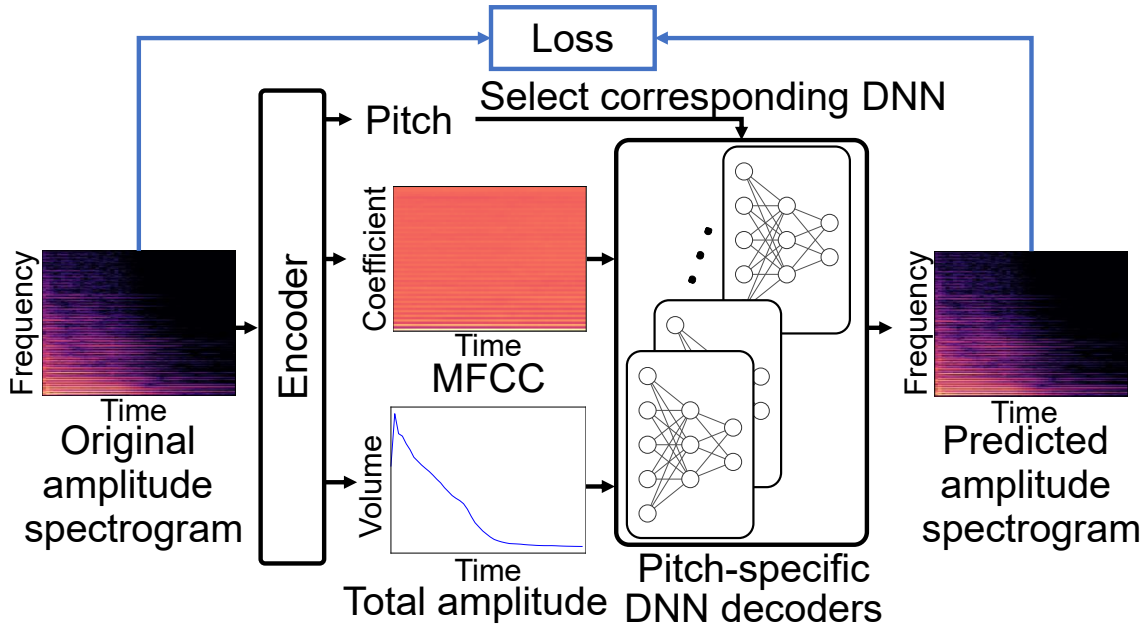


Fig. 3.3. Training process flow of the proposed DNN-based timbre decoder.

ここで、 $E[\cdot]$  及び  $D_{KL}[\cdot|\cdot]$  はそれぞれ期待値及び Kullback–Leibler 情報量である。また、(3.8) の第 1 項及び第 2 項はそれぞれ、再構成誤差及び正規化項と呼ばれる。学習後は、潜在変数  $z$  を乱数等から与えることで、1 章で示したような新しいデータを生成できる。即ち、Fig. 3.2 の場合は学習済みの VAE から新しい音色 (MFCC) を生成することができる。

最後に、音高、VAE から出力された MFCC  $\hat{C}$ 、及び total amplitude の 3 つの特徴量をデコーダに入力し、振幅スペクトログラム  $\hat{X}$  を生成する。このデコーダが本論文で取り扱う主題であり、次節でその詳細を述べる。

以上が提案音生成システムの詳細である。学習時は Fig. 3.1 に示す入力と出力 (Original wave と Generated wave) 間の損失が少なくなるように VAE 及びデコーダをそれぞれ学習する。学習後は、Fig. 1.3 に示すように、乱数等から作成した潜在変数  $z$  から新しい楽器音の音響信号を生成できるようになることが期待される。

### 3.3 提案音生成システムにおける問題

前節で述べたように、Fig. 3.1 の提案音生成システムでは、音高、VAE から生成された MFCC、及び total amplitude の 3 つの特徴量から振幅スペクトログラムを生成する必要がある。しかしながら、MFCC は音色のみを低次元空間で表現した特徴量であることから、音高、MFCC、及び total amplitude の 3 つの特徴量から解析的に振幅スペクトログラムを求めることはできない。従って、3 つの特徴量から振幅スペクトログラムを予測する何らかの非線形な変換をデコーダに用いる必要が生じる。

本論文では、前述の問題を取り扱うことを主目的とし、Fig. 3.1 に示す提案音生成システムを実装する上で必要不可欠なデコーダを DNN で実現する方法について実験的に検討する。す

なわち、音高、MFCC、及び total amplitude の 3 つの特徴量から振幅スペクトログラムを高精度に予測する DNN の構築を目指す。この本論文で取り扱う主目的を Fig. 3.3 に示す。なお、本論文で取り扱うデコーダは MFCC 及び total amplitude を入力とする DNN を想定している。音高の特徴量は離散的であることから、DNN の入力に与えるのではなく、各音高専用に学習した DNN を選択するために用いる。すなわち、予め学習された音高依存の DNN を複数用意し、いずれかの DNN が入力の音高により選択される。このような方式を取ることで、DNN に基づくデコーダは音高に対する汎化性能を獲得する必要がなくなり、より高精度な振幅スペクトログラムの予測が可能になると考えられる。最後に、デコーダとして用いる DNN は、最も基本的な DNN である MLP 型 DNN 及び時系列データの繋がりを考慮した学習が可能な BiLSTM 型 DNN 及び BiGRU 型 DNN の 3 種類を本論文では取り扱う。

## 3.4 DNN に基づくデコーダ

本節では、本論文で実験的に調査する DNN に基づくデコーダの詳細について述べる。3.2 節及び 3.3 節で述べた通り、このデコーダは VAE で生成された MFCC  $\hat{C}$  及び total amplitude  $v$  を入力とし、そのような MFCC と total amplitude を持つ振幅スペクトログラム  $\hat{X}$  を予測する DNN である。このとき、音高は入出力ともに既知として固定しており、音高の種類の数だけ DNN を学習することを想定している（即ち、C3 音専用の DNN デコーダや E4 音専用の DNN デコーダ等を別々に学習して用意する）。また、本来の提案音生成システムでは VAE で生成された MFCC  $\hat{C}$  を入力に用いるが、本実験では基礎的な調査として、入力の音響信号から計算される MFCC  $C$  そのものを DNN デコーダの入力に用い、入力の音響信号の振幅スペクトログラム  $X$  そのものをラベルとして学習する。この手続きは、VAE で非常に高精度な（その MFCC を持つ音響信号が実際に存在するような）MFCC が生成されることを仮定していることに相当する。

DNN デコーダのアーキテクチャについては、3.4.1 項から 3.4.3 項で述べる通り、MLP 型、BiLSTM 型、及び BiGRU 型の 3 種類を用い、どのアーキテクチャが高精度に振幅スペクトログラム  $X$  を予測できるか調査する。なお、これまでの章・節においてイタリック体で定義してきた各変数の文字と、本節での DNN の説明に用いる各変数の文字が重複・混同することを避けるため、本節の説明に限って登場する変数はローマン体（例えば  $x$ ,  $y$ ,  $p$  等）で定義し使用する。

### 3.4.1 MLP 型 DNN

デコーダとして最も標準的な DNN である MLP を用いる場合について説明する。MLP の入力層に与える入力ベクトル、出力層から得られる出力ベクトル、及び出力ベクトルの正解となるラベルベクトルをそれぞれ  $x$ ,  $y$ , 及び  $p$  と定義する。これまでに説明した通り、DNN デコーダは MFCC と total amplitude を入力特徴量とするため、MFCC  $C$  を 1 次元に整形（ベクトル化）し、total amplitude ベクトル  $v$  を結合した  $KJ + J$  次元のベクトルを入力ベ

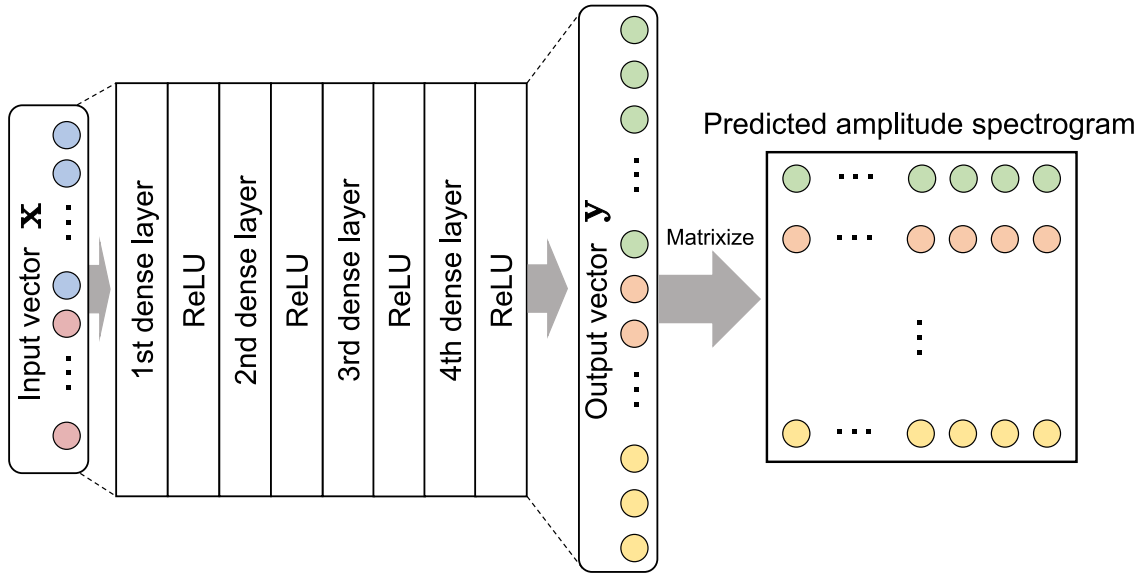


Fig.3.4. Architecture of MLP used as the DNN decoder.

クトル  $\mathbf{x}$  とする。このとき、MFCC  $\mathbf{C}$  は入力の音響信号から 2.3 節の方法で直接計算したものをを用いる。またラベルベクトル  $\mathbf{p}$  は、入力の音響信号から直接計算される振幅スペクトログラム  $\mathbf{X}$  をベクトル化して得られる  $IJ$  次元のベクトルとする。従って、 $IJ$  次元のベクトルである出力ベクトル  $\mathbf{y}$  を、 $I \times J$  次元の行列に整形（行列化）することで、予測結果の振幅スペクトログラムを得ることができる。

MLP の学習には、次式に示す MSE 損失関数を用いる。

$$\mathcal{L}_{\text{MSE}}(\mathbf{y}, \mathbf{p}) = \|\mathbf{y} - \mathbf{p}\|_2^2 \quad (3.10)$$

ここで、 $\|\cdot\|_2$  は行列の  $L_2$  ノルムである。MLP の学習では、この損失関数の値が小さくなるよう、誤差逆伝播によって全パラメータ（重み係数及びバイアス）を最適化する。

Fig. 3.4 に本論文で用いた MLP の構造を示す。入力層、隠れ層 4 層、及び出力層の計 6 層からなる全結合型 DNN となっている。DNN 学習時の勾配消失問題を回避するために、各隠れ層の非線形関数には rectified linear unit (ReLU) を使用している。これにより、出力ベクトルは 0 以上の要素を持つことが担保される。また、各隠れ層の次元数は入力側から順番に 1024, 512, 512, 及び 52275 と設定している。

### 3.4.2 BiLSTM 型 DNN

デコーダとして双方向 RNN (bidirectional RNN: BiRNN) の一種である BiLSTM を用いる場合について説明する。BiRNN は、MFCC やスペクトログラム等のように時間という物理量の次元を持つ入力に対して、時間方向の再帰性を考慮した学習が可能な DNN である。BiRNN の構造を Fig. 3.5 に示す。通常の RNN とは異なり、過去から未来及び未来から過去の双方向で 2 つの RNN を駆動し、各時刻の結果を統合したものを出力とする構造を持っている。

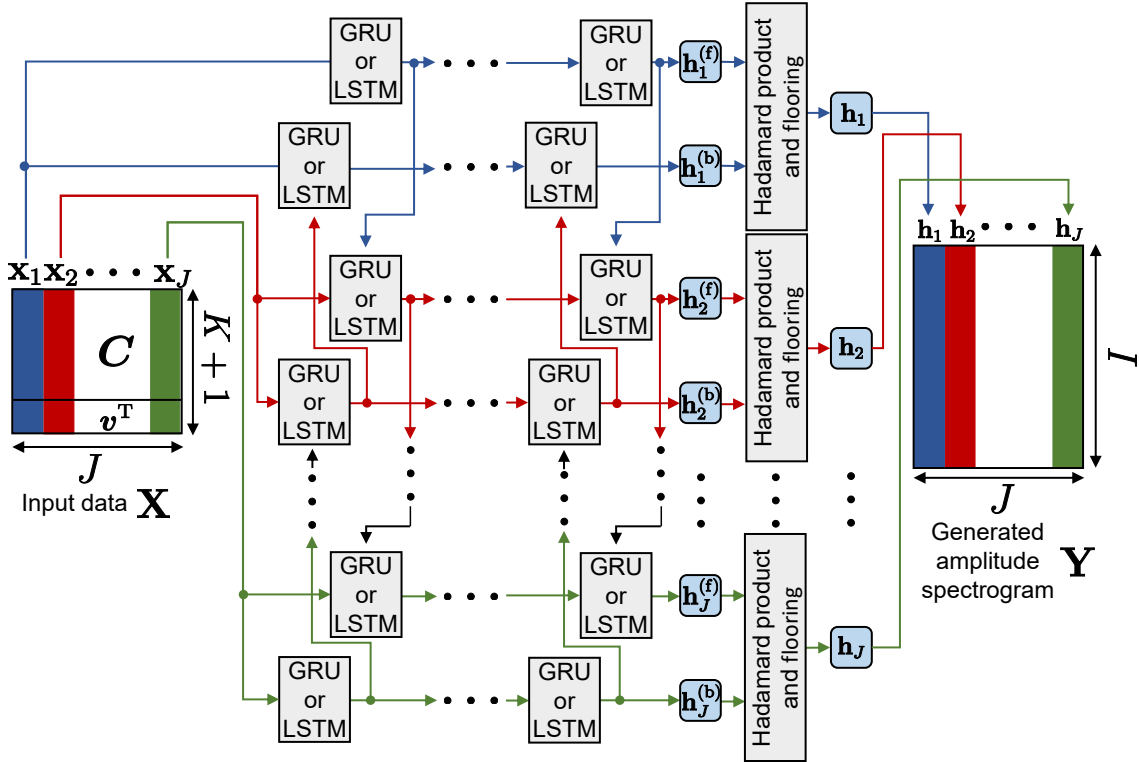


Fig.3.5. Architecture of BiRNN used as the DNN decoder.

る。この時、過去から未来の方向（以後、順方向と呼ぶ）の各 RNN の出力を過去側から順に  $\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \dots, \mathbf{h}_J^{(f)}$ 、未来から過去の方向（以後、逆方向と呼ぶ）の各 RNN の出力を過去側から順に  $\mathbf{h}_1^{(b)}, \mathbf{h}_2^{(b)}, \dots, \mathbf{h}_J^{(b)}$  とする。そのため、通常の RNN における時系列予測のようなオンラインの入力には対応していないが、時間の次元を持つ MFCC 及び total amplitude のバッチ入力に対しては、MLP よりも少ないパラメータで強力な表現力を持つことが期待できる。

BiRNN の入力、出力、及びラベルをそれぞれ  $\mathbf{X}$ 、 $\mathbf{Y}$ 、及び  $\mathbf{P}$  と定義する。これらは時間という次元を持つため、少なくとも 2 次元のデータとなる。本論文においては、MFCC と total amplitude を次のように結合した行列を入力  $\mathbf{X}$  と定義する。

$$\mathbf{X} = \begin{bmatrix} \mathbf{C} \\ \mathbf{v}^T \end{bmatrix} \in \mathbb{R}^{(K+1) \times J} \quad (3.11)$$

また、出力  $\mathbf{Y}$  及びラベル  $\mathbf{P}$  は振幅スペクトログラムそのものであり、いずれも  $I \times J$  の非負行列となる。

BiLSTM は、BiRNN における双方向の RNN に LSTM ユニット [26] を用いた構造を持つ DNN である。LSTM ユニットの内部を Fig. 3.6 に示す。ここで、LSTM ユニットへの入力及び出力をそれぞれ  $\bar{\mathbf{X}} \in \mathbb{R}^{D \times J}$  及び  $\mathbf{H} \in \mathbb{R}^{D' \times J}$  と定義し、これらの時刻  $j$  におけるベクトルをそれぞれ  $\bar{\mathbf{x}}_j$  及び  $\mathbf{h}_j$  と定義している。また、 $D$  及び  $D'$  はそれぞれこの LSTM ユニットに入力された時点での  $\bar{\mathbf{X}}$  の特徴量の次元数及びこの LSTM ユニットから出力される時点での

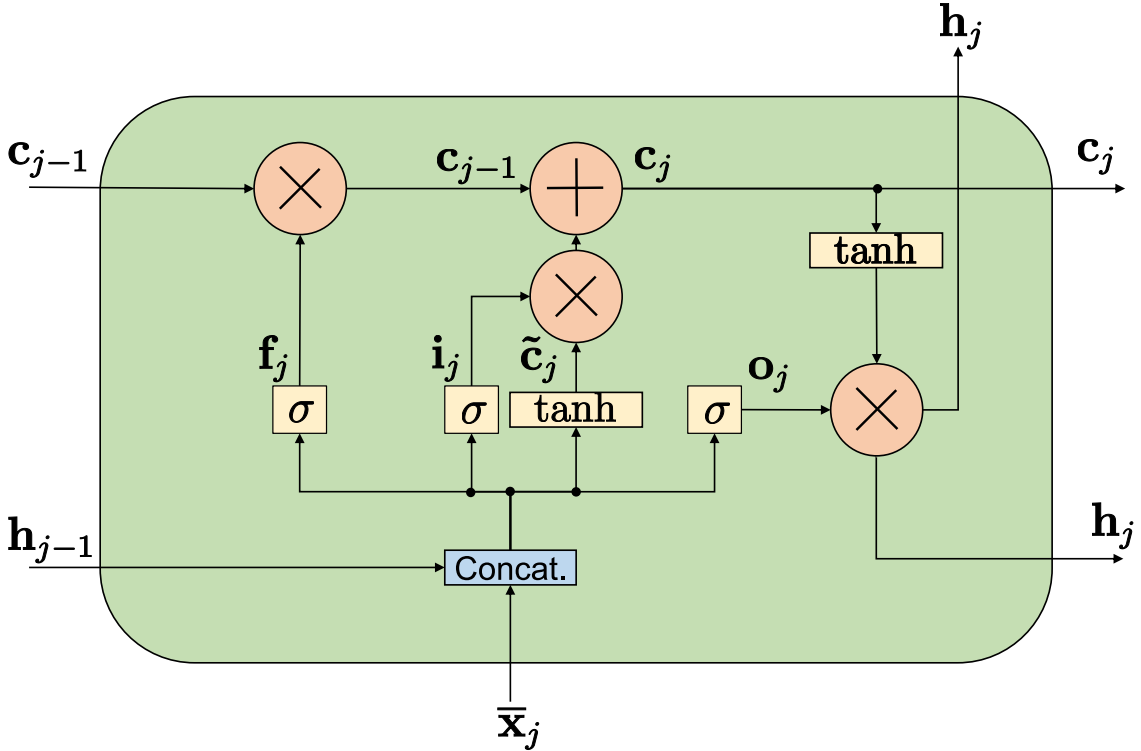


Fig.3.6. Structure of LSTM unit.

$\mathbf{H}$  の特徴量の次元数である。Fig. 3.6 の各変数の計算は次式となる。

$$\mathbf{f}_j = \sigma(\mathbf{W}^f \bar{\mathbf{x}}_j + \mathbf{R}^f \mathbf{h}_{j-1} + \mathbf{b}^f) \quad (3.12)$$

$$\mathbf{i}_j = \sigma(\mathbf{W}^i \bar{\mathbf{x}}_j + \mathbf{R}^i \mathbf{h}_{j-1} + \mathbf{b}^i) \quad (3.13)$$

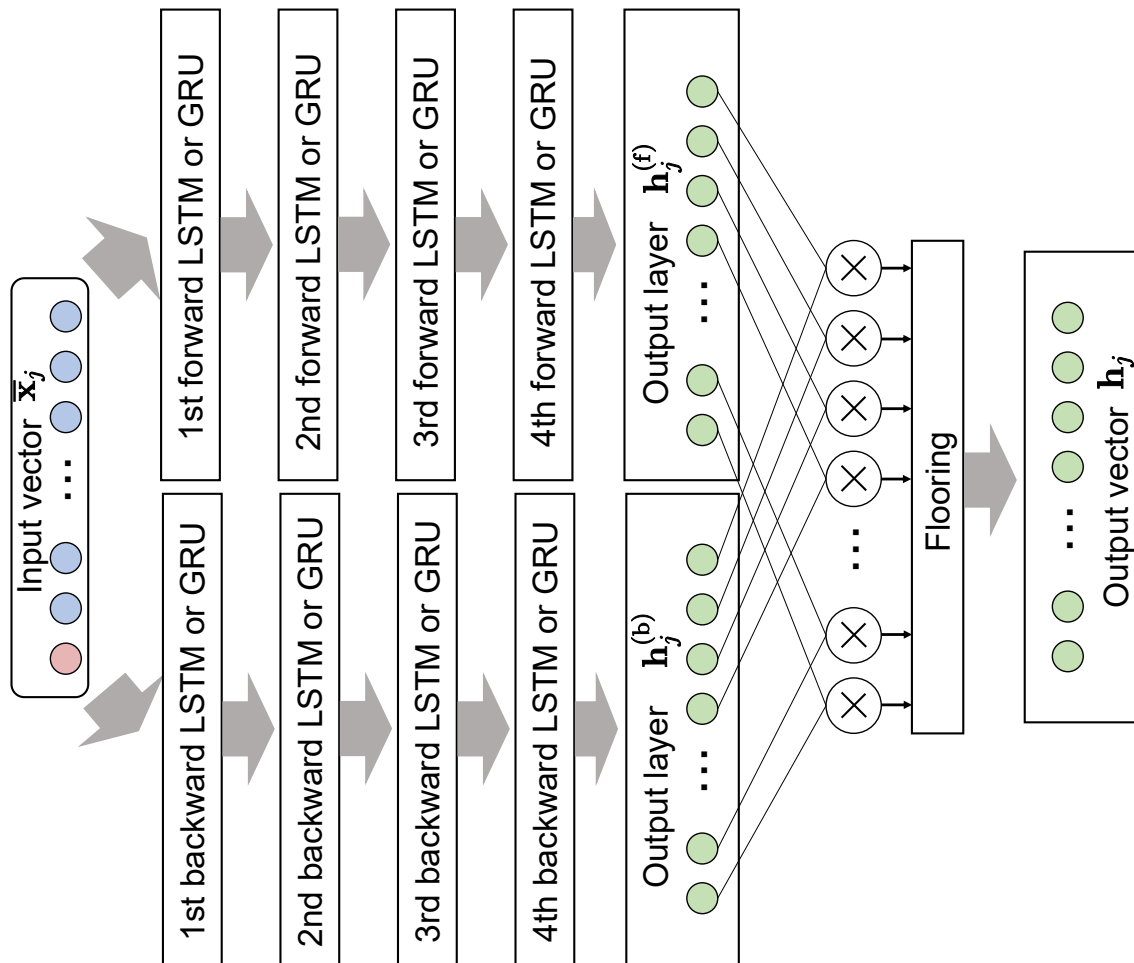
$$\tilde{\mathbf{c}}_j = \tanh(\mathbf{W}^{\tilde{c}} \bar{\mathbf{x}}_j + \mathbf{R}^{\tilde{c}} \mathbf{h}_{j-1} + \mathbf{b}^{\tilde{c}}) \quad (3.14)$$

$$\mathbf{c}_j = \mathbf{f}_j \circ \mathbf{c}_{j-1} + \mathbf{i}_j \circ \tilde{\mathbf{c}}_j \quad (3.15)$$

$$\mathbf{o}_j = \sigma(\mathbf{W}^o \bar{\mathbf{x}}_j + \mathbf{R}^o \mathbf{h}_{j-1} + \mathbf{b}^o) \quad (3.16)$$

$$\mathbf{h}_j = \mathbf{o}_j \circ \tanh(\mathbf{c}_j) \quad (3.17)$$

ここで、 $\sigma(\cdot)$  はベクトルに対するシグモイド関数、 $\mathbf{W}^f$ 、 $\mathbf{W}^i$ 、 $\mathbf{W}^{\tilde{c}}$ 、及び  $\mathbf{W}^o$  は時間  $j$  における入力ベクトル  $\bar{\mathbf{x}}_j$  に対する重み係数行列、 $\mathbf{R}^f$ 、 $\mathbf{R}^i$ 、 $\mathbf{R}^{\tilde{c}}$ 、及び  $\mathbf{R}^o$  は、時間  $j-1$  における出力ベクトル  $\mathbf{h}_{j-1}$  に対する重み係数行列、 $\mathbf{b}^f$ 、 $\mathbf{b}^i$ 、 $\mathbf{b}^{\tilde{c}}$ 、及び  $\mathbf{b}^o$  は、それぞれの係数に対するバイアスベクトル、 $\tanh(\cdot)$  はベクトルに対する双曲線正接関数、 $\circ$  はベクトルの要素毎の積をそれぞれ示す。 $\mathbf{f}_j$  は、忘却ゲートと呼ばれ、時間  $j-1$  における長期記憶ベクトル  $\mathbf{c}_{j-1}$  からの要素を保持するか決定するベクトルである。 $\mathbf{i}_j$  は入力ゲートと呼ばれ、時間  $j$  における入力ベクトル  $\bar{\mathbf{x}}_j$  及び時間  $j-1$  における出力ベクトル  $\mathbf{h}_{j-1}$  からの要素を保持するか決定するベクトルである。 $\mathbf{o}_j$  は出力ゲートと呼ばれ、時間  $j$  における出力ベクトル  $\mathbf{h}_j$  を求めるためのベクトルである。忘却ゲート  $\mathbf{f}_j$  及び入力ゲート  $\mathbf{i}_j$  を用いて、時間  $j$  における長期記憶  $\mathbf{c}_j$  が得られる。これを順方向及び逆方向で行うことで時系列データとして学習する。上記は順方向の場合であり、逆方向の場合、時間  $j$  における長期記憶ベクトル  $\mathbf{c}_j$  及び短期記憶ベクトル  $\mathbf{h}_j$

Fig.3.7. Data flow in the multi-layer BiRNN at time frame  $j$ .

は、時間  $j + 1$  における長期記憶ベクトル  $\mathbf{c}_{j+1}$  及び短期記憶ベクトル  $\mathbf{h}_{j+1}$  を用いて求める。

BiLSTM (及び BiGRU) を多層化した場合の構造を、Fig. 3.7 に示す。但し、簡便のために Fig. 3.5 のある時間フレーム  $j$  に対する入力から出力までの拡大図を示している。本論文では、Fig. 3.7 に示すように、時間  $j$  における入力  $\bar{x}_j$  から 4 つの LSTM ユニットを通して、 $\mathbf{h}_j$  を出力する。この時、BiLSTM では順方向の出力ベクトル  $\mathbf{h}_j^{(f)}$  及び逆方向の出力ベクトル  $\mathbf{h}_j^{(b)}$  が出力され、その要素毎の積を取り、フロアリングを行ったベクトルを時間  $j$  における出力ベクトル  $\mathbf{h}_j$  として扱う。なお、Fig. 3.7 の LSTM ユニットの出力ベクトルの要素数は全て  $I$  と設定している。

### 3.4.3 BiGRU 型 DNN

デコーダとして BiRNN の一種である BiGRU を用いる場合について説明する。入力、出力、及びラベルは 3.4.2 項で述べた BiLSTM の場合と同様である。BiGRU は、BiRNN における

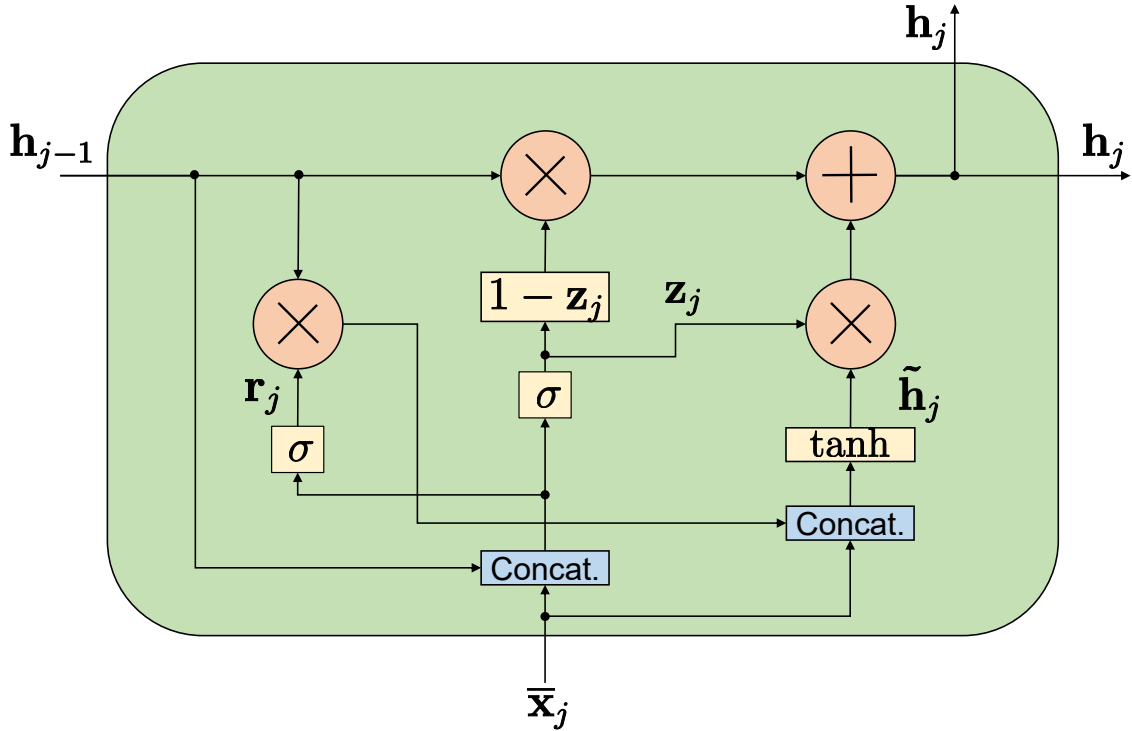


Fig.3.8. Structure of GRU.

双方向の RNN に GRU[27] を用いた構造を持つ DNN である。GRU の内部を Fig. 3.8 に示す。ここで、GRU への入力及び出力をそれぞれ LSTM と同様に  $\bar{\mathbf{X}} \in \mathbb{R}^{D \times J}$  及び  $\mathbf{H} \in \mathbb{R}^{D' \times J}$  と定義し、これらの時刻  $j$  におけるベクトルをそれぞれ  $\bar{\mathbf{x}}_j$  及び  $\mathbf{h}_j$  と定義している。また、 $D$  及び  $D'$  はそれぞれこの GRU に入力された時点での  $\bar{\mathbf{X}}$  の特徴量の次元数及びこの GRU から出力される時点での  $\mathbf{H}$  の特徴量の次元数である。Fig. 3.8 の各変数の計算は次式となる。

$$\mathbf{r}_j = \sigma(\mathbf{W}^r \mathbf{x}_j + \mathbf{R}^r \mathbf{h}_{j-1} + \mathbf{b}^r) \quad (3.18)$$

$$\tilde{\mathbf{h}}_j = \tanh(\mathbf{W}^{\tilde{\mathbf{h}}} \mathbf{x}_j + \mathbf{R}^{\tilde{\mathbf{h}}}(\mathbf{r}_j \circ \mathbf{h}_{j-1}) + \mathbf{b}^{\tilde{\mathbf{h}}}) \quad (3.19)$$

$$\mathbf{z}_j = \sigma(\mathbf{W}^z \mathbf{x}_j + \mathbf{R}^z \mathbf{h}_{j-1} + \mathbf{b}^z) \quad (3.20)$$

$$\mathbf{h}_j = (1 - \mathbf{z}_j) \circ \mathbf{h}_{j-1} + \mathbf{z}_j \circ \tilde{\mathbf{h}}_j \quad (3.21)$$

ここで、 $\mathbf{W}^z$ ,  $\mathbf{W}^r$ , 及び  $\mathbf{W}^{\tilde{\mathbf{h}}}$  は時間  $j$  における入力ベクトル  $\bar{\mathbf{x}}_j$  に対する重み係数行列、 $\mathbf{R}^z$ ,  $\mathbf{R}^r$ , 及び  $\mathbf{R}^{\tilde{\mathbf{h}}}$  は、時間  $j-1$  における出力ベクトル  $\mathbf{h}_{j-1}$  に対する重み係数行列、 $\mathbf{b}^z$ ,  $\mathbf{b}^r$ , 及び  $\mathbf{b}^{\tilde{\mathbf{h}}}$  は、それぞれの係数に対するバイアスベクトルをそれぞれ示す。 $\mathbf{r}_j$  はリセットゲートと呼ばれ、時間  $j-1$  における出力ベクトル  $\mathbf{h}_{j-1}$  からどの要素を保持するか決定するベクトルである。 $\mathbf{z}_j$  は更新ゲートと呼ばれ、時間  $j$  における入力ベクトル  $\bar{\mathbf{x}}_j$  からどの要素を保持するか決定するベクトルである。リセットゲート  $\mathbf{r}_j$  及び更新ゲート  $\mathbf{z}_j$  を用いて、出力ベクトル  $\mathbf{h}_j$  を得る。これを順方向及び逆方向で行うことで時系列データとして学習する。BiLSTM の説明と同様に、上記は順方向の場合であり、逆方向の場合、時間  $j$  における長期記憶ベクトル及び出力ベクトル  $\mathbf{h}_j$  は、時間  $j+1$  における長期記憶ベクトル  $\mathbf{h}_{j+1}$  を用いて求める。



多層化した場合の BiGRU の構造は BiLSTM と同じく Fig. 3.7 のようにあらわされる。各 GRU の出力ベクトルの要素数も、BiLSTM の時と同様に全て  $I$  次元と設定している。

### 3.5 本章のまとめ

本章では、1 章で概要の説明を行った提案音生成システムの詳細な説明と、提案音生成システムの問題点を上げ、解決方法の提案を行った。3.2 節では、提案音生成システムのエンコーダで行う音高、音色、及び音量の抽出方法の説明と複数楽器音の音色抽出に用いる VAE の基本的な形について説明した。3.3 節では、提案音生成システムの問題である音高、音色、及び音量の 3 つの音響特徴量から振幅スペクトログラムの復元できない問題について言及し、解決する方法として非線形な変換が行える MLP 型 DNN、BiLSTM 型 DNN、及び BiGRU 型 DNN を用いたデコーダの作成を提案した。3.4 節では、デコーダとして用いる MLP 型 DNN、BiLSTM 型 DNN、及び BiGRU 型 DNN についての説明を行い、本論文で行う実験で用いた層の構成を説明した。次章では、3.4 節で説明した各 DNN の構成を用いて、音高、音色、及び音量の 3 つの音響特徴量から振幅スペクトログラムの予測する DNN を学習し、テストデータに対する予測精度の評価を行う。

## 第 4 章

# 振幅スペクトログラム予測実験

### 4.1 まえがき

本章では、3章で説明した提案音生成システムの解決すべき問題である DNN に基づくデコーダの実験の詳細及び結果について説明する。4.2 節では、DNN の学習に用いる楽器音データ及び各変換の条件を説明する。4.3 節では、3.4 節及び 4.2 節で説明した実験条件に基づき、各 DNN に対して実験を行い、テストデータに対する振幅スペクトログラムの予測結果を示す。4.4 節では、予測された振幅スペクトログラムの MFCC 相対誤差及び振幅相対誤差を用いて、各 DNN の予測精度の比較評価を行う。4.5 節では、予測された楽器音信号が本物の楽器音信号とどの程度違って聞こえるのか主観評価実験及びその評価を行う。4.6 節では、本章のまとめを行う。

### 4.2 実験条件

本実験で用いる楽器音信号には、neural audio synthesis (NSynth) [33] と呼ばれるデータセットを用いる。Nsynth は、サンプリング周波数が 16 kHz、信号長 4 s の楽器音信号である。11 種類の楽器と 3 種類の演奏方法の楽器音信号が収録されている。データセット数は 305,979 個用意されており、289,205 個の学習データ、12,678 個の検証データ、4,096 個のテストデータに分けられて用意されている。Nsynth の楽器種類毎のデータ数を Table. 4.1 に示す。音高は note number と呼ばれる C-1 (note number: 0) から G9 (note number: 127) までの 128 段階で定義され、Fig. 4.2 に示すように各楽器音で適切な範囲の音源が用意されている。また、各音響信号に適用する STFT 及び MFCC への変換に用いた条件を Table 4.3 に示す。

各 DNN の学習条件を説明する。各 DNN の学習において、損失関数は MSE ロスを用いる。また、学習のエポック数は全ての DNN において 50000 エポックに設定する。DNN の最適化

Table4.1. Number of data for NSynth [33]

Family	Training data	Validation data	Test data	Total
Bass	65,474	2,638	843	68,955
Brass	12,675	886	269	13,830
Flute	8,773	470	180	9,423
Guitar	32,690	2,081	652	35,423
Keyboard	51,821	2,404	766	54,991
Mallet	34,201	663	202	35,066
Organ	34,477	1,598	502	36,577
Reed	13,911	720	235	14,866
String	19,474	814	306	20,594
Synth-lead	5,501	0	0	5,501
Vocal	10,208	404	141	10,753
Total	289,205	12,678	4,096	305,979

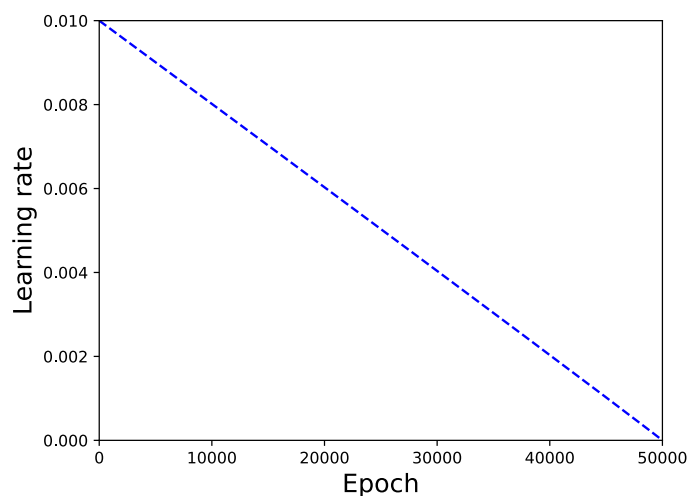


Fig.4.1. Relationship between the learning rate and the number of epochs.

法には Adam[34] を用いる. Adam の重みの最適化アルゴリズムを次式に示す.

$$\mathbf{g}^{(t)} = \nabla \mathcal{L}(\mathbf{w}^{(t)}) \quad (4.1)$$

$$\mathbf{m}_t = \rho_1 \mathbf{m}_{t-1} + (1 - \rho_1) \mathbf{g}^{(t)} \quad (4.2)$$

$$\mathbf{v}_t = \rho_2 \mathbf{v}_{t-1} + (1 - \rho_2) (\mathbf{g}^{(t)})^2 \quad (4.3)$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \rho_1^t} \quad (4.4)$$

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \rho_2^t} \quad (4.5)$$

$$\Delta \mathbf{w}^{(t)} = - \frac{\eta}{\sqrt{\hat{\mathbf{v}}_t + \varepsilon}} \hat{\mathbf{m}}_t \quad (4.6)$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta \mathbf{w}^{(t)} \quad (4.7)$$

Table4.2. Range of pitch for NSynth [33]

Family	Training data		Validation data		Test data	
	Lowest	Highest	Lowest	Highest	Lowest	Highest
Bass	9	109	9	108	10	108
Brass	21	96	24	108	25	108
Flute	21	108	21	108	21	108
Guitar	9	120	21	108	21	108
Keyboard	9	110	21	108	21	108
Mallet	9	120	21	120	21	119
Organ	9	120	9	108	9	108
Reed	21	108	31	97	31	97
String	23	105	24	96	24	96
Synth-lead	9	120	-	-	-	-
Vocal	9	108	21	108	22	108

Table4.3. Experimental conditions used in STFT and MFCC calculations

Window length in STFT	23.2 ms
Shift length in STFT	11.6 ms
Window function in STFT	Hann window
Maximum frequency of mel-filter bank	22.05 kHz
Minimum frequency of mel-filter bank	0.00 kHz
Number of mel-filters ( $K$ )	64

ここで、 $\mathcal{L}(\mathbf{w})$  及び  $\mathbf{w}$  はそれぞれ損失関数及び DNN の最適化変数をまとめたベクトルであり、 $\mathbf{m}_t$  及び  $\mathbf{v}_t$  はいずれも過去の勾配変化を表すモーメントと呼ばれる量である。また、上付き文字の  $t$  は最適化（変数更新）の反復回数を表す。 $\mathbf{m}_t$  及び  $\mathbf{v}_t$  が考慮されていることにより、変数更新における振動を抑えながら高速かつ安定な最適化が可能となる。式 (4.2)–(4.6) 中のハイパーパラメータはそれぞれ標準的な設定値である  $\varepsilon = 1.0 \times 10^{-8}$ ,  $\rho_1 = 0.9$ , 及び  $\rho_2 = 0.999$  に設定する。学習率  $\eta$  は、初期値を 0.01 と設定し、Fig. 4.1 に示すようにエポック数 50000 で 0 となるように線形なスケジューリングを設定する。

本論文の実験では、以下の条件で各 DNN の学習を行った。

- 学習回数は最大 50000 エポックとする
- 全ての DNN で必ず 1000 エポックは学習を行う
- 過去 100 エポック以内に、検証データに対する損失関数の最小値を更新できなければ早

期終了する

さらに、学習中に検証データを用いて求める検証誤差が最小となった際の DNN のパラメータを保持しておき、さらに小さい検証誤差を記録した際にのみこれを更新した。従って、学習終了時に最も検証誤差の小さかった DNN のパラメータを保存した。

## 4.3 実験結果

本節では、3.4 節で説明した各 DNN の条件及び 4.2 節で説明したデータ及び各変換の条件に基づいた実験結果の評価を行う。4.2 節で述べた音高の内、本論文では A4 (note number: 69) の楽器音のみに対して実験を行った予測結果を示す。

Figs. 4.2–4.4 には、最も予測精度の高かった Flute、平均的な精度であった Keyboard、最も予測精度の低かった Mallet の結果をそれぞれ示している。また、各結果の図の (a), (b), (c), 及び (d) は、入力音響信号の振幅スペクトログラム、MLP の予測結果、BiGRU の予測結果、及び BiLSTM の予測結果をそれぞれ示している。なお、これらの楽器音以外の結果については付録 A にまとめて示す。

まず、Fig. 4.2 は、テストデータの中で最も高精度な予測ができた Flute の結果である。MLP の予測結果は、調波構造及び振幅の減衰を予測できている。しかし、予測結果に零値 (Fig. 4.2(b) における黒) が多く見られ、この零値は時間波形に変換した際に人工的な歪みとして残り、音質の劣化につながってしまう。この零値が多く見られる原因として、MLP の第 4 層の重みパラメータが  $512 \times 52275 = 26764800$  と多く、正確なパラメータの更新が行えていない可能性があげられる。それに対して、BiGRU 及び BiLSTM は調波構造及び振幅の減衰を高精度に予測しており、MLP に比べて零値が少ないため予測精度が高いといえる。BiGRU 及び BiLSTM を比較すると、周波数 0.2 kHz 以下の予測は BiGRU の方が精度が高く、周波数 1.3 kHz 以上の予測は BiLSTM の方が精度が高いと言える。

次に、Fig. 4.3 は、テストデータの中でも平均的な予測精度であった Keyboard の結果である。MLP は、周波数 0.2–0.3 kHz で入力音響信号の振幅スペクトログラムに存在しない周波数の振幅を予測している。また、この Keyboard の特徴である振幅の増減や高周波成分の振幅ほど減衰が早いことを予測できていないことがわかる。それに対して、BiGRU 及び BiLSTM は調波構造及び振幅の減衰を高精度に予測しており、また前述したこの Keyboard の特徴も予測できている。BiGRU 及び BiLSTM を比較すると、周波数 0.4 kHz 以下及び 1.3–1.7 kHz において、BiGRU は入力音響信号の振幅スペクトログラムに存在しない振幅を予測しているが、BiLSTM は入力音響信号の振幅スペクトログラムに存在しない振幅があまり見られないことから BiLSTM の方が予測精度が高いといえる。

最後に、Fig. 4.4 は、テストデータの中で最も予測精度が低かった Mallet の結果である。MLP は、周波数構造及び振幅の減衰を全く予測できていない。それに対して、BiGRU 及び BiLSTM は、ある程度周波数構造及び振幅の減衰を予測できている。BiGRU 及び BiLSTM を比較すると、周波数 0.4 kHz 以下において、BiGRU は入力音響信号の振幅スペクトログラ

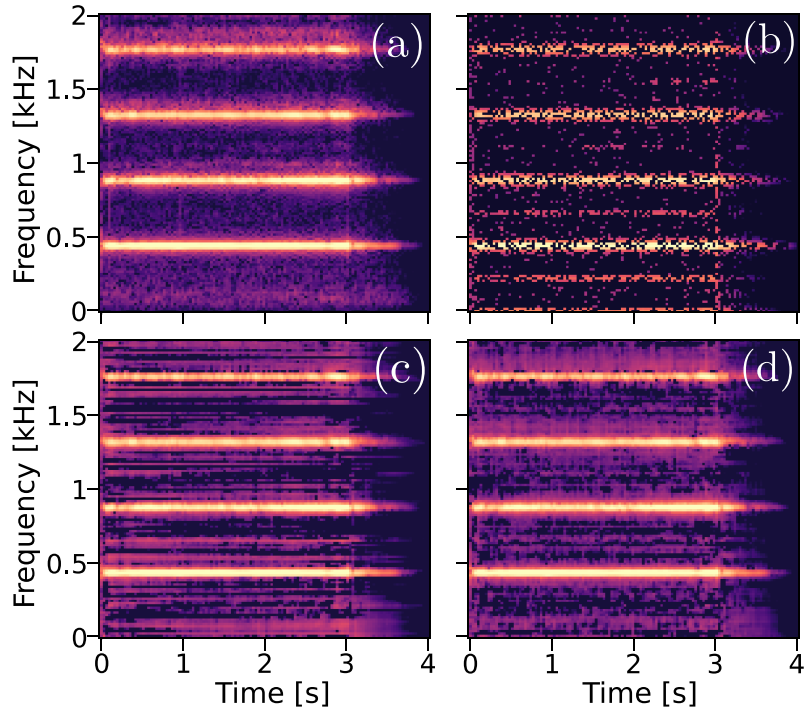


Fig.4.2. Example of spectrograms for test data (acoustic flute sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

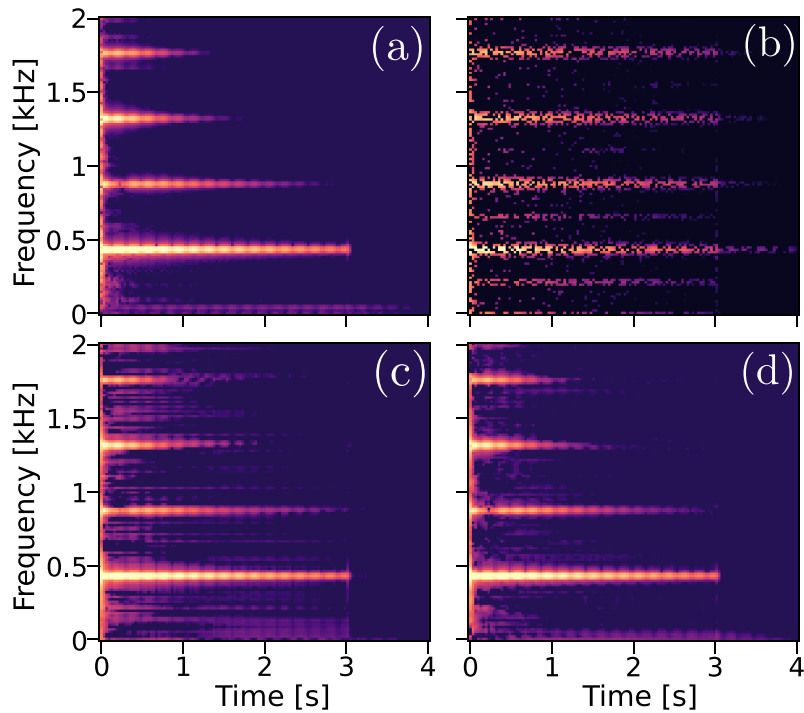


Fig.4.3. Example of spectrograms for test data (synth. keyboard sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

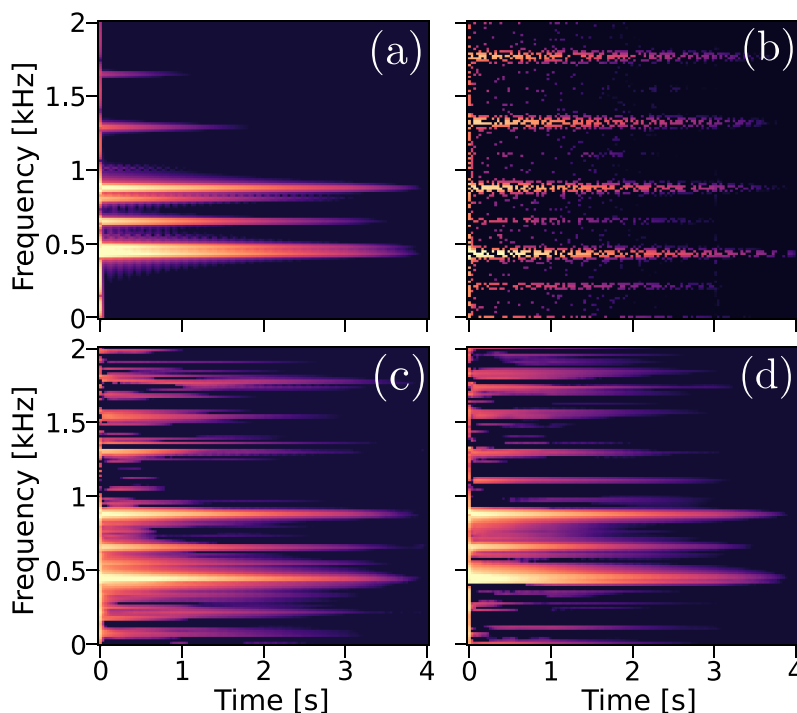


Fig.4.4. Example of spectrograms for test data (acoustic mallet sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

ムに存在しない振幅を予測しているが、BiLSTM は入力音響信号の振幅スペクトログラムに存在しない振幅があまり見られないことから BiLSTM の方が予測精度が高いといえる。しかし、Figs. 4.2 及び 4.3 の Flute 及び Keyboard の各結果と比較すると、BiGRU 及び BiLSTM での予測精度が低いといえる。これは、Mallet は打楽器であり、その音色は複雑な周波数構造と時間推移を持っていることが原因であると考えられる。

#### 4.4 振幅相対誤差及び MFCC 相対誤差に基づく実験評価

前節では、入力と予測結果のパワースペクトログラムを比較し評価及び考察を行った。本節では、予測された振幅スペクトログラムに対する予測精度の客観的評価を行う。本論文では、客観評価指標として、振幅相対二乗誤差 (amplitude relative squared error: ARSE) 及び MFCC 相対二乗誤差 (MFCC relative squared error: MRSE)[35] を用いる。

まず ARSE は、入力に用いた音響信号から直接計算される振幅スペクトログラムを真値とし、また DNN デコーダから予測される振幅スペクトログラムを推定値として、これらの相対的な二乗誤差を客観評価尺度に用いる。ARSE は次式で表される。

$$\text{ARSE} = 10 \log \frac{\sum_{j=1}^J \sum_{i=1}^I (x_{ij} - \hat{x}_{ij})^2}{\sum_{j=1}^J \sum_{i=1}^I (x_{ij})^2} \text{ [dB]} \quad (4.8)$$

次に MRSE は、入力に用いた音響信号から直接計算される MFCC を真値とし、また DNN デコーダから予測される振幅スペクトログラムの MFCC を推定値として、これらの相対的な

二乗誤差を客観評価尺度に用いる。MRSE は次式で表される。

$$\text{MRSE} = 10 \log \frac{\sum_{j=1}^J \sum_{k=2}^{14} (c_{kj} - \hat{c}_{kj})^2}{\sum_{j=1}^J \sum_{k=2}^{14} (c_{kj})^2} [\text{dB}] \quad (4.9)$$

ここで、 $k = 1, 2, \dots, K$  は MFCC の次元のインデックスを示す。また、 $c_{kj}$  及び  $\hat{c}_{kj}$  はそれぞれ  $C$  及び  $\hat{C}$  の要素である。MFCC は 1 次元目が音量を表す特徴量であり、高次元になるほど情報量が減衰する。そこで MRSE では、最も音色の情報を含んでいる次元である MFCC の 2 から 14 次元を用いて計算する。ARSE 及び MRSE は値が小さいほど、予測精度が高いことを示す。ARSE は、入力音響信号の振幅スペクトログラムと生成した振幅スペクトログラムの誤差を求めることはできるが、提案音生成システムの本質である音色の類似度を求めるものではない。そこで、ARSE に加えて、音色の類似度のみを求めることのできる MRSE を用いることで、振幅スペクトログラムの予測精度が高く、音色を最も再現できる DNN デコーダを調査する。

4.3 節で示した結果を含め、4.2 節で説明した NSynth に含まれる A4 の 53 個のテストデータに対して、各 DNN デコーダで予測を行い ARSE 及び MRSE を算出した。算出した ARSE 及び MRSE の楽器毎の平均値をそれぞれ Figs. 4.5(a) 及び (b) に示す。まずは、DNN ごとの比較を行う。ARSE の結果では、MLP は BiGRU 及び BiLSTM に比べると振幅スペクトログラムの予測精度が低く、BiGRU は BiLSTM と同様もしくは少し劣る振幅スペクトログラムの予測精度であるという評価を得ている。それに対して、MRSE の結果では、MLP は BiGRU 及び BiLSTM に比べると音色の予測精度が低く、BiGRU の音色の予測精度は MLP 及び BiLSTM の中間的なものであるという評価を得ている。このことから以下のことがわかる。

- MLP は振幅スペクトログラムの予測精度は低く、音色の予測精度も他の DNN デコーダには劣る
- BiGRU は振幅スペクトログラムの予測精度は高いが、音色の予測精度は高いとはいえない
- BiLSTM は振幅スペクトログラムの予測精度及び音色の予測精度のどちらも最も高い

次に、楽器ごとの比較を行う。まず、4.3 節で取り上げた Flute は、振幅スペクトログラムの予測精度は高いが、音色の予測精度は高いとはいえない。次に、4.3 節で取り上げた Keyboard は、振幅スペクトログラムの予測精度は高くないが、音色の予測精度は高いといえる。最後に、4.3 節で取り上げた Mallet は、両評価指標で予測精度が低いと評価されていることが分かる。

これらの客観的評価の結果から、Mallet のような特徴的な周波数構造を持つ楽器音を除く楽器音は、BiLSTM を DNN デコーダとして用いることで音色を再現した振幅スペクトログラムの予測が可能であるということが分かった。



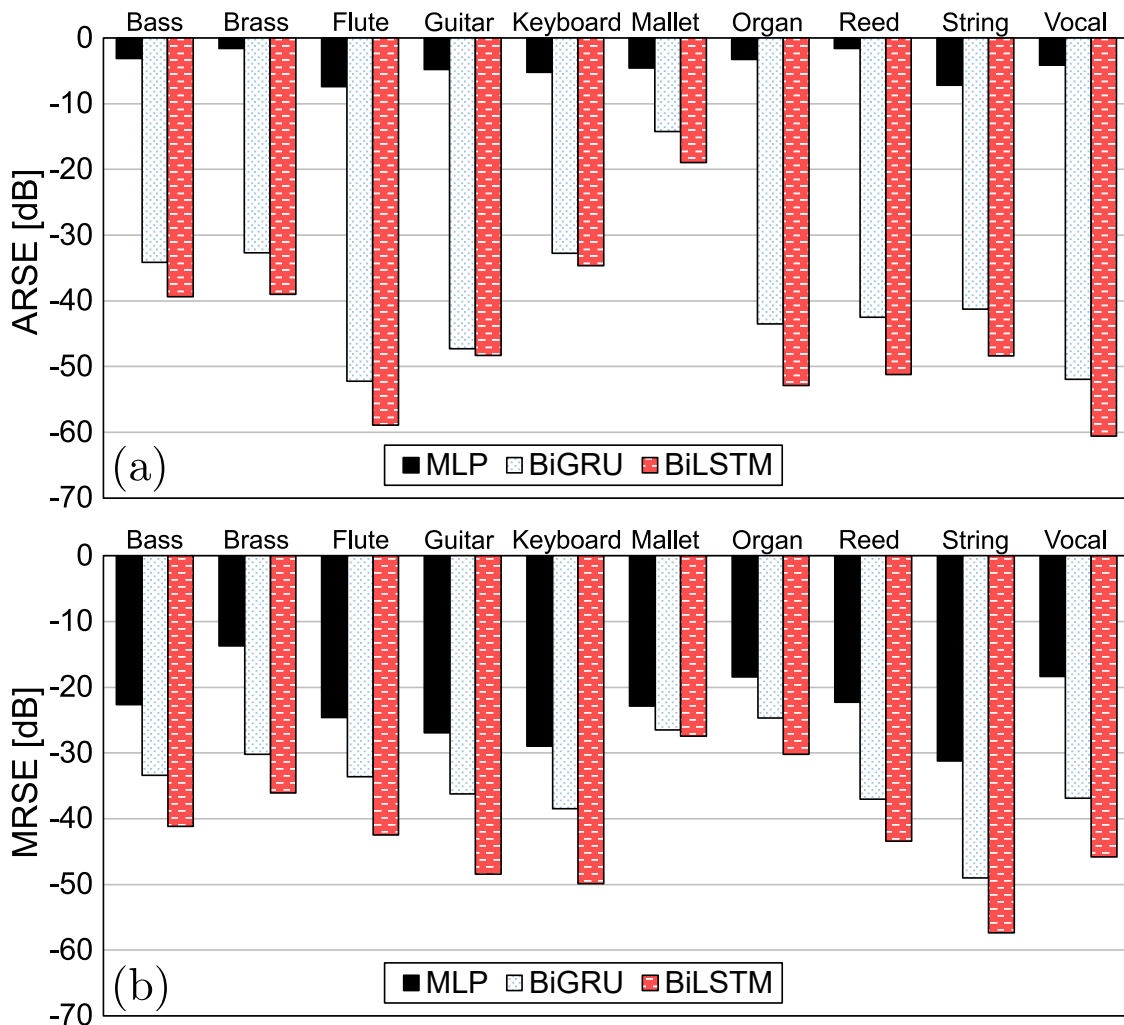


Fig.4.5. (a)ARSE and (b)MRSE results averaged over test data for each musical instrument.

## 4.5 予測楽器音の主観的評価実験

前節では、予測された振幅スペクトログラムに対する予測精度の客観的評価を行い、各 DNN デコーダで予測精度に差があることが分かった。しかし、提案音生成システムの目的は、音楽という人間の感性と関連する芸術につながるものであるため、各 DNN デコーダの予測精度は実際に人間が聞いたうえで評価するべきである。そこで本節では、被験者による主観的評価実験での予測精度の評価を行う。

主観評価実験を行う目的として、下記がある。

- 実際に人間が聞いたうえで違和感のない楽器音を DNN デコーダが生成できるのかを評価する
- 客観評価人間の感性に合った評価が行えていたのかを確認する

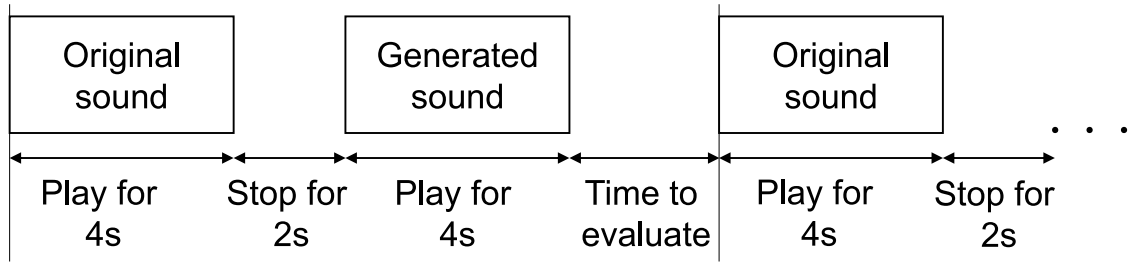


Fig.4.6. Evaluation flow of DMOS test.

Table4.4. Evaluation opinion and its value used in DMOS test

Evaluation opinion	Evaluation value
Degradation is inaudible. (劣化を全く感じない)	5
Degradation is audible but not annoying. (劣化を認められるが気にならない)	4
Degradation is slightly annoying. (劣化がわずかに気になる)	3
Degradation is annoying. (劣化が気になる)	2
Degradation is very annoying. (劣化が非常に気になる)	1

- BiGRU 型 DNN 及び BiLSTM 型 DNN のどちらが高精度な予測が可能かを評価する

そこで本論文では、客観的評価手法として degradation mean opinion score (DMOS) 法 [36] を用いる。DMOS 法とは、Fig. 4.6 に示すように、先に本物の音響信号を被験者に提示し、次に予測した音響信号を被験者に提示することで、予測した音響信号が本物の音響信号からどの程度劣化しているかを Table 4.5 に示す表の評価値を用いて評価する方法である。

主観的評価実験では、4.4 節の客観的評価実験で高い評価を得た BiGRU 及び BiLSTM のみを対象とし、再生する音響信号は 4.4 節で評価した音響信号を用いる。再生する音響信号は、本物の音響信号 53 個に対して、「本物の音響信号と BiGRU が予測した音響信号を比較する音源」及び「本物の音響信号と BiLSTM が予測した音響信号を比較する音源」を用意し、合計 106 個の主観評価用音源を被験者に提示する。この時、主観評価用音源は本物の音響信号と予測した音響信号の間に Fig. 4.6 のように 2 秒の無音区間を挿入した。さらに、主観評価実験において主観評価用音源数の半分である 53 個の評価が終了した時、5 分間の休憩を取るようにした。

主観的評価実験は、MATLAB で作成した GUI アプリケーションを用いて行った。以下に実験手順を示す。

1. MATLAB の GUI アプリケーションのファイルを実行する
2. Fig. 4.7(b) に示す画面が表示され、被験者は Fig. 4.7(a) に示すインストラクションを読む
3. Fig. 4.7(b) のテキストボックスにユーザーネームを入力し、音量調整を行ったうえで開始ボタンを押下する
4. Fig. 4.7(c) に示す画面へ遷移し、被験者は再生ボタンを押下し、音源を視聴する
5. 視聴後に Fig. 4.7(d) の画面へ遷移し、被験者は評価を行う
6. 手順 4 及び手順 5 を指定の問題数繰り返す
7. 問題数の半分 (53 問) 終了時点で 5 分間の休憩を取る
8. ウィンドウを閉じて終了する

このとき、Fig. 4.7(b) の画面で入力したユーザーネームをシード値とした擬似乱数を生成し、106 個の音源を提示する順番を被験者ごとにランダムにしている。本実験では、CREATIVE 社のヘッドフォンである Aurvana Live! HP-AURVN-LV を用い、被験者 15 名 (香川高等専門学校に所属する教職員及び学生、20 代から 40 代の男性 12 名及び女性 3 名) による主観評価実験を行った。

Fig. 4.8 に被験者 15 名の回答を DNN 及び楽器ごとに箱ひげ図で示す。この時、Fig. 4.8 の図下部の (.) はテストデータに含まれる各楽器の数を示しており、各楽器の箱ひげ図は楽器数  $\times 15$  の標本で作製されていることに注意する。(Bass の場合、テストデータに 13 種類の Bass 音源が含まれており、箱ひげ図は  $13 \times 15 = 195$  個の標本から作成されていることを意味する。) まず、実際に人間が聞いたうえで違和感のない楽器音を DNN デコーダが生成できるかを評価する。Fig. 4.8 において、DMOS 値の 5 は本物の音響信号と予測した音響信号の違いが分からないことを示している。そのことから、Flute, String, 及び Vocal は評価値が 5 に近いので、これらの楽器音は人間が聞いたうえで違和感のない予測が行えていることが分かる。それに対して、4.3 節の客観評価実験の結果と同様に Mallet は低い評価となった。また、実験後に被験者に対してヒアリングを行ったところ、「短音の楽器音は、本来の音の後にうなりのようなものがあり、全く異なる楽器音に聞こえた」という意見があった。これらのことから、一部の音響信号を除き、人間が聞いたうえで違和感のない楽器音を DNN デコーダが生成できることが分かった。

次に、客観評価指標が人間の感性に合った評価が行えていたのかを確認する。Figs. 4.9(a) 及び (b) に、主観評価実験結果及び客観評価実験結果の楽器別平均を示す (Fig. 4.9(b) は Fig. 4.5(b) の再掲である)。Figs. 4.9(a) 及び (b) を比較すると、どちらの評価実験でも Mallet が最も低く、String が最も高い評価を得ていることが分かる。また、どちらの評価実験でもほとんどの楽器において BiLSTM 型 DNN の方が BiGRU 型 DNN よりも高い評価を得ていることが分かる。このことから、客観評価指標が人間の感性に合った評価が行えていたことが分かった。

最後に、BiGRU 型 DNN 及び BiLSTM 型 DNN のどちらが高精度な予測が可能かを評価

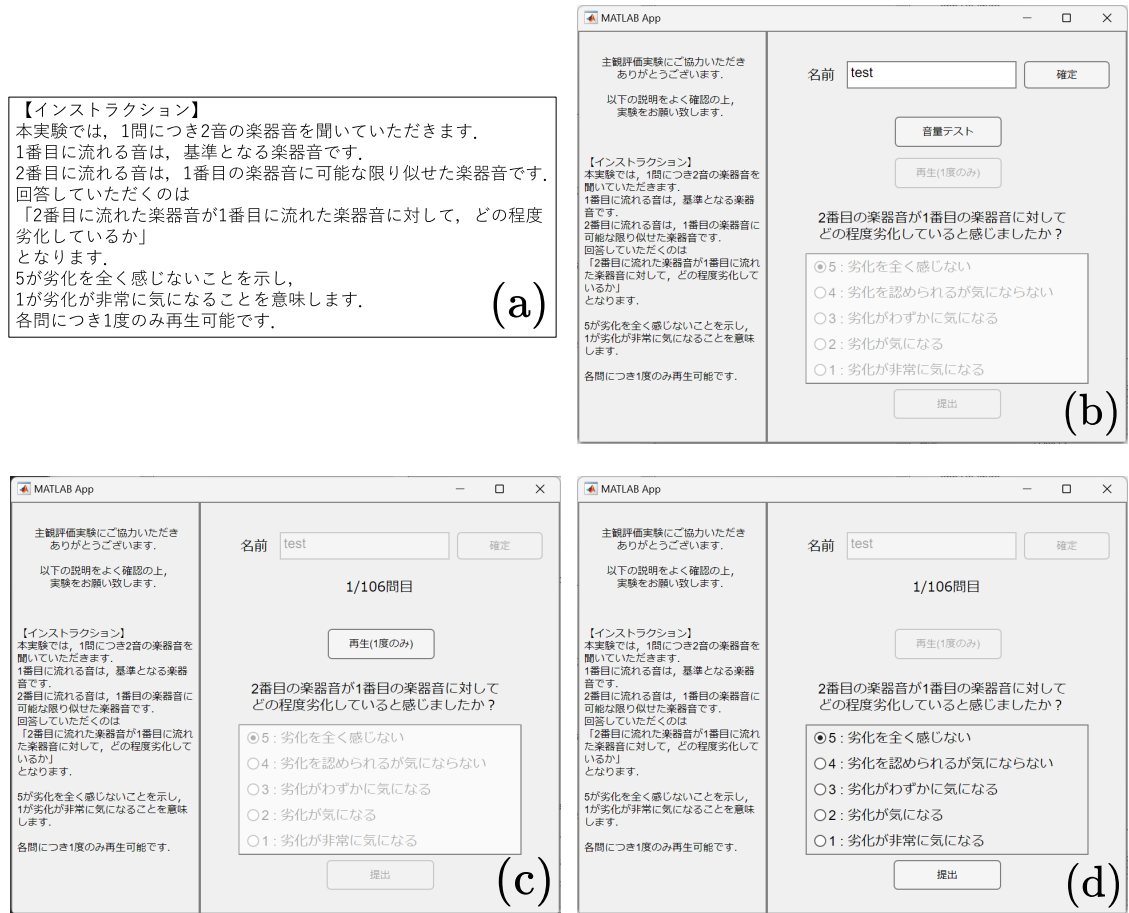


Fig.4.7. MATLAB-based GUI application used in subjective test: (a) instruction presented to subjects, (b) initial setup screen, (c) questioning screen, and (d) answering screen.

する。Figs. 4.9(a) 及び (b) の結果より、主観評価実験及び客観評価実験の両方で BiLSTM 型 DNN の方が BiGRU 型 DNN よりも高精度な予測が可能であることが分かる。

## 4.6 本章のまとめ

本章では、3章で提案音生成システムの問題に対する解決策として上げた、DNN を用いた音高、音色、及び音量の3つの特徴量から振幅スペクトラムの予測実験を行い、テストデータに対する予測精度の評価を行った。4.2節では、実験に用いる音響信号の詳細、STFT 及び MFCC への変換に用いた条件、及び全ての DNN における共通実験条件を示した。4.3節では、テストデータの中で最良、平均的、及び最悪の結果を示し、DNN ごとに比較を行った。4.4節では、ARSE 及び MRSE の2種類の客観的評価指標を用いて客観的評価を行った。結果として、BiGRU 型 DNN 及び BiLSTM 型 DNN を用いることで高精度に振幅スペクトログラムを予測することができることが分かった。4.5節では、主観評価方法として DMOS 法を用い、被験者による主観的評価実験を行った。結果として、BiLSTM 型 DNN が人間が聞

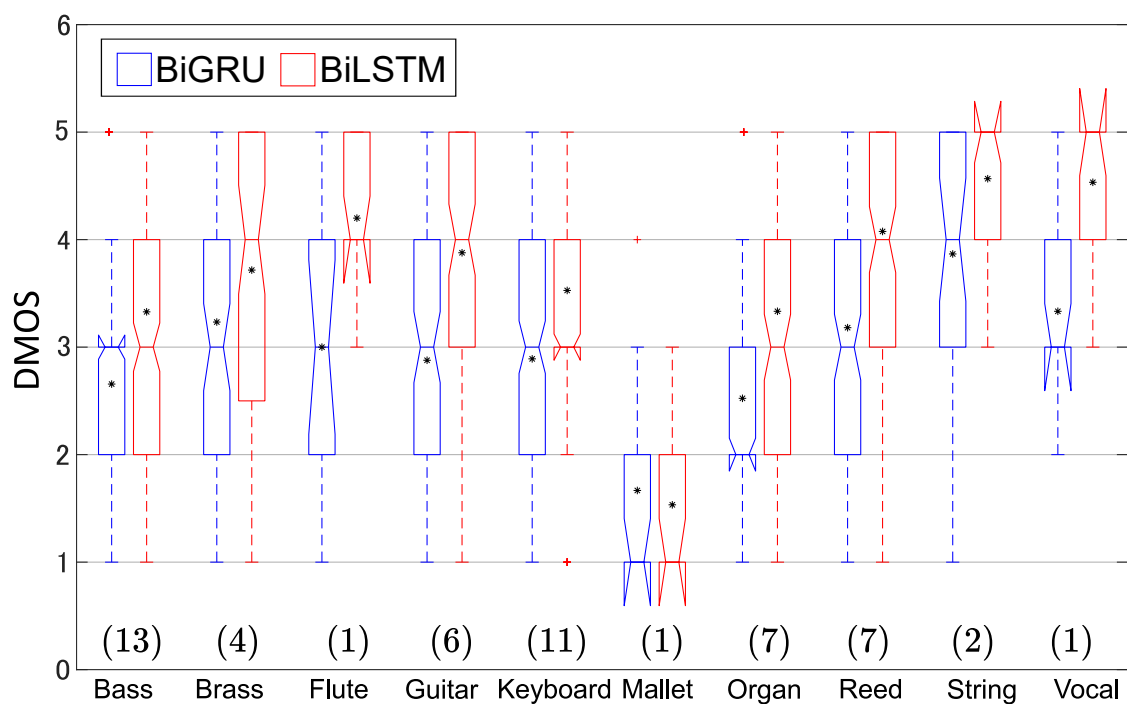


Fig.4.8. Box plot of DMOS values for each musical instrument. Central line of box plot indicates median, asterisk mark in box indicates average, bottom and top edges of box indicates 25th and 75th percentilis, (·) indicates the number of test data for each instrument, and notch of each box indicates difference between the medians of the groups, respectively. If two notches overlap, the true medians of their groups do not differ with 95% confidence.

いたうえで違和感の少ない音響信号を予測できていることに加え、客観的評価指標が人間の感性に沿った評価が行えていることも分かった。また、4.3, 4.4, 及び 4.5 節より BiLSTM 型 DNN を用いることで高精度な振幅スペクトログラムの予測が可能であることが分かった。次章では、本論文で作成した DNN デコーダ及び VAE を用いて、提案音生成システムで各楽器音の特徴量を併せ持つ新しい音響信号が生成可能かを実験的に検証する。

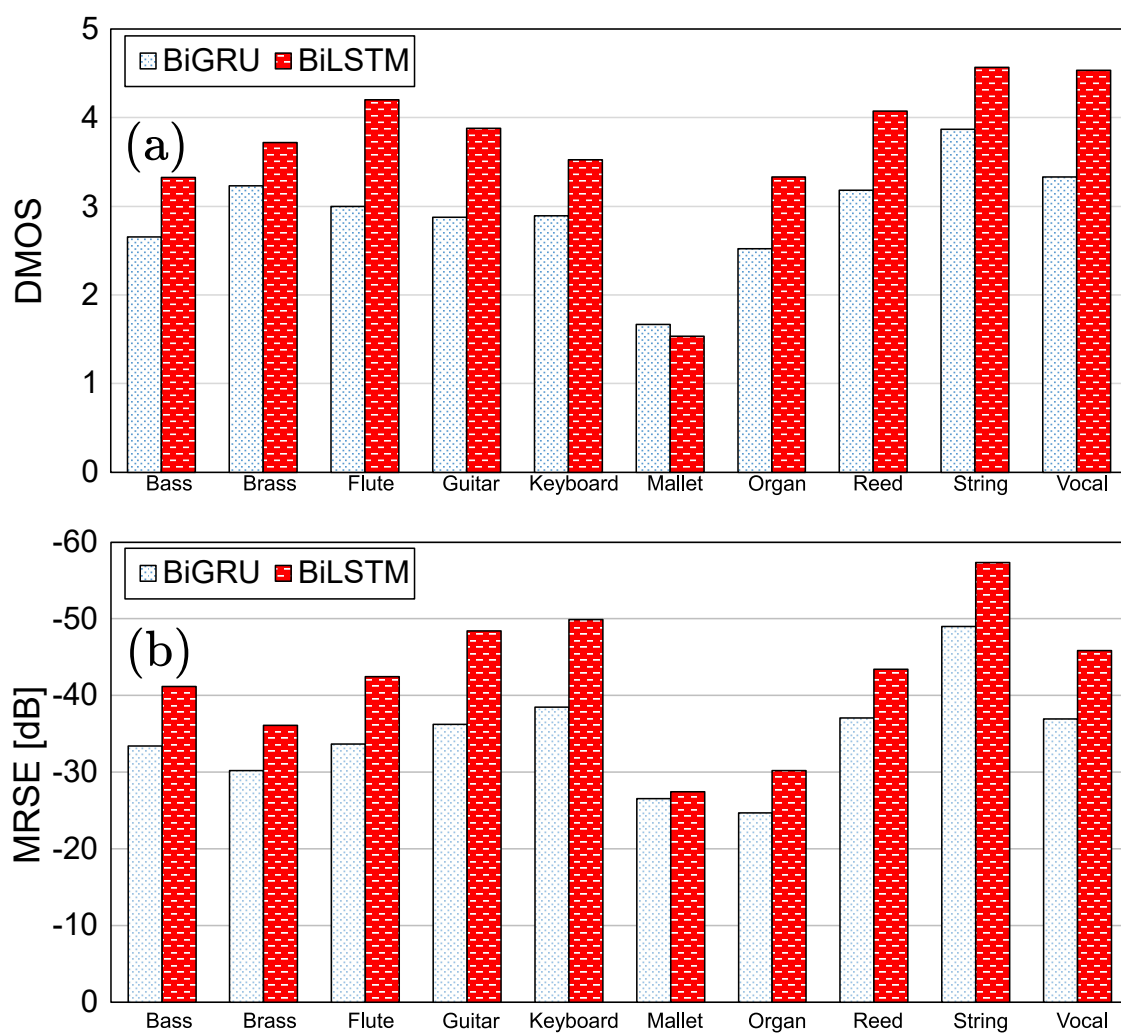


Fig.4.9. (a) DMOS and (b) MRSE results averaged over test data for each musical instrument.

## 第 5 章

# MFCC 及び音響信号生成実験

### 5.1 まえがき

本章では、4章にて作成した DNN デコーダを用いて、3章で説明した提案音生成システムを作成し、音響信号を生成及び評価を行う。5.2 節では、実験で VAE の構造及びデータセットについて説明する。5.3 節では、VAE 及び DNN デコーダを用いて音響信号を生成した結果を示し、評価及び考察を行う。5.4 節では、本章のまとめを行う。

### 5.2 実験条件

本実験では、4章の実験で用いたデータセットである NSynth を用いる。データセットのうち、VAE で学習する MFCC は 4章の実験で用いた A4 (note number: 69) の楽器音のみを用いる。また、音響信号から MFCC への変換は Table 4.3 と同様とする。

本章で学習する VAE の構成について説明する。この VAE は音色特徴量である MFCC を

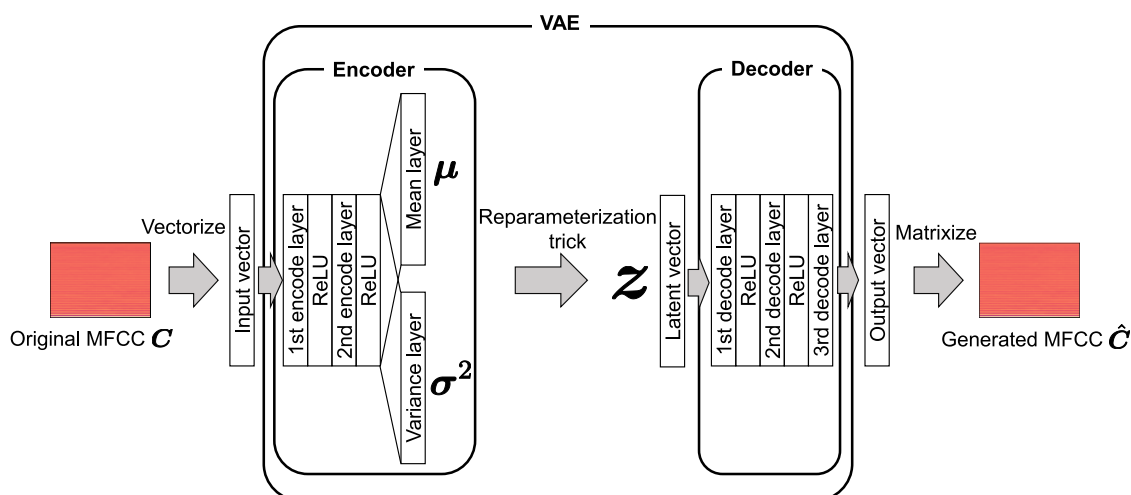


Fig.5.1. Architecture of VAE.

入力とし、その入力の MFCC を再現する構成である。従って、Fig. 1.3 に示すように、この VAE の出力 MFCC を、前章まで議論していた振幅スペクトログラムに変換する DNN へと入力する流れとなる。本論文では、Fig. 5.1 に示す構造の VAE を用いた。層構成としては、入力層、エンコーダ 3 層、デコーダ 3 層、及び出力層の計 8 層からなる。エンコーダの隠れ層の次元数は入力側から 2048, 512 であり、エンコーダの出力層の次元数 ( $\mu$  や  $\sigma^2$  の次元数) は  $D = 16$  と設定する。Fig. 5.1 の Encoder に示すように、エンコーダの出力層は平均  $\mu$  及び分散  $\sigma^2$  を出力するための独立した全結合層が存在する。また、デコーダの各層の隠れ層は入力側から 512, 2048, 及び 8064 と設定した。各層の非線形関数には ReLU を使用した。

次に VAE の学習条件を説明する。VAE の学習において、損失関数には次式に示すものを用いた。

$$\mathcal{L}(C, \hat{C}) = \mathcal{L}_{\text{MSE}}(C, \hat{C}) + \frac{1}{2} \sum_{d=1}^D (\log(\sigma_d^2) - \sigma_d^2 - \mu_d^2 + 1) \quad (5.1)$$

これは (3.9) の第 1 項 (再構成損失) として MSE 損失関数を用いたものである。DNN の最適化法には、4 章と同様に Adam を用いる。ハイパーパラメータは、 $\varepsilon = 1.0 \times 10^{-8}$ ,  $\rho_1 = 0.9$ ,  $\rho_2 = 0.999$ , 及び  $\eta = 0.0001$  に設定した。

また本章の実験では、4 章の実験と同様に、以下の条件で VAE の学習を行った。

- 学習回数は最大 10000 エポックとする
- 全ての DNN で必ず 1000 エポックは学習を行う
- 過去 100 エポック以内に、検証データに対する損失関数の最小値を更新できなければ早期終了する

さらに、学習中に検証データを用いて求める検証誤差が最小となった際の DNN のパラメータを保持しておき、さらに小さい検証誤差を記録した際にのみこれを更新した。従って、学習終了時に最も検証誤差の小さかった DNN のパラメータを保存した。

### 5.3 実験結果

本節では、5.2 節で説明した条件に基づいて作成した VAE 及び 4 章で作成した BiLSTM 型 DNN に基づく DNN デコーダを用いた音響信号生成実験の結果を示し、評価を行う。Figs. 5.3 及び 5.4 の (a) 及び (b) には、入力音響信号の振幅スペクトログラム及び提案音生成システムを通して再現された音響信号の振幅スペクトログラムをそれぞれ示す。

Fig. 5.3 には、String の生成結果を示す。Figs. 5.3(a) 及び (b) を比較すると、調波構造は正確に再現できており、周波数 0.8, 1.3, 及び 1.8 kHz における振幅の減衰はある程度再現できている。しかし、Fig. 5.3(a) では周波数 0.4 kHz の振幅が 1 秒で減衰しきっているのに対して、Fig. 5.3(b) では 1.5 秒まで続いている。さらに、Fig. 5.3(a) の周波数 0.4 kHz 以下において 1.5 秒までに見られる振幅が Fig. 5.3(b) では全く再現できていない。Fig. 5.4 には、Bass の生成結果を示す。Figs. 5.4(a) 及び (b) を比較すると、調波構造はある程度再現できて



いることが分かる．また、この音響信号の特徴である約 0.2 秒間隔で見られる振幅の増減もある程度再現できている．しかし、Fig. 5.4(a) では周波数 0.4, 0.8, 1.3, 及び 1.8 kHz の振幅はほぼ同じなのに対して、Fig. 5.4(b) では周波数が高くなるほど振幅が小さくなっている．さらに、3 秒以降に注目すると、再現精度が 0–3 秒に比べて低いことがわかる．

Figs. 5.3(b) 及び 5.4(b) と付録 A の Figs. A.48(d) 及び A.1(d) に示す 4 章で行った振幅スペクトログラムの予測結果と比較する．Figs. 5.3(b) 及び A.48(d) を比較すると、Fig. 5.3(b) は 1 秒までの振幅の減衰及び周波数 0.4 kHz 以下における振幅の予測精度は Fig. A.48(d) に劣っているといえる．次に、Figs. 5.4(b) 及び A.1(d) を比較すると、Fig. 5.3(b) は周波数 0.4, 0.8, 1.3, 及び 1.8 kHz における振幅の違い及び 3 秒以降の振幅の予測精度は Fig. A.1(d) に劣っているといえる．これらは、VAE での MFCC の生成精度が低いことが原因として考えられる．実際に入力音響信号の MFCC と VAE で再現された MFCC の MRSE を算出したところ、String は  $-31.77$  dB, Bass は  $-33.27$  dB と、VAE での MFCC の生成精度が低いことがわかったため、今後の課題としてあげられる．

次に、MFCC を学習した VAE を用いた複数楽器音の MFCC を併せ持つ楽器音の生成実験を行う．本論文では、Figs. 5.3 及び 5.4 に示した 2 種類の楽器音を用いる．この時、Figs. 5.3 及び 5.4 の音色を Fig. 5.2 のように VAE の潜在空間上で線形補間を行うことで、String から Bass までの音色の推移の生成が可能か検証する．また、String の音色が強い音響信号、Bass の音色が強い音響信号、及び 2 種類の音色を均等に持った音響信号のような様々な条件での音響信号の生成が可能か検証する．

まず、VAE の潜在空間上での音色の線形補間の方法を説明する．Figs. 5.3, 5.4, 及び生成された音響信号の音色の潜在変数をそれぞれ  $\mathbf{z}^{(\text{string})} = [z_1^{(\text{string})}, z_2^{(\text{string})}, \dots, z_D^{(\text{string})}]^T \in \mathbb{R}^D$ ,  $\mathbf{z}^{(\text{bass})} = [z_1^{(\text{bass})}, z_2^{(\text{bass})}, \dots, z_D^{(\text{bass})}]^T \in \mathbb{R}^D$ , 及び  $\mathbf{z}^{(\lambda)} = [z_1^{(\lambda)}, z_2^{(\lambda)}, \dots, z_D^{(\lambda)}]^T \in \mathbb{R}^D$  とすると潜在空間上での線形補間は以下の式で表される．

$$\mathbf{z}^{(\lambda)} = \mathbf{z}^{(\text{string})} \times \lambda + \mathbf{z}^{(\text{bass})} \times (1 - \lambda) \quad (5.2)$$

ここで、 $\lambda$  は生成する音色に含まれる string の音色の割合を表し、本実験では  $\lambda = 1.0$  (100%), 0.8 (80%), 0.6 (60%), 0.4 (40%), 0.2 (20%), 及び 0.0 (0%) の 6 通りを用いて生成を行う．また、DNN デコーダに入力する total amplitude も Figs. 5.3(a) 及び 5.4(a) から式 (3.2) により計算された total amplitude を用いて線形補間を行う．Figs. 5.3(a), 5.4(a), 及び生成された音響信号の total amplitude をそれぞれ  $\mathbf{v}^{(\text{string})}$ ,  $\mathbf{v}^{(\text{bass})}$ , 及び  $\mathbf{v}^{(\lambda)}$  とすると、total amplitude の線形補間は以下の式で表される．

$$\mathbf{v}^{(\lambda)} = \mathbf{v}^{(\text{string})} \times \lambda + \mathbf{v}^{(\text{bass})} \times (1 - \lambda) \quad (5.3)$$

Figs. 5.5(a)–(f) にそれぞれ  $\lambda = 1.0, 0.8, 0.6, 0.4, 0.2$ , 及び 0.0 で生成した振幅スペクトログラムを示す．Figs. 5.5(a) 及び (b) を比較すると、Fig. 5.5(a) では周波数 0.4, 0.8, 1.3, 及び 1.8 kHz の振幅が 1.5 秒までに減衰していることに対して、Fig. 5.5(b) では 3 秒まで振幅の減衰が見られない．また、周波数 0.4, 0.8, 1.3, 及び 1.8 kHz の振幅に注目すると、string

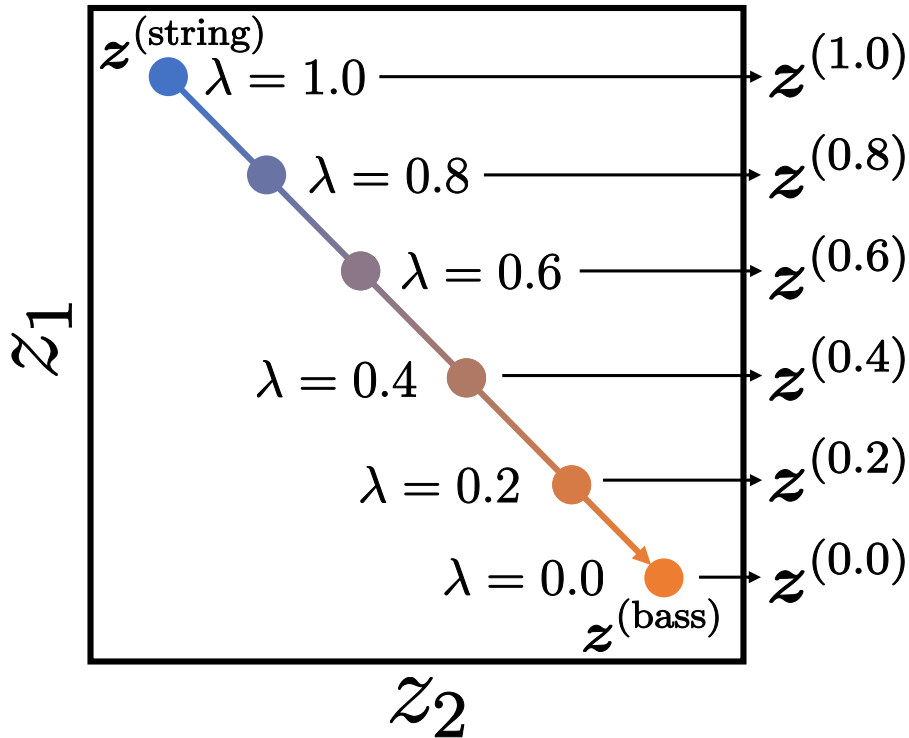


Fig.5.2. Latent variables of timbre transition in the latent space of VAE when  $D = 2$ .

の音色の割合  $\lambda$  が小さくなるにつれて、周波数 0.4, 0.8, 1.3, 及び 1.8 kHz の振幅が大きくなっていることがわかる。原因として、DNN デコーダで振幅スペクトログラムを生成するときに用いる線形補間を行った total amplitude が大きく影響していると考えられる。これは、本実験のように複数楽器音の音色の推移及び音色の合成を行う上では問題ないが、Fig. 1.3 に示す乱数を用いた生成の際に total amplitude をどのように設定するかが問題となる。さらに、string では打音を除いて周波数 0.4, 0.8, 1.3, 及び 1.8 kHz 以外の周波数帯域で振幅を持たないが、string の音色の割合が小さくなるにつれて周波数 0.4, 0.8, 1.3, 及び 1.8 kHz 以外の周波数帯域の振幅が見られるようになった。特に Figs. 5.5(b) 及び (c) の 0.3 kHz 以下の周波数帯域でその変化が顕著に見られる。これらのことから、string の音色の割合を示す  $\lambda$  を用いて、2 種類の楽器音の音色の推移及び様々な条件での音響信号は生成可能であることがわかった。

主観的な印象の参考として、Fig. 5.5 に示した振幅スペクトログラムを inverse STFT 及び位相復元法である Griffin-Lim [37] を用いて時間波形に変換し、試験的に 3 名に聴取させたところ、3 名から共通して「元の楽器音と比較して雑音が大きく、同じ楽器音とは」という意見があった。入力音響信号を  $\mathbf{y} = [y_1, y_2, \dots, y_L]^T \in \mathbb{R}^L$ 、提案音生成システムで再現した音響信号を  $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_L]^T \in \mathbb{R}^L$  とすると、SNR は以下の式で表される。

$$\text{SNR} = 10 \log_{10} \frac{\sum_{l=1}^L |y_l|^2}{\sum_{l=1}^L |\hat{y}_l - y_l|^2} \text{ [dB]} \quad (5.4)$$

SNR は値が大きいほど雑音が少ないことを示す, Figs. 5.5(a) 及び (f) の SNR をそれぞれ算出したところ,  $-0.7181$  dB 及び  $-2.1628$  dB であった. これは, 入力音響信号よりも大きな音量のノイズが再現した音響信号に含まれていることを示している. 原因として, 先述した VAE での MFCC の生成精度が低いことに加え, 振幅スペクトログラムを時間波形に変換する際に用いている位相復元方法の精度が低いことがあげられる. 位相復元方法の改善も提案音生成システムを作成する上で必要であり, その方法として振幅スペクトログラムと同様に位相スペクトログラムを DNN デコーダを用いて予測する方法があげられる.

## 5.4 本章のまとめ

本章では, 3 章で説明した提案音生成システムで各楽器音の特徴量を併せ持つ新しい音響信号が生成可能かを実験的に検証した. 5.2 節では, 実験で用いるデータ及び VAE の学習条件について説明した. 5.3 節では, 入力音響信号の再現実験及び各楽器音の特徴量を併せ持つ音響信号の生成実験を行った結果を示し, 評価を行った. 結果として, 入力音響信号の再現は可能だが, VAE の MFCC 生成精度に大きく依存するため, 生成精度の改善が必要であると分かった. また, 潜在空間上で推移させた音色の特徴量から時間波形を生成すると, 音色が徐々に推移することが確認できた. 次章では, 本論文をまとめる.

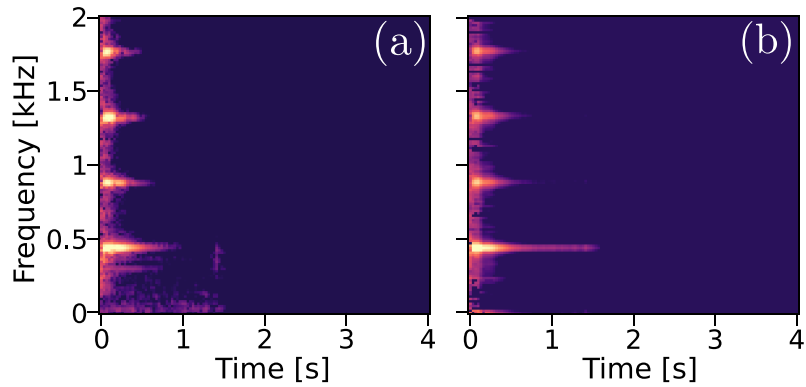


Fig.5.3. Example of spectrograms for test data (acoustic string sound): (a) original and (b) generated by the proposed sound generation system.

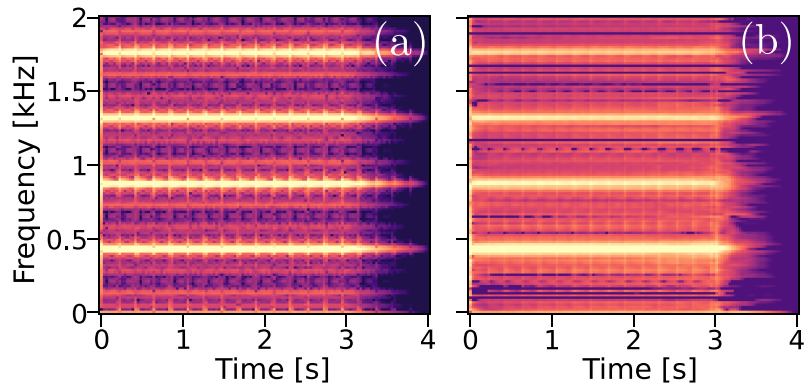


Fig.5.4. Example of spectrograms for test data (synthetic bass sound): (a) original and (b) generated by the proposed sound generation system.

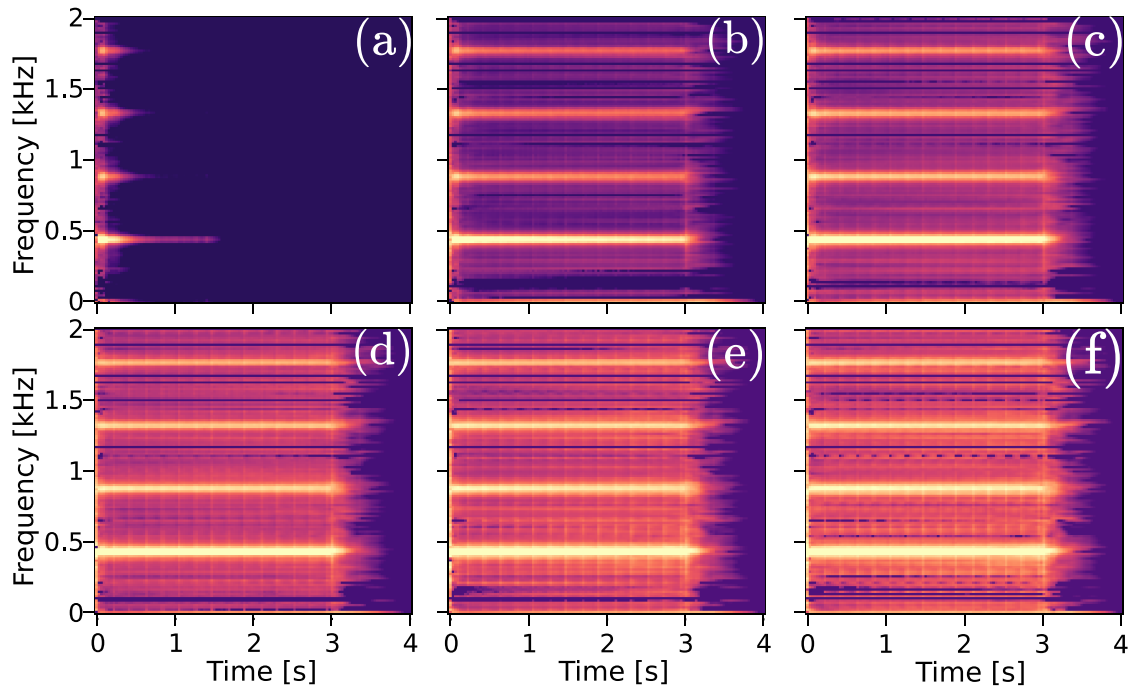


Fig.5.5. Timbre transition of sound generated by DNN decoder using transition of latent variable from string to bass in latent space of VAE: (a)  $\lambda = 1.0$ , (b)  $\lambda = 0.8$ , (c)  $\lambda = 0.6$ , (d)  $\lambda = 0.4$ , (e)  $\lambda = 0.2$ , and (f)  $\lambda = 0.0$ , respectively.

## 第6章

# 結言

本論文では、音色の抽出及び変換を行える提案音生成システムについて説明し、提案音生成システムを実現に必要な部分システムとして、DNNを用いた音高、音色、及び音量の3つの特徴量から振幅スペクトログラムを予測するデコーダを提案した。非線形変換を行えるDNNをデコーダとして用いることで、低次元な情報である音色、音高、及び音量から高次元な情報である振幅スペクトログラムの予測が可能となった。デコーダにはMLP型DNN、BiLSTM型DNN、及びBiGRU型DNNの3種類を用い、テストデータを用いて行った予測実験を行い、2つの客観評価指標を用いた客観評価及び被験者による主観評価実験を行った。結果としては、双方向再帰型DNNであるBiLSTM型DNNを用いた音高、音色、及び音量の3つの特徴量から振幅スペクトログラム予測は高い精度で行えることが分かった。加えて、提案音生成システムを用いた音響信号生成が可能かを実験的に検証した。結果としては、提案音生成システムを用いることで入力音響信号の再現はある程度可能であることが分かった。また、複数楽器音間の音色推移の再現も可能であることが分かった。

最後に今後の課題を述べる。最も優先される課題は、音響信号の生成精度をあげるためにVAEのMFCC生成精度の向上があげられる。方法として、MFCCは時系列データであるため、エンコーダ及びデコーダにRNNを用いたVAEで学習することで性能の改善が可能であると考える。また、別の課題として、振幅スペクトログラムを時間波形に変換する手法の改善も必要である。改善方法としては、振幅スペクトログラムをDNNデコーダで予測するように、DNNを用いた位相スペクトログラム予測が考えられる。

## 謝辞

本論文は、香川高等専門学校電気情報工学科北村研究室にて行われた研究に基づくものです。

まず、本研究を進めるにあたり、ご多忙のところ熱心にご指導くださいました指導教員の北村大地講師に心より感謝申し上げます。北村大地講師には、論文執筆や研究に関する議論など、細部にわたるまで丁寧にご指導いただきました。国内での学会発表をはじめ、国際会議や原著論文執筆など貴重な経験をさせていただくことができました。

本論の副査である村上幸一准教授及び柿元健准教授には、論文の構成及び記述に関して大変有益な助言を頂き、大変お世話になりました。ここに厚く御礼申し上げます。

北村研究室の同期である蓮池氏、溝渕氏、村田氏には深層学習、Python、プログラムの最適化などのアドバイスを頂くなど、様々な面で支えていただきました。また、後輩の綾野氏、加藤氏、鈴木氏、松本氏、和気氏には日頃の会話をはじめとし、様々な面で研究室生活を暖かく支えていただきました。ここに感謝申し上げます。

最後になりますが、現在に至るまで私の学生生活に支え、暖かく見守って下さった両親には感謝の念に堪えません。これまで本当にありがとうございました。

## 参考文献

- [1] D. O’Shaghnessy, “Linear predictive coding,” *Trans. IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.
- [2] F. Zheng, G. Zhang, and Z. Song, “Comparison of different implementations of MFCC,” *Journal of Computer Science and Technology*, vol. 16, pp. 582–589, 2001.
- [3] K. D. Martin, “Sound-Source Recognition: A Theory and Computation Model,” PhD Thesis, Massachusetts institute of technology, 1999.
- [4] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [5] M. S. Nagawade and V. R. Ratnaparkhe, “Musical instrument identification using MFCC,” *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology*, pp. 2198–2202, 2017.
- [6] K. Kashino and H. Murase, “A sound source identification system for ensemble music based on template adaptation and music stream extraction,” *Speech Communication*, vol. 27, pp. 337–349, 1999.
- [7] T. Kitahara, M. Goto, K. Komatani, T. Odata, and H. Okuno, “Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps,” *European Association for Signal Processing Journal on Advances in Signal Processing*, vol. 2007, pp. 1–15, 2007.
- [8] T. Kitahara, M. Goto, K. Komatani, T. Odata, and H. Okuno, “Instrogram: Probabilistic Representation of Instrument Existence for Polyphonic Music,” *Information and Media Technologies*, vol. 2, no. 1, pp. 279–291, 2007.
- [9] D. Bogdanov, M. Haro, F. Fuhrmann, E. Gómez, and P. Herrera, “Content-based music recommendation based on user preference examples,” *Proc. Workshop on Music Recommendation and Discovery*, 2010.
- [10] F. D. Leon and K. Martinez, “Enhancing timbre model using MFCC and its time derivatives for music similarity estimation,” *Proc. European Signal Processing Conference*, pp. 2005–2009, 2012.
- [11] M. Spiertz and V. Gnam, “Source-filter based clustering for monaural blind source separation,” *Proc. International Conference on Digital Audio Effects*, 2009.

- [12] T. Barker and T. Virtanen, “Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation,” *Proc. INTERSPEECH*, pp. 827–831, 2013.
- [13] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, “Differentiable Digital Signal Processing,” *Proc. International Conference on Learning Representations Conference*, 2020.
- [14] I. Goodfellow, J. P.-Abadie, M. Mirza, B. Xu, D. W.-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” *Proc. Advances in Neural Information Processing Systems*, 2014.
- [15] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, “A review on generative adversarial networks: Algorithms, theory, and applications,” *IEEE Trans. Knowledge and Data Engineering*, 2021.
- [16] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” *Proc. International Conference on Machine Learning*, pp. 1530–1538, 2015.
- [17] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, “Normalizing flows: An introduction and review of current methods,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.
- [18] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *Proc. International Conference on Learning Representations*, 2014.
- [19] R. Yu, “A tutorial on VAEs: From Bayes’ rule to lossless compression,” arXiv: 2006.10273, 2020.
- [20] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, “Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7327–7347, 2021.
- [21] P. Esling, A. Chemla-RomenuSantos, and A. Bitton, “Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics,” *Proc. International Conference on Digital Audio Effects*, 2018.
- [22] Y. J. Luo, K. Agres, and D. Herremans, “Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders,” *Proc. International Society for Music Information Retrieval Conference*, 2019.
- [23] Y. J. Luo, K. W. Cheuk, T. Nakano, M. Goto, and D. Herremans, “Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds,” *Proc. International Society for Music Information Retrieval Conference*, pp. 700–707, 2020.
- [24] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii, and S. Morishima, “Pitch-timbre disentanglement of musical instrument sounds based on VAE-based metric learning,”



- IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 111–115, 2021.
- [25] ISO 532-1:2017, “Acoustics – Methods for calculating loudness – Part1: Zwicker method,” 2017.
- [26] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” *IEEE Trans. Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.
- [27] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXiv: 1406.1078, 2014.
- [28] D. O’Shaughnessy, *Speech Communication: Human and Machine*. Addison-Wesley, 1987.
- [29] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “Crepe: A convolutional representation for pitch estimation,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 161–165, 2018.
- [30] L. Hantrakul, J. Engel, A. Roberts, and C. Gu, “Fast and flexible neural audio synthesis,” *Proc. International Society for Music Information Retrieval Conference*, pp 524–530, 2019.
- [31] S. McAdams, S. Winsberg, S. Donnadieu, G. D. Seoete, and J. Krimphoff, “Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes,” *Psychological research*, vol. 58, no. 3, pp. 177–192, 1995.
- [32] S. Lakatos, “A common perceptual space for harmonic and percussive timbres,” *Perception & psychophysic*, vol. 62, no. 7, pp. 1426–1439, 2000.
- [33] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, “Neural audio synthesis of musical notes with WaveNet autoencoders,” *Proc. International Conference on Machine Learning*, pp. 1068–1077, 2017.
- [34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv*, pp. 1412–6980, 2014.
- [35] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, “Minimum mean-squared error estimation of mel-frequency cepstral coefficients using a novel distortion model,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1654–1661, 2008.
- [36] ITU-T Recommendation P.800, “Methods for subjective determination of transmission quality,” 1996.
- [37] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 235–243, 1984.

# 発表文献一覧

## 原著論文

1. Shoya Kawaguchi and Daichi Kitamura, “Amplitude spectrogram prediction from mel-frequency cepstrum coefficients using deep neural networks,” *Journal of Signal Processing*, vol. 27, no. 6, pp. 207–211, 2023.

## 査読付き国際会議

1. Shoya Kawaguchi and Daichi Kitamura, “Amplitude spectrogram prediction from mel-frequency cepstrum coefficients and loudness using deep neural networks,” in Proceedings of *RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP 2023)*, pp. 225–228, 2023.

## 国内学会

1. 川口翔也, 北村大地, “双方向 RNN による MFCC 及びラウドネスからの振幅スペクトログラム予測,” 情報処理学会 第 134 回音楽情報科学研究会 (音学シンポジウム 2022), vol. 2022-MUS-134, no. 60, pp. 1–6, 2022.
2. 川口翔也, 北村大地, “双方向 LSTM によるラウドネス及び MFCC からの振幅スペクトログラム予測と評価,” 日本音響学会 2022 年秋季研究発表会講演論文集, 1-1-16, pp. 1471–1474, 2022.
3. 蓮池郁也, 北村大地, 渡辺瑠伊, 川口翔也, “周波数双方向再帰に基づく深層パーミュテーション解決法,” 電子情報通信学会 第 37 回信号処理シンポジウム, A13-2, pp. 308–313, 2022.
4. 加藤大輝, 川口翔也, 北村大地, “ブラインド音源分離のための単一話者発話区間検出,” 第 26 回 日本音響学会関西支部 若手研究者交流研究発表会, p. 17, 2023.

## 付録 A

# 振幅スペクトラムの予測結果

本付録では、本文で掲載しなかった残りの結果についてまとめて掲載する。Figs. A.1–A.50(a), (b), (c), 及び (d) には、入力音響信号の振幅スペクトログラム、MLP の予測結果、BiGRU の予測結果、及び BiLSTM の予測結果を示す。Figs. A.1–A.13 には bass, Figs. A.14–A.15 には brass, Figs. A.18–A.23 には guitar, Figs. A.24–A.33 Figs. A.34–A.40 には organ, Figs. A.41–A.47 には reed, Figs. A.48–A.49 には string Fig. A.50 には vocal の結果をそれぞれ示す。

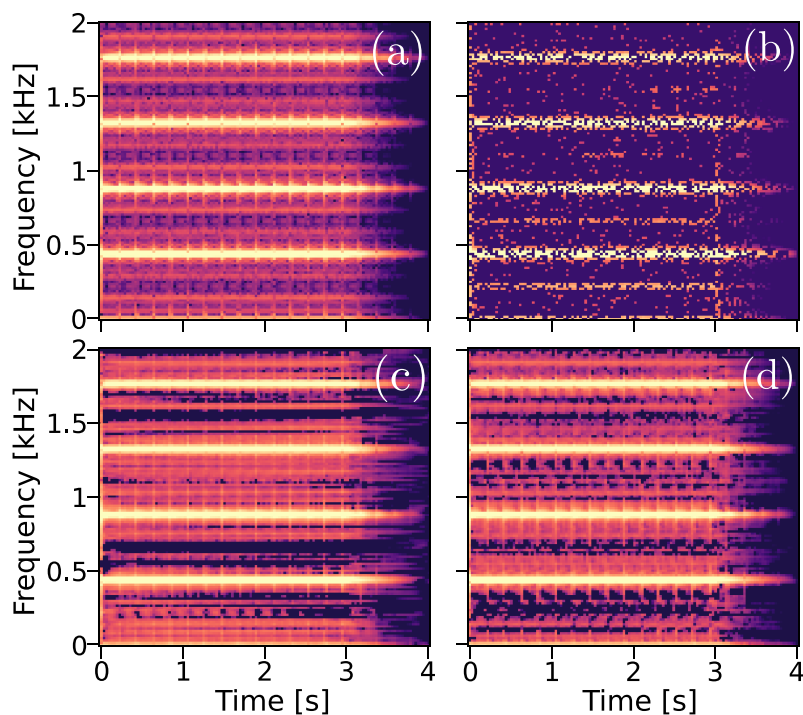


Fig.A.1. Example of spectrograms for test data (synthetic bass1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

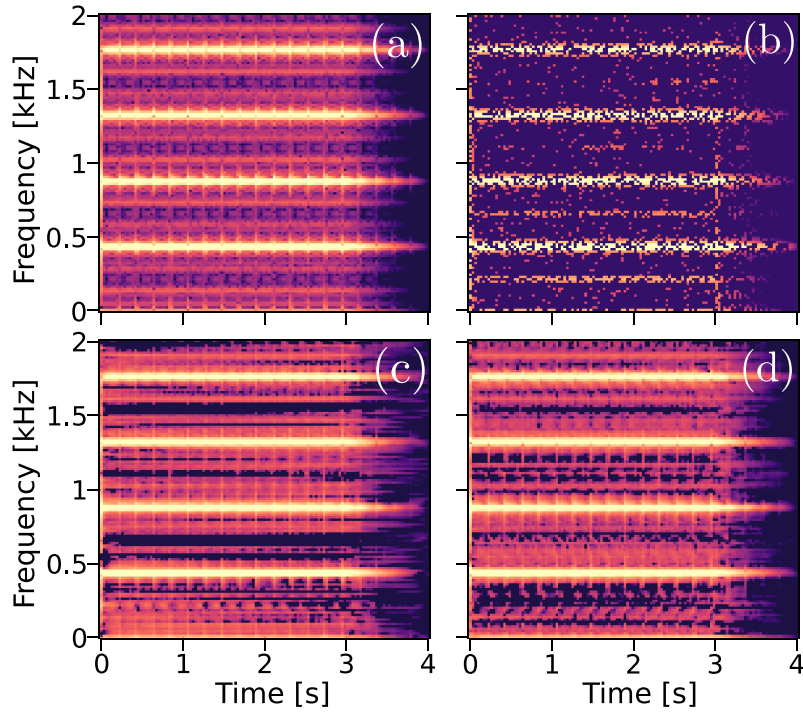


Fig.A.2. Example of spectrograms for test data (synthetic bass2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

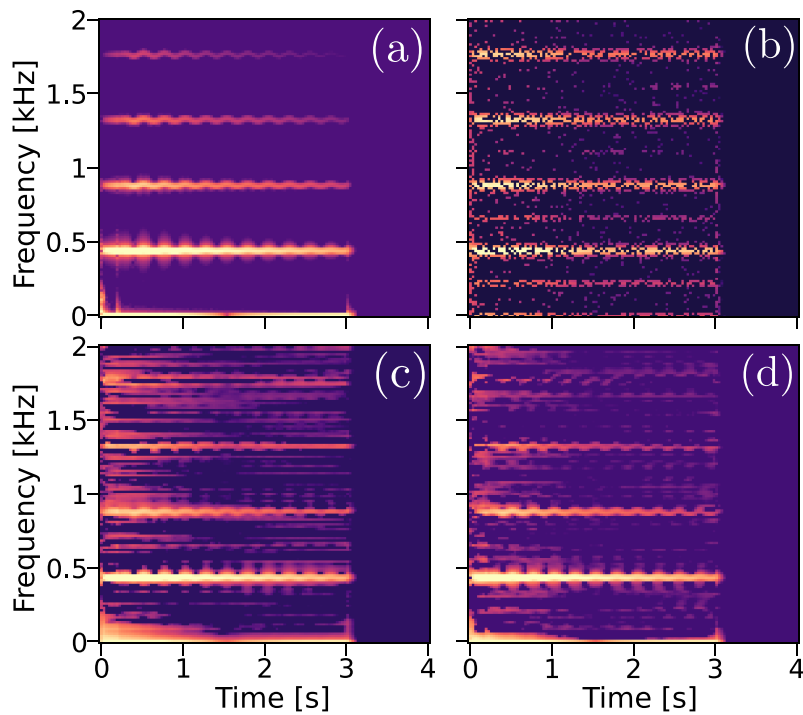


Fig.A.3. Example of spectrograms for test data (synthetic bass3 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

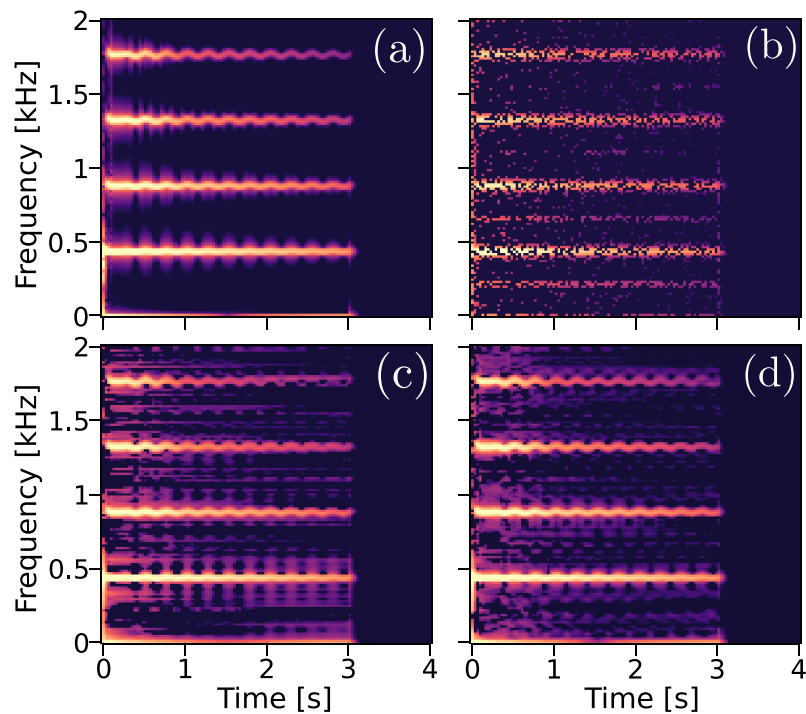


Fig.A.4. Example of spectrograms for test data (synthetic bass4 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

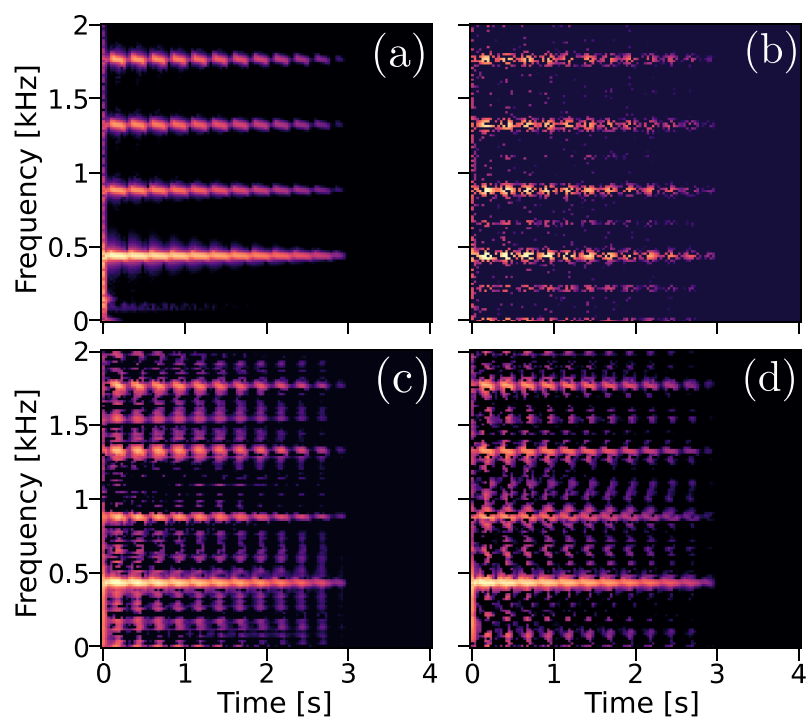


Fig.A.5. Example of spectrograms for test data (synthetic bass5 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

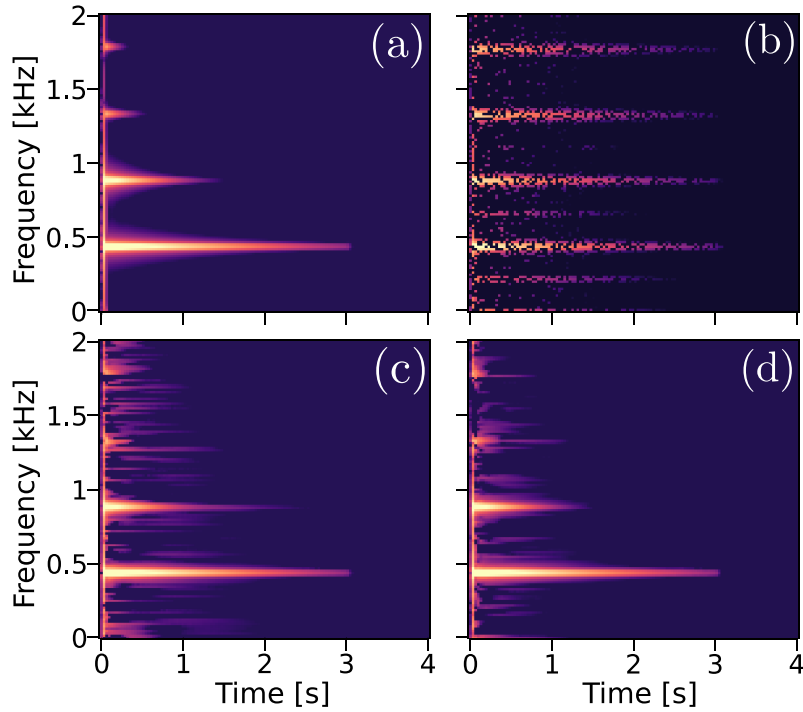


Fig.A.6. Example of spectrograms for test data (synthetic bass6 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

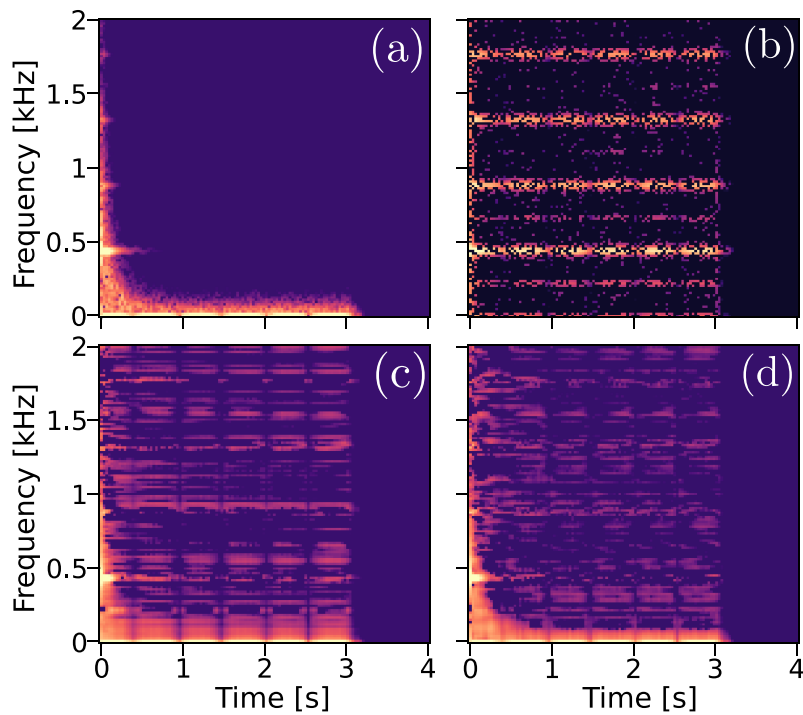


Fig.A.7. Example of spectrograms for test data (synthetic bass7 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

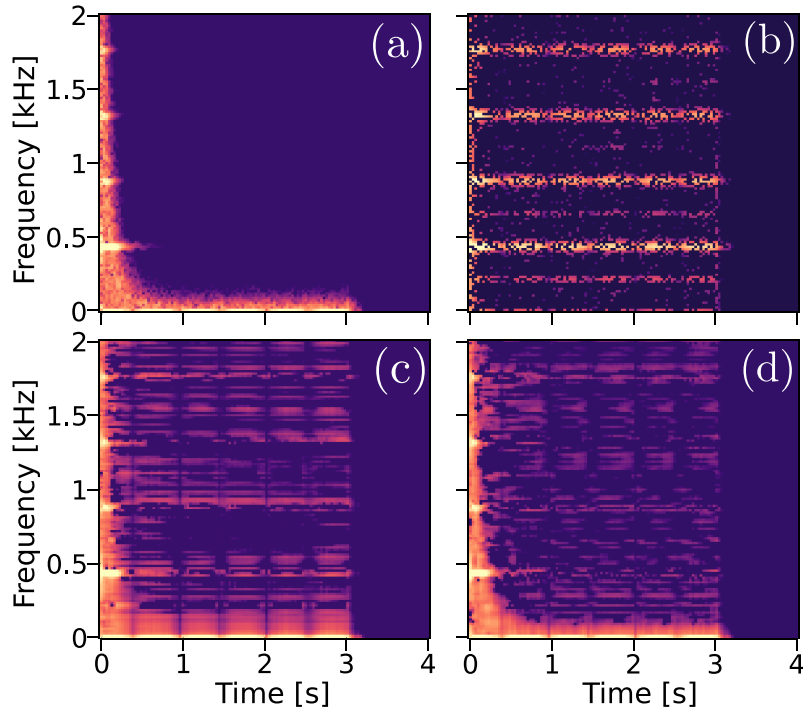


Fig.A.8. Example of spectrograms for test data (synthetic bass8 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

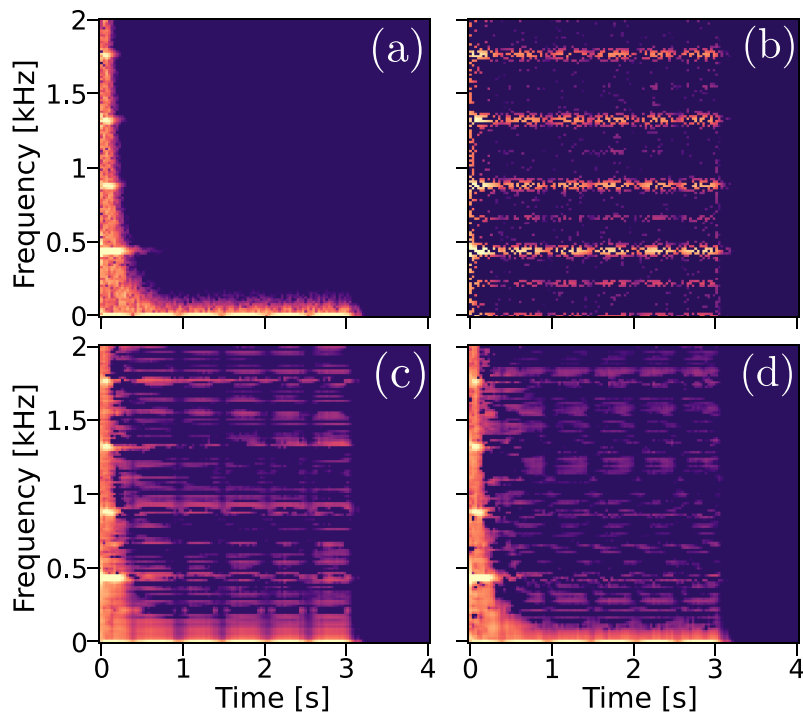


Fig.A.9. Example of spectrograms for test data (synthetic bass9 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

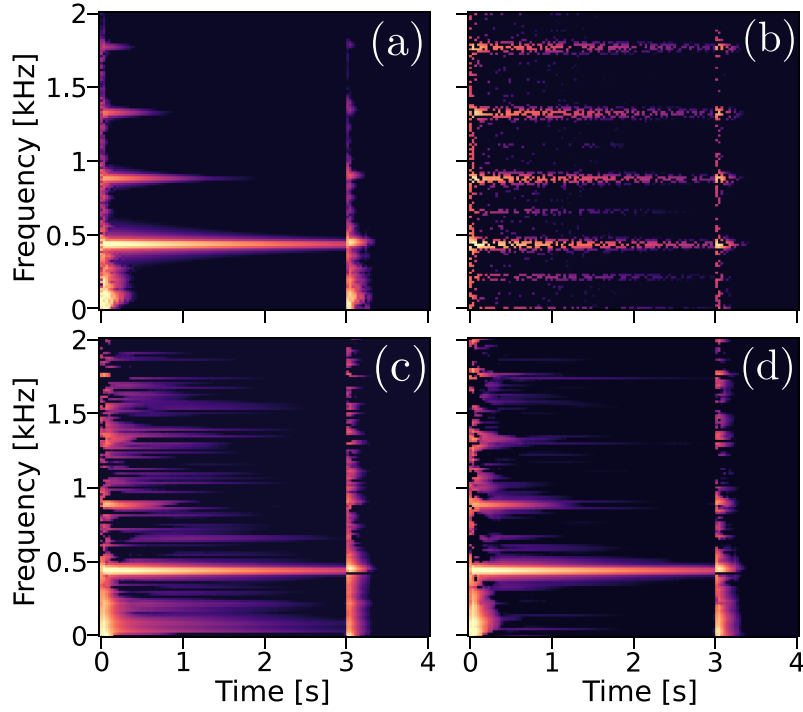


Fig.A.10. Example of spectrograms for test data (synthetic bass10 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

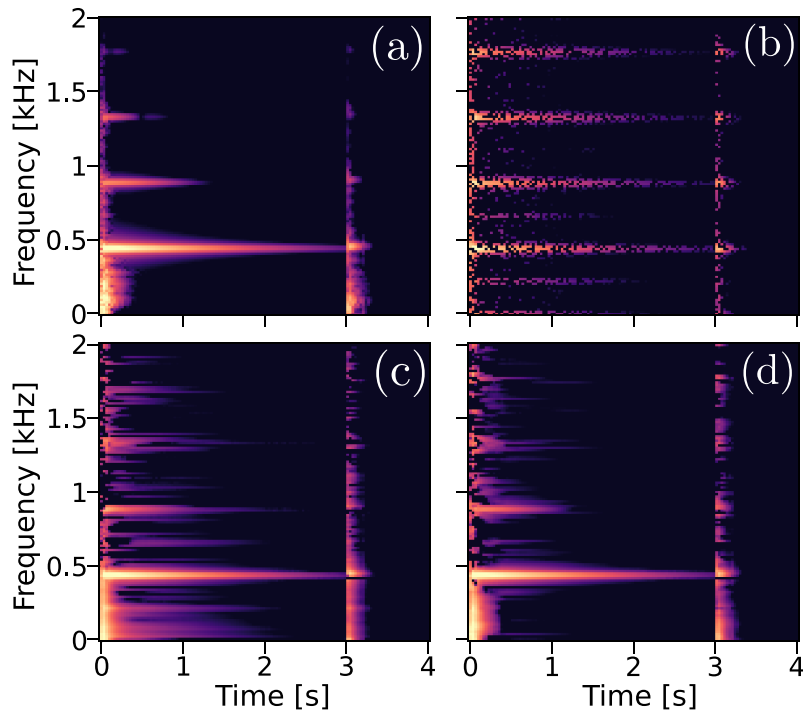


Fig.A.11. Example of spectrograms for test data (synthetic bass11 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.



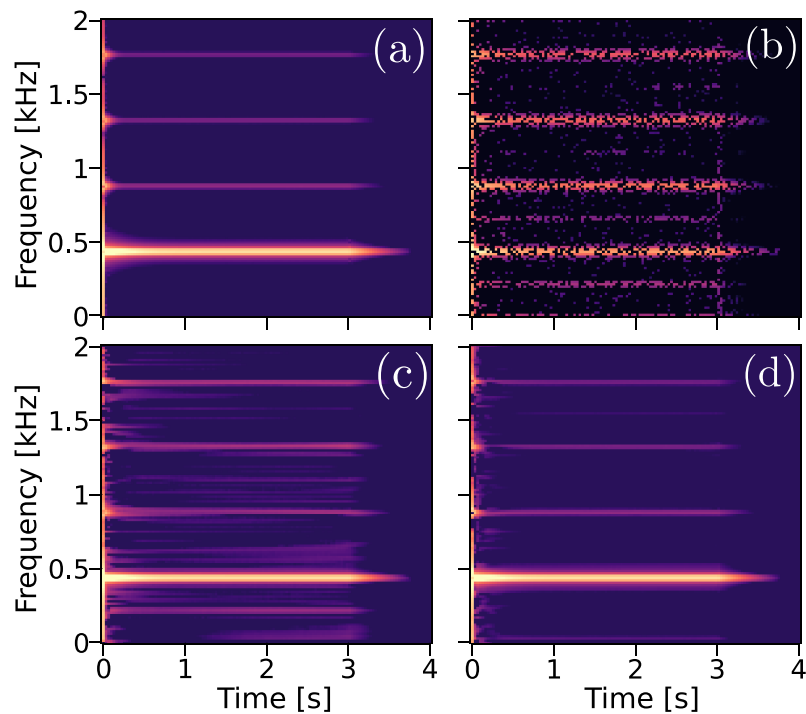


Fig.A.12. Example of spectrograms for test data (synthetic bass12 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

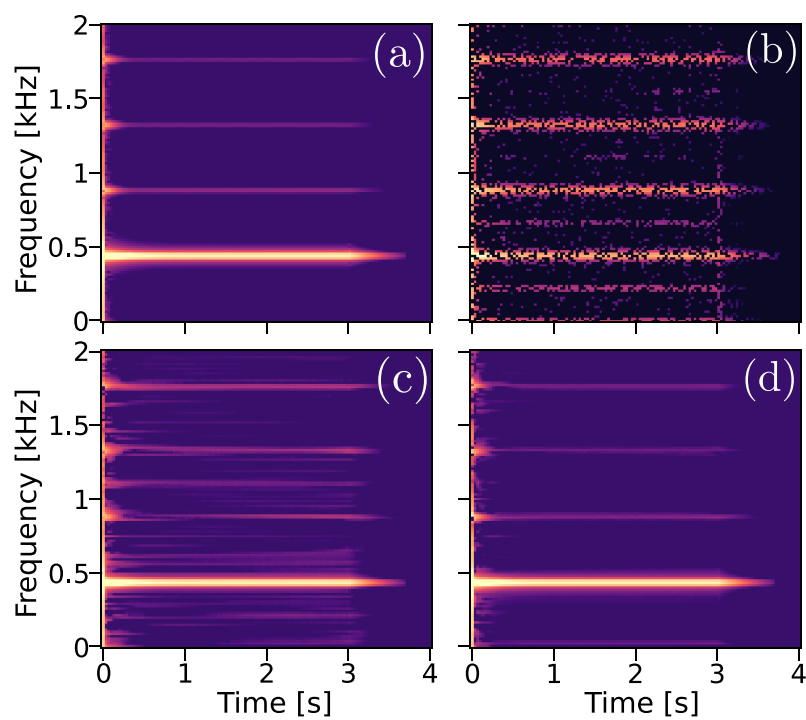


Fig.A.13. Example of spectrograms for test data (synthetic bass13 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

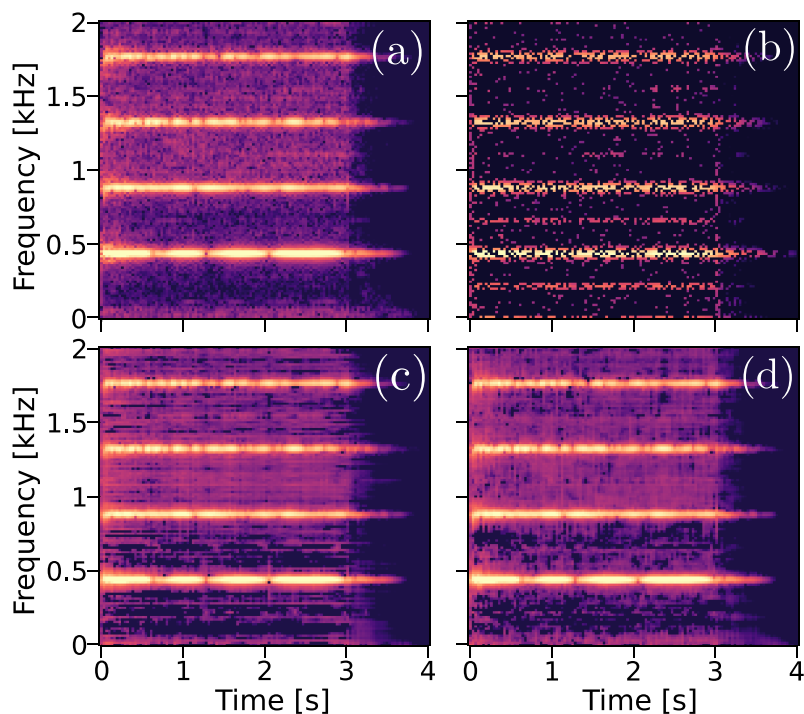


Fig.A.14. Example of spectrograms for test data (acoustic brass1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

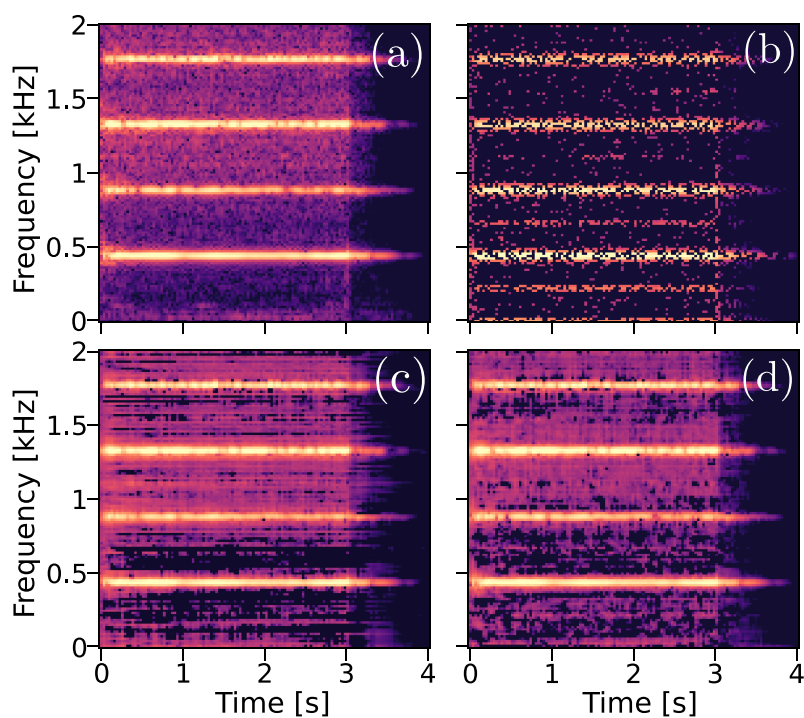


Fig.A.15. Example of spectrograms for test data (acoustic brass2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

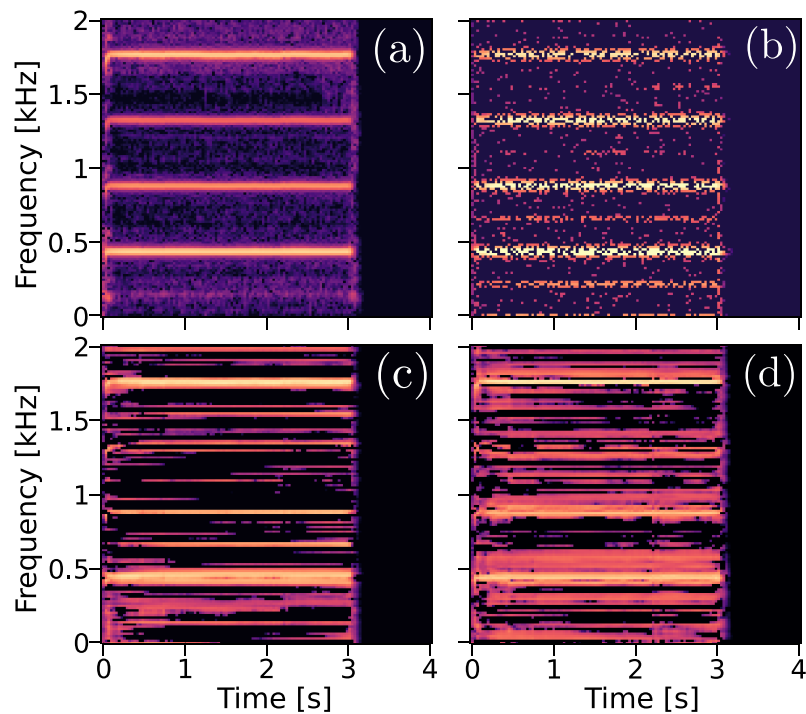


Fig.A.16. Example of spectrograms for test data (acoustic brass3 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

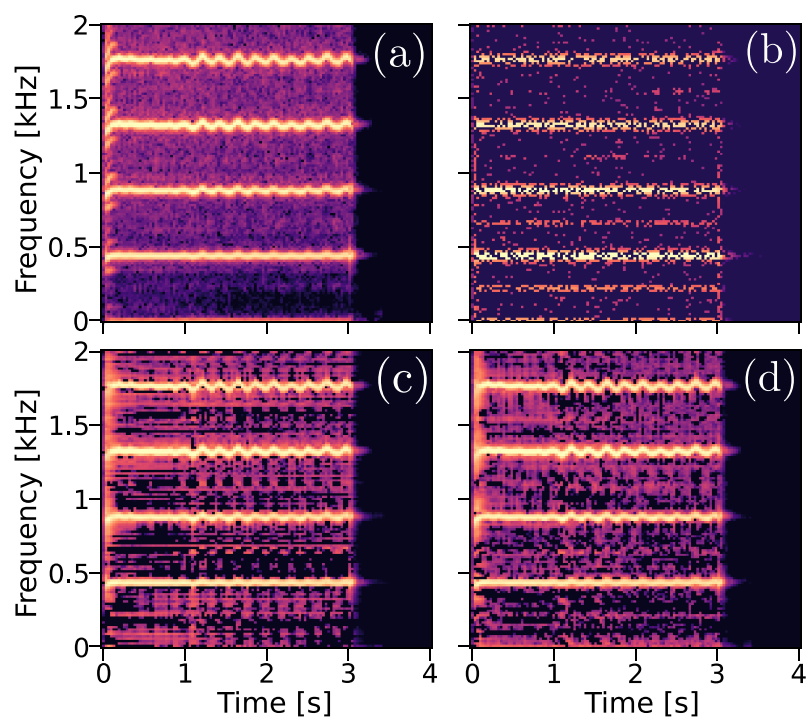


Fig.A.17. Example of spectrograms for test data (acoustic brass4 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

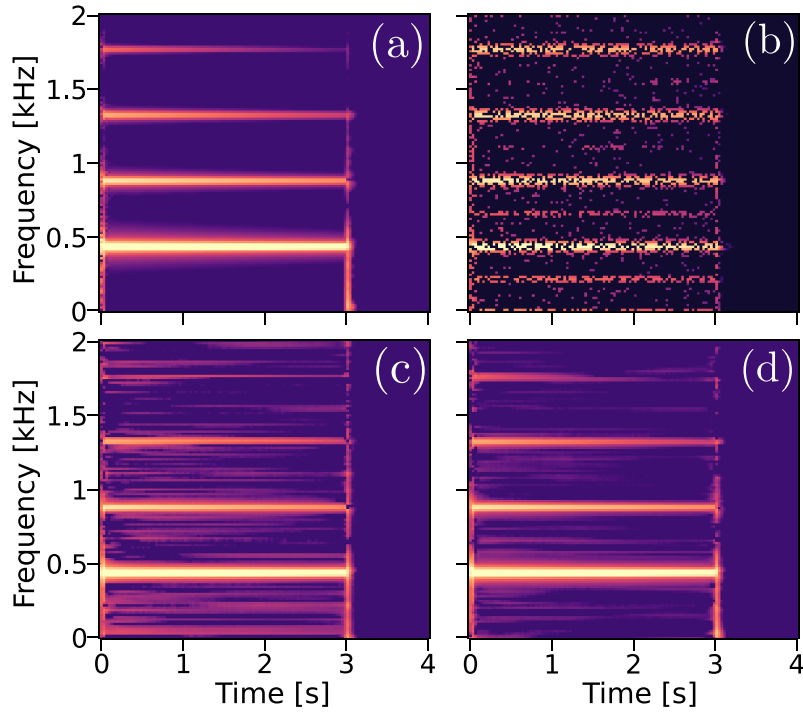


Fig.A.18. Example of spectrograms for test data (acoustic guitar1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

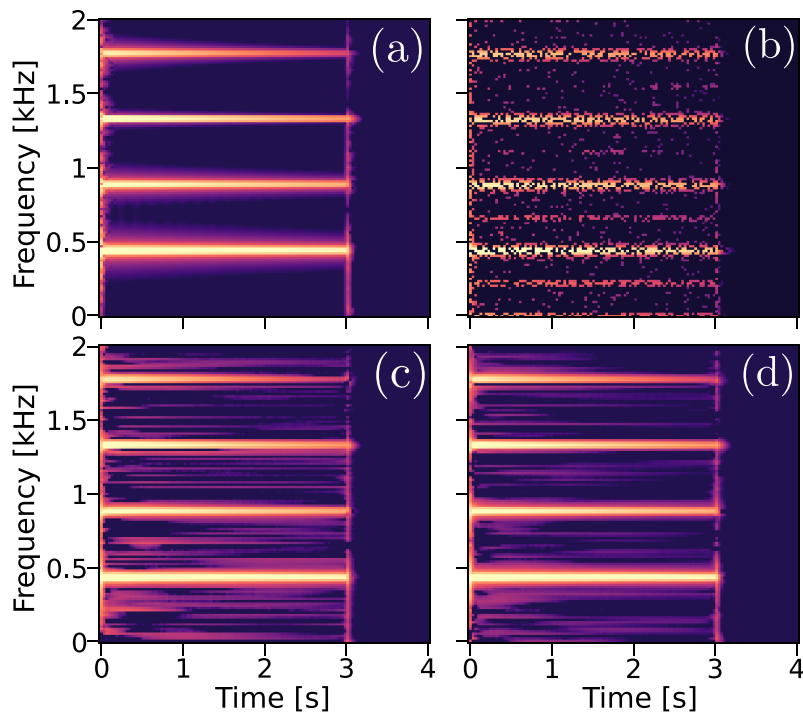


Fig.A.19. Example of spectrograms for test data (acoustic guitar2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

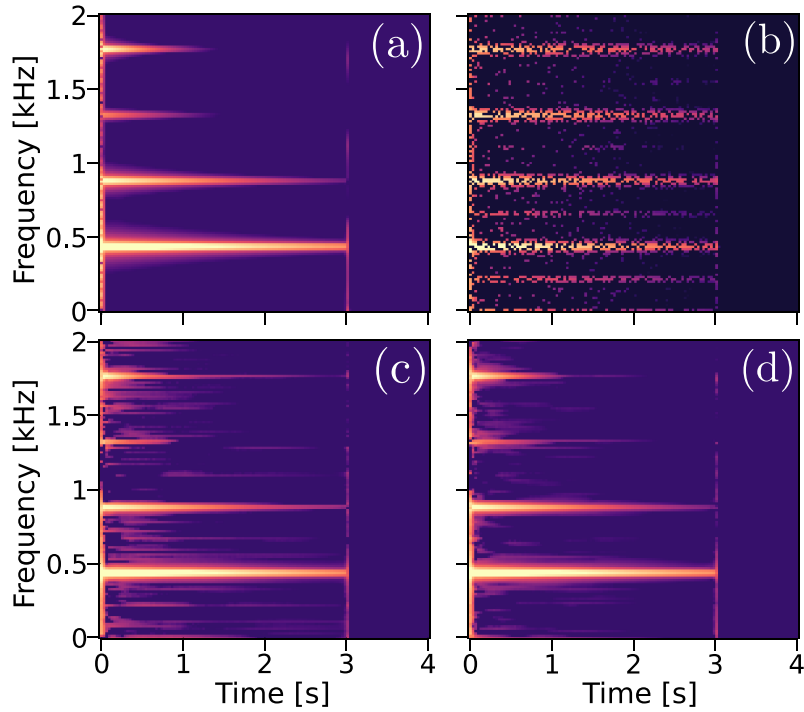


Fig.A.20. Example of spectrograms for test data (acoustic guitar3 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

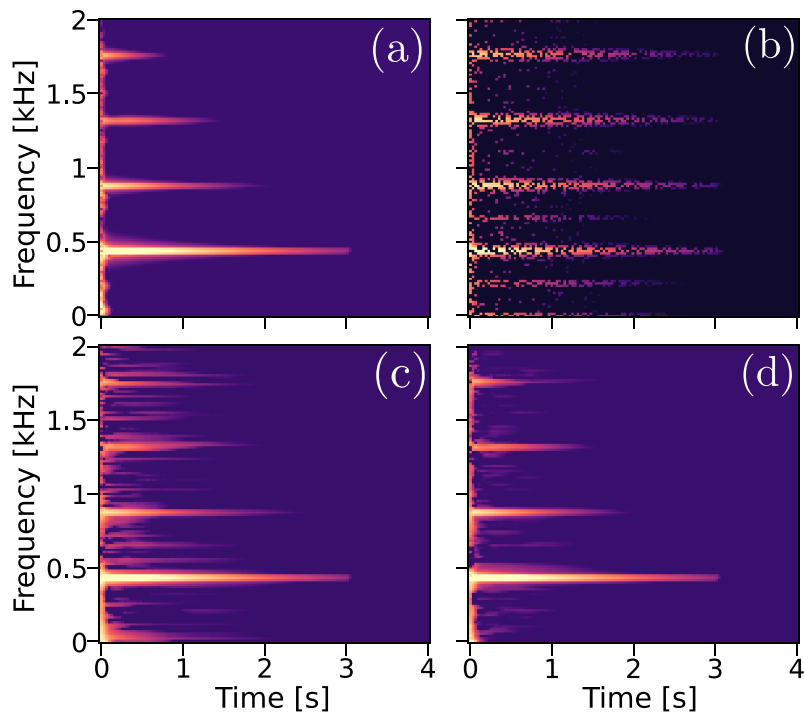


Fig.A.21. Example of spectrograms for test data (acoustic guitar4 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

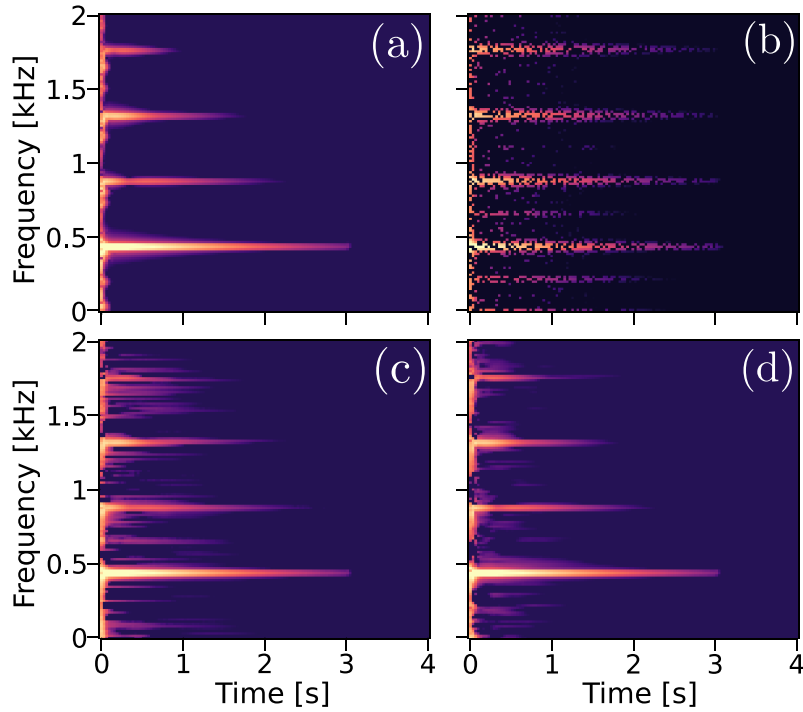


Fig.A.22. Example of spectrograms for test data (acoustic guitar5 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

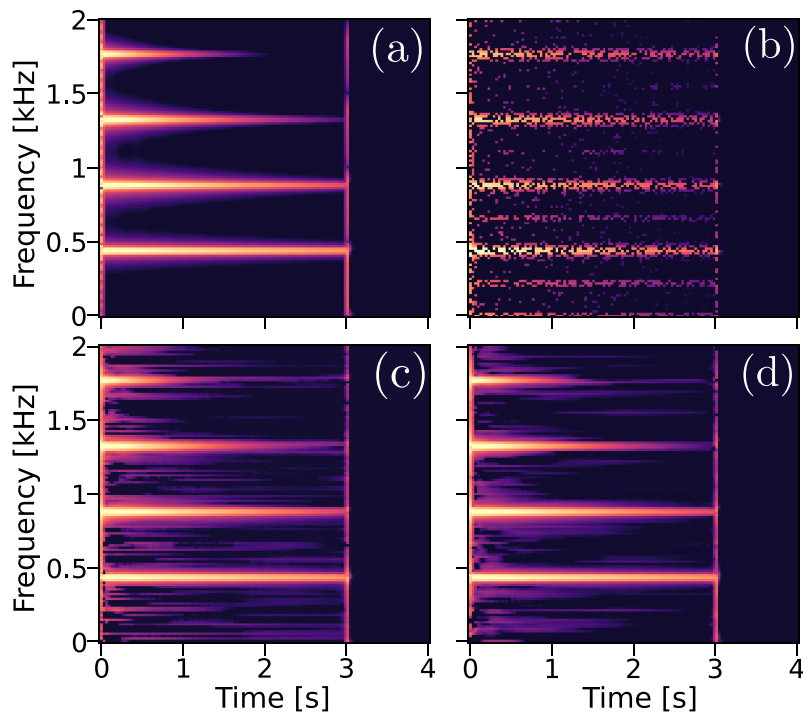


Fig.A.23. Example of spectrograms for test data (electronic guitar sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

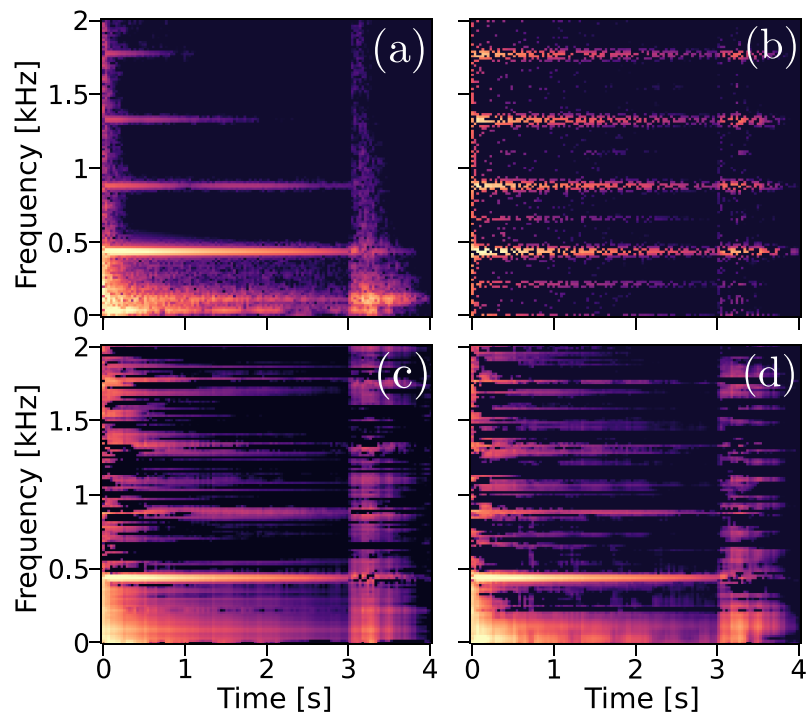


Fig.A.24. Example of spectrograms for test data (acoustic keyboard1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

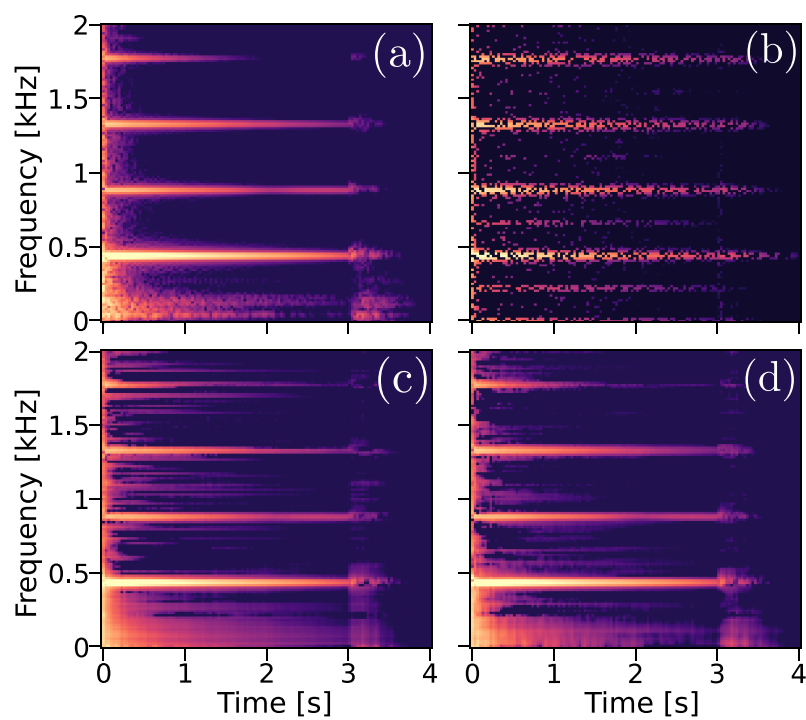


Fig.A.25. Example of spectrograms for test data (acoustic keyboard2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.



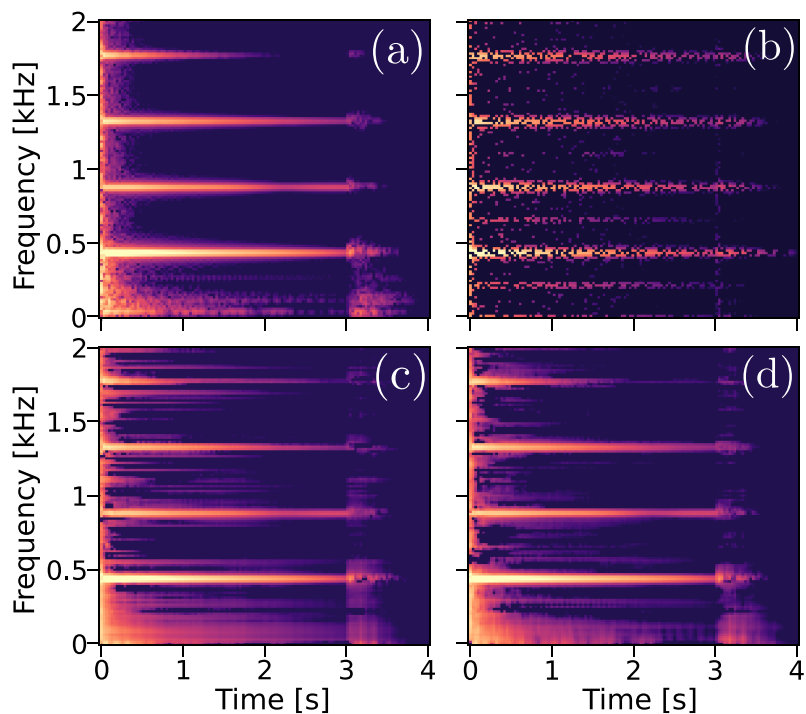


Fig.A.26. Example of spectrograms for test data (acoustic keyboard3 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

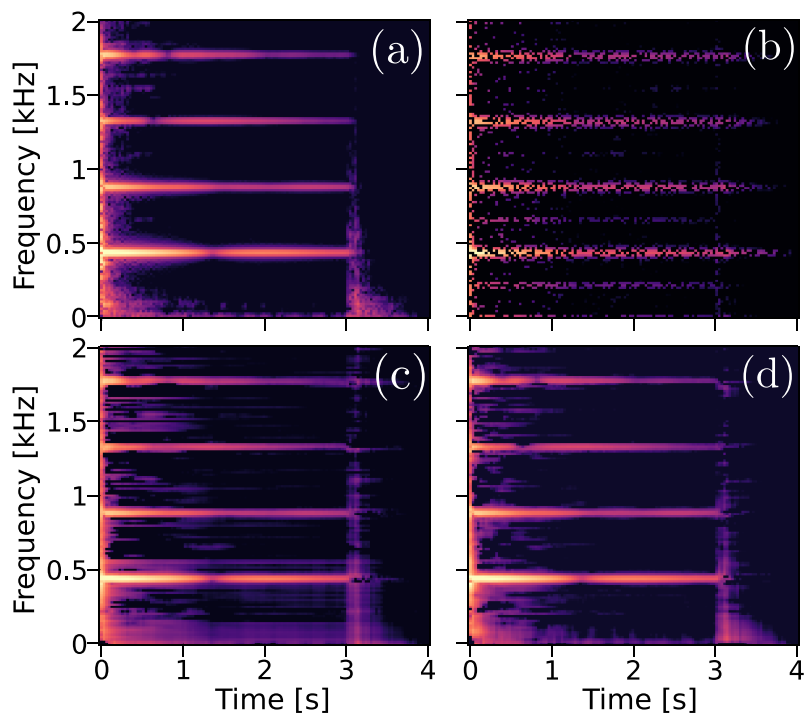


Fig.A.27. Example of spectrograms for test data (electronic keyboard1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.



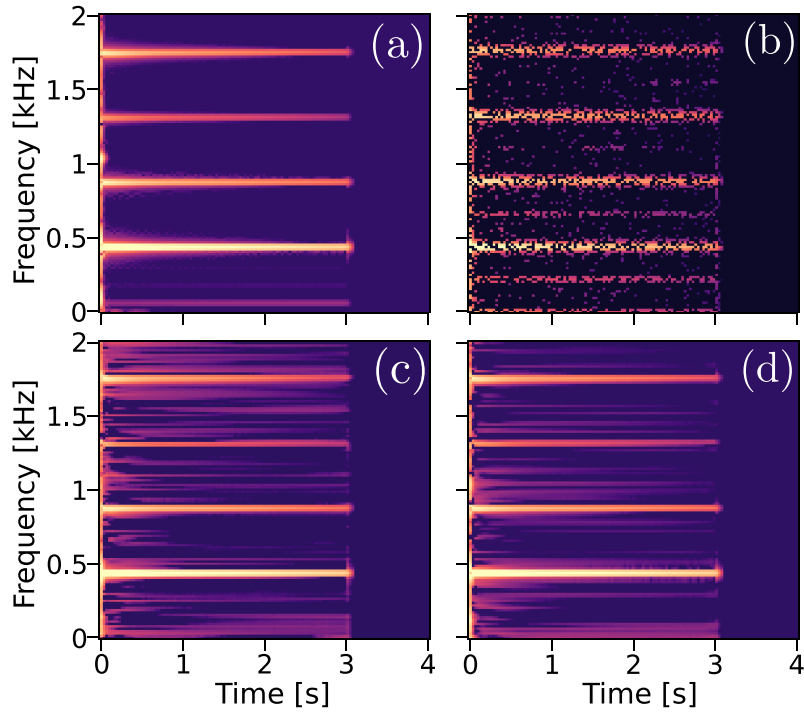


Fig.A.28. Example of spectrograms for test data (electronic keyboard2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

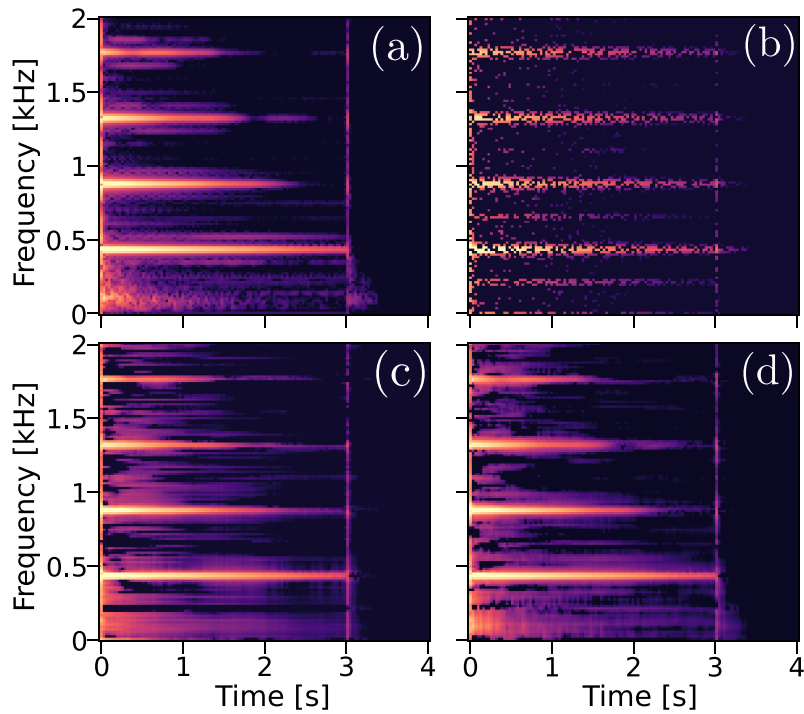


Fig.A.29. Example of spectrograms for test data (electronic keyboard3 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

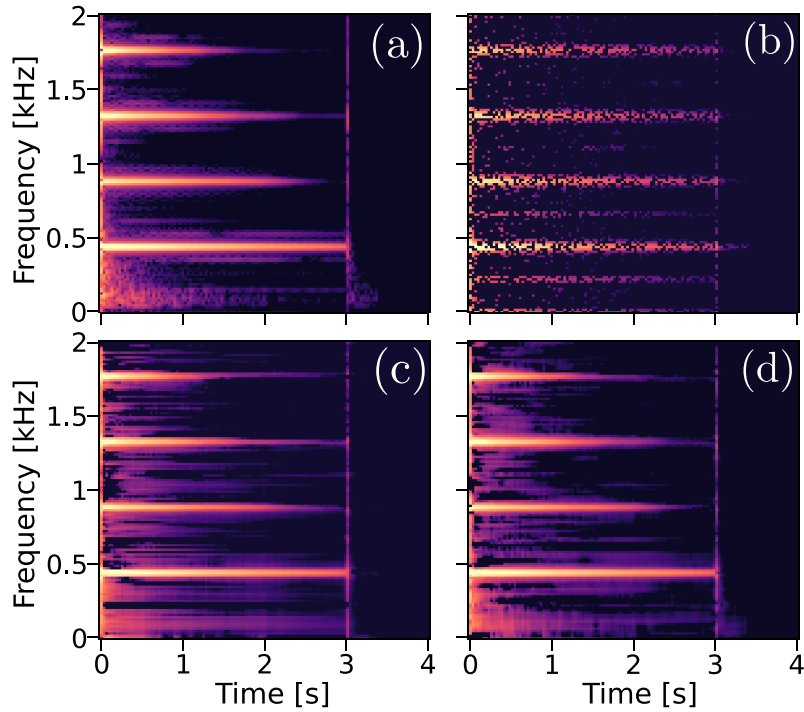


Fig.A.30. Example of spectrograms for test data (electronic keyboard4 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

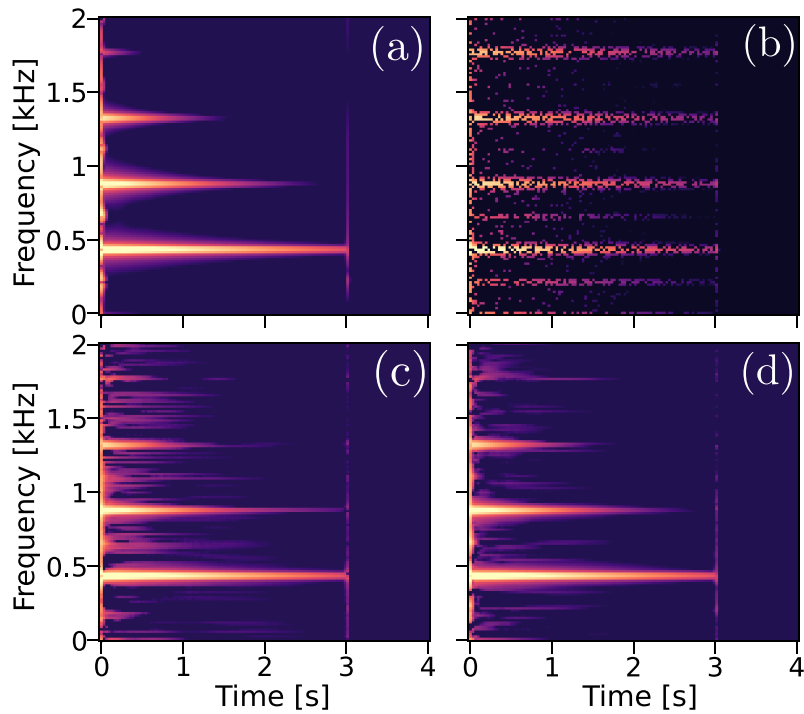


Fig.A.31. Example of spectrograms for test data (electronic keyboard5 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

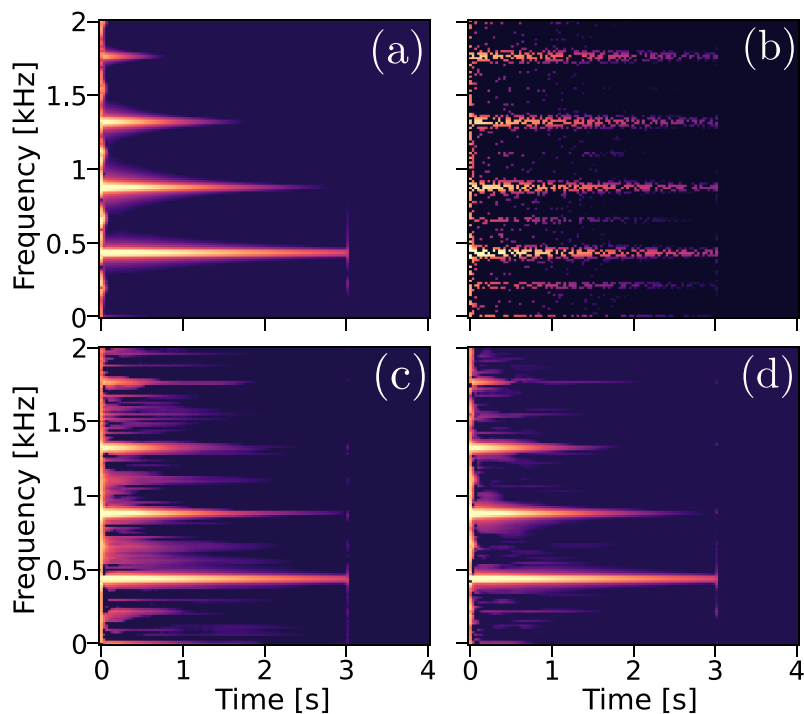


Fig.A.32. Example of spectrograms for test data (electronic keyboard6 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

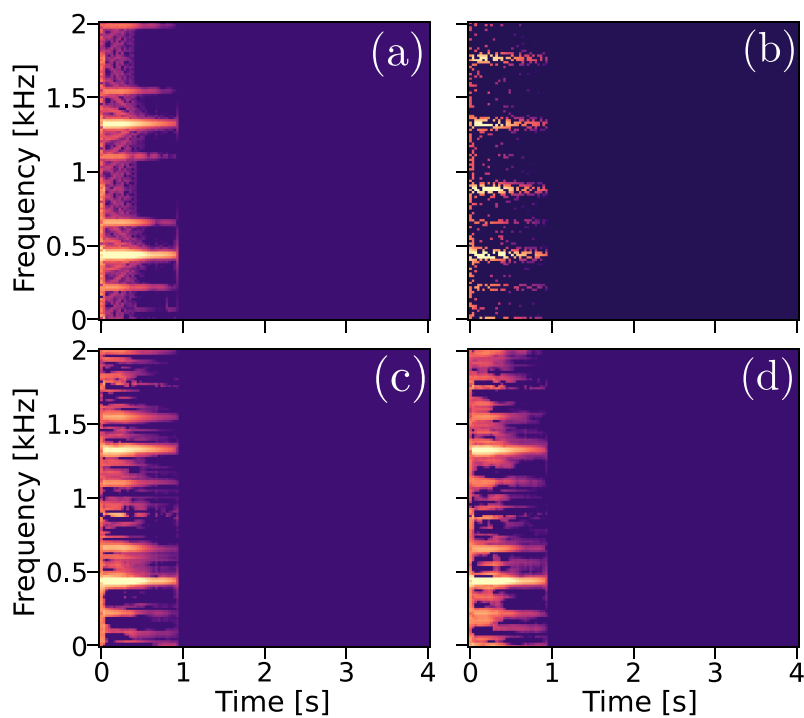


Fig.A.33. Example of spectrograms for test data (synthetic keyboard sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

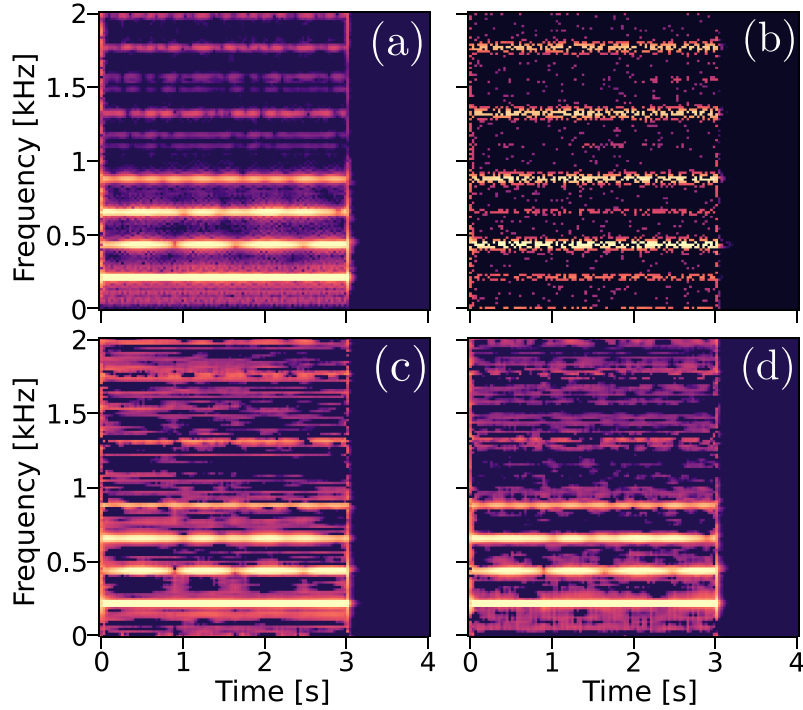


Fig.A.34. Example of spectrograms for test data (electronic organ1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

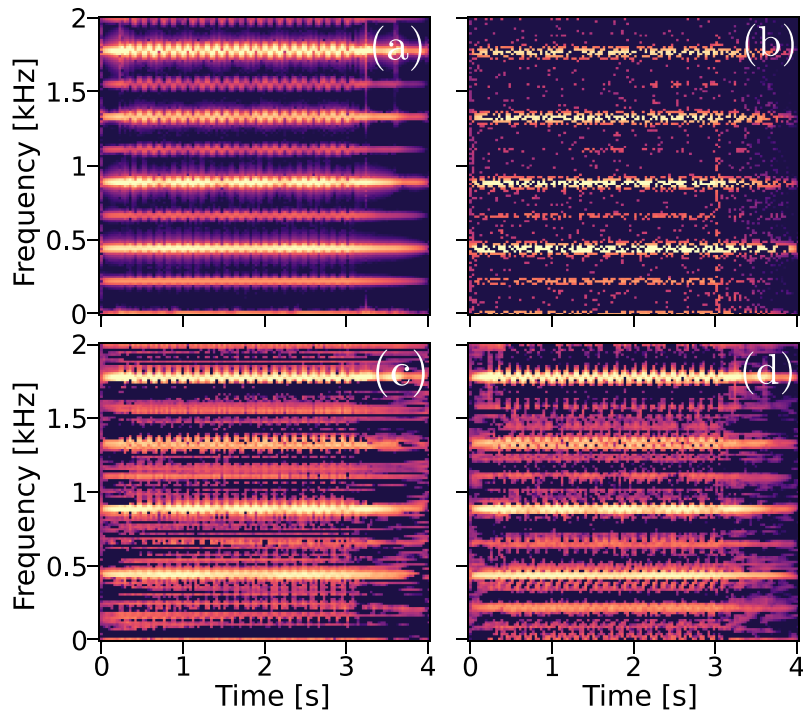


Fig.A.35. Example of spectrograms for test data (electronic organ2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

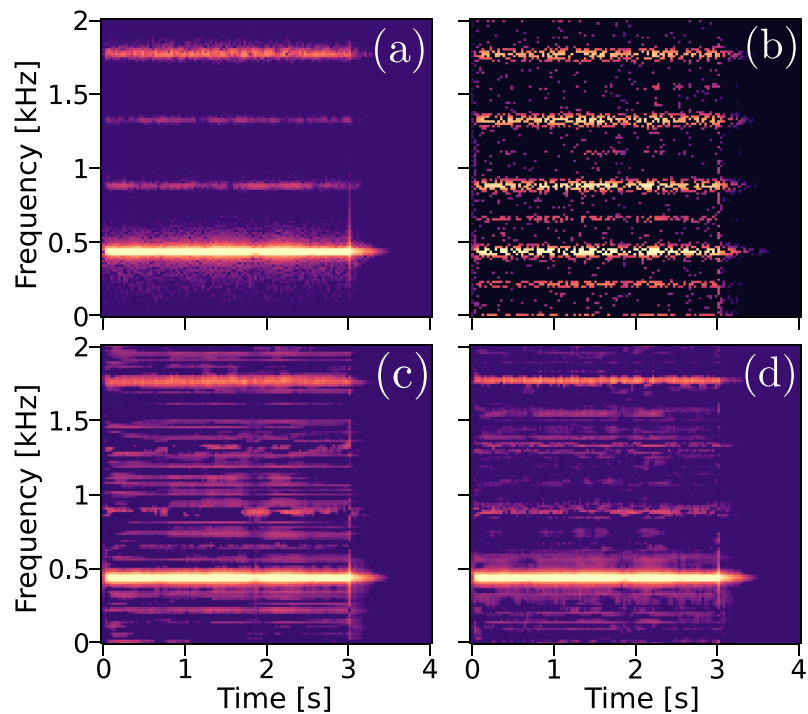


Fig.A.36. Example of spectrograms for test data (electronic organ3 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

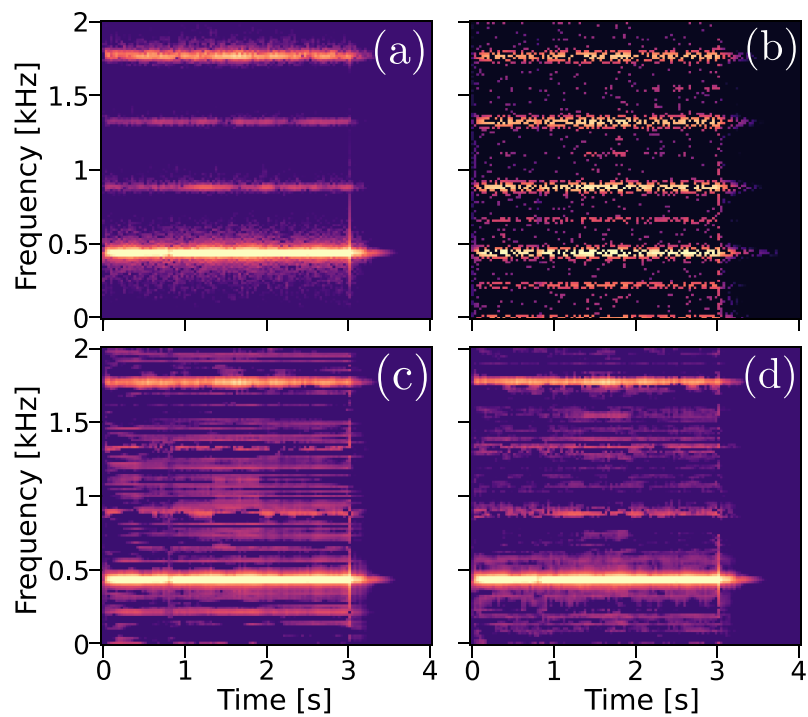


Fig.A.37. Example of spectrograms for test data (electronic organ4 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

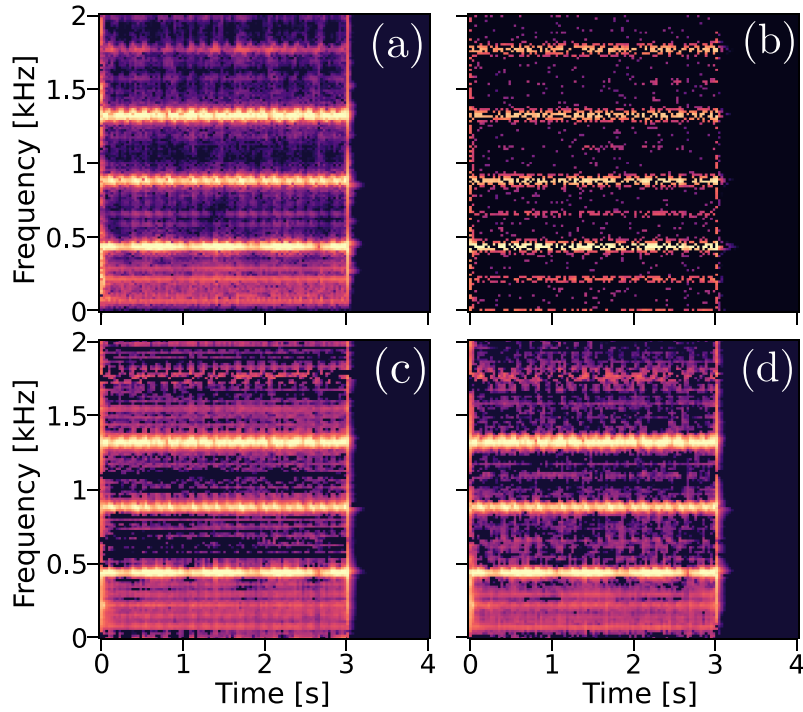


Fig.A.38. Example of spectrograms for test data (electronic organ5 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

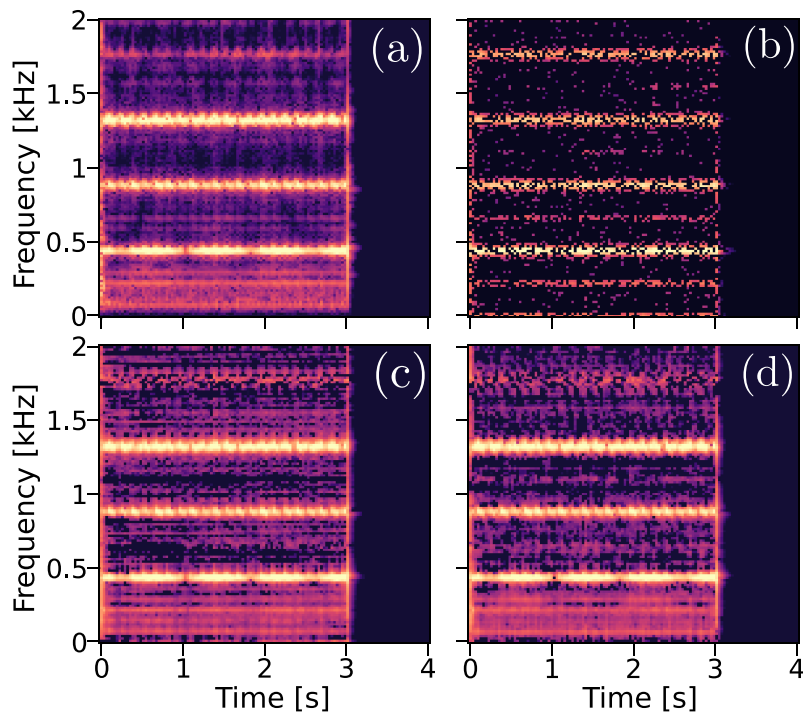


Fig.A.39. Example of spectrograms for test data (electronic organ6 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.



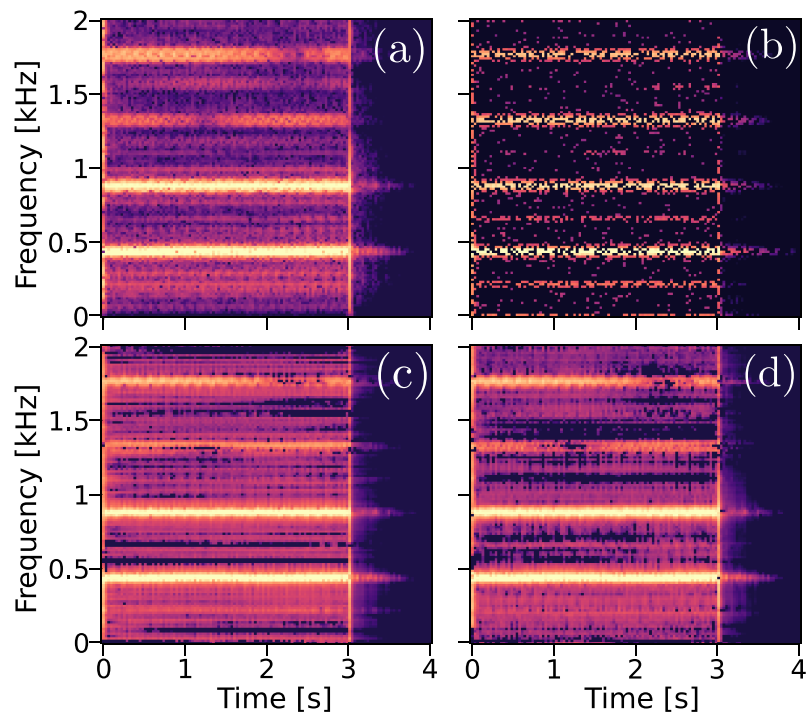


Fig.A.40. Example of spectrograms for test data (electronic organ7 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

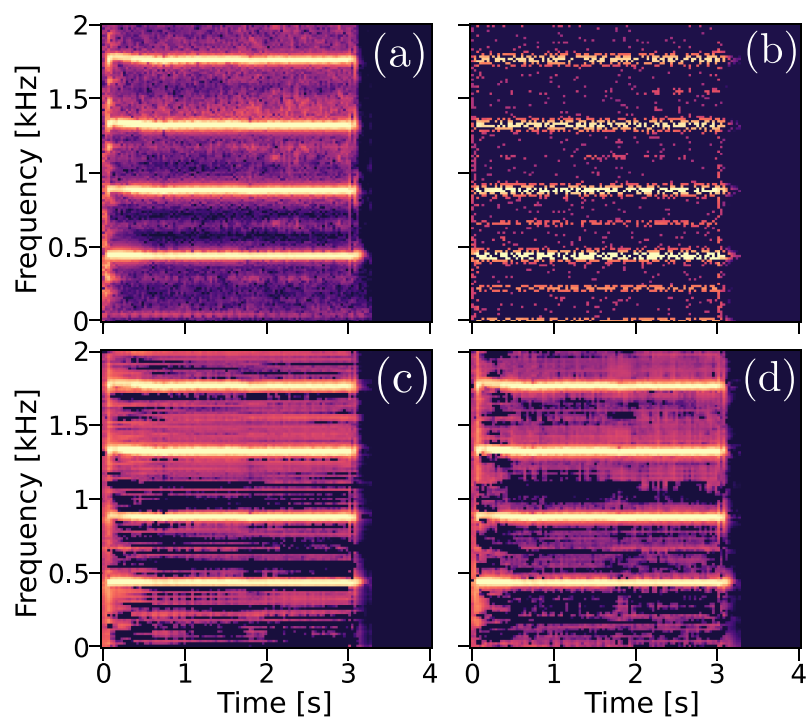


Fig.A.41. Example of spectrograms for test data (acoustic reed1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

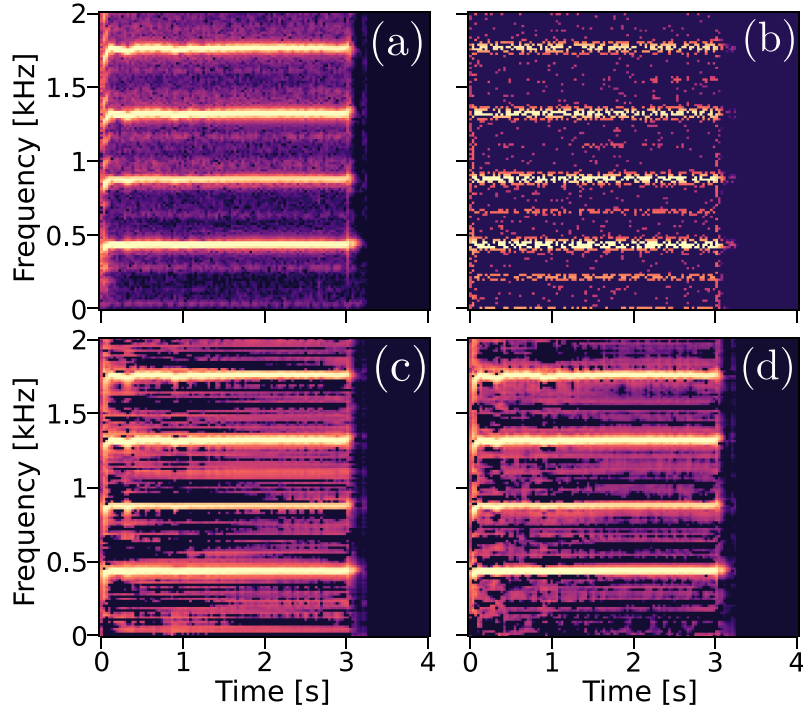


Fig.A.42. Example of spectrograms for test data (acoustic reed2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

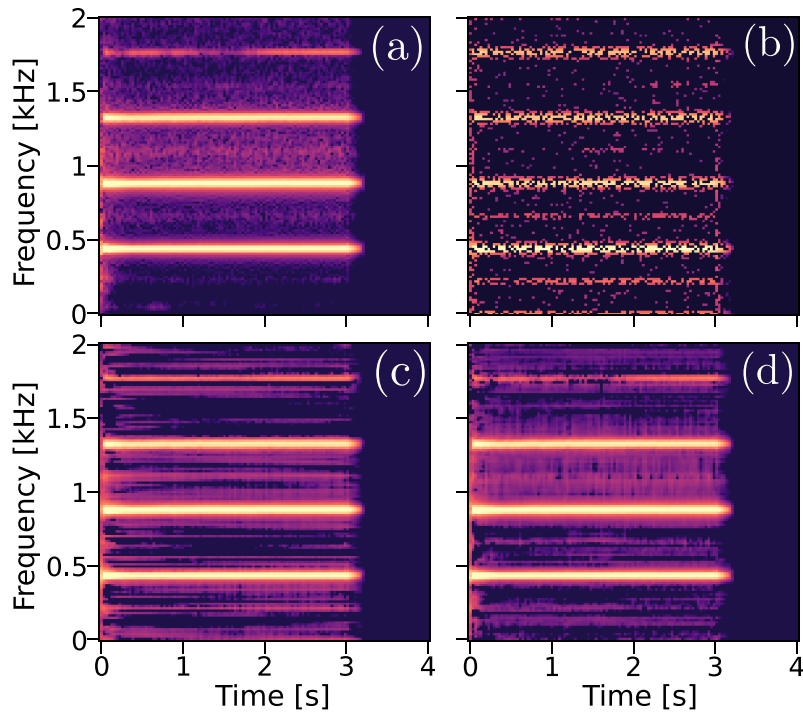


Fig.A.43. Example of spectrograms for test data (acoustic reed3 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.



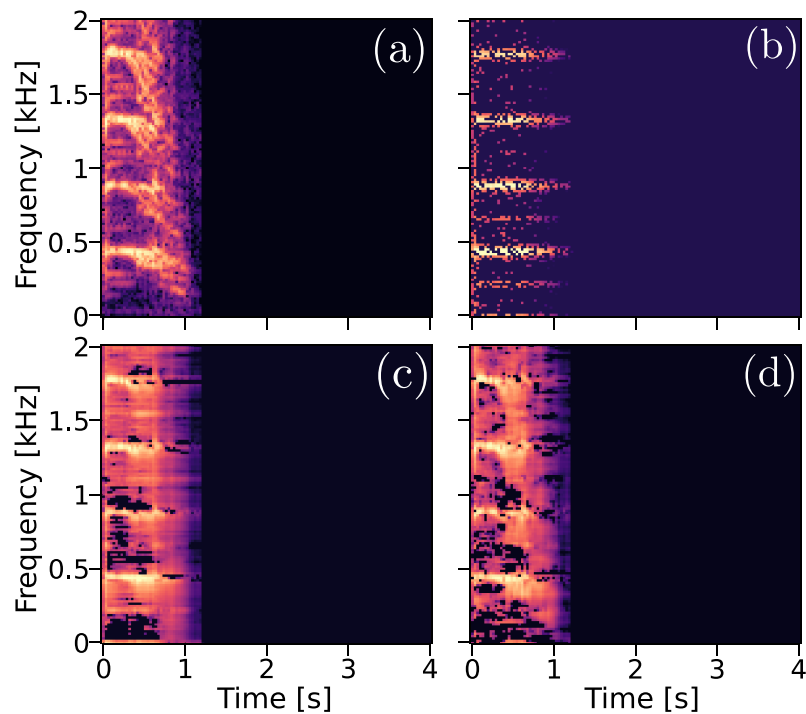


Fig.A.44. Example of spectrograms for test data (acoustic reed4 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

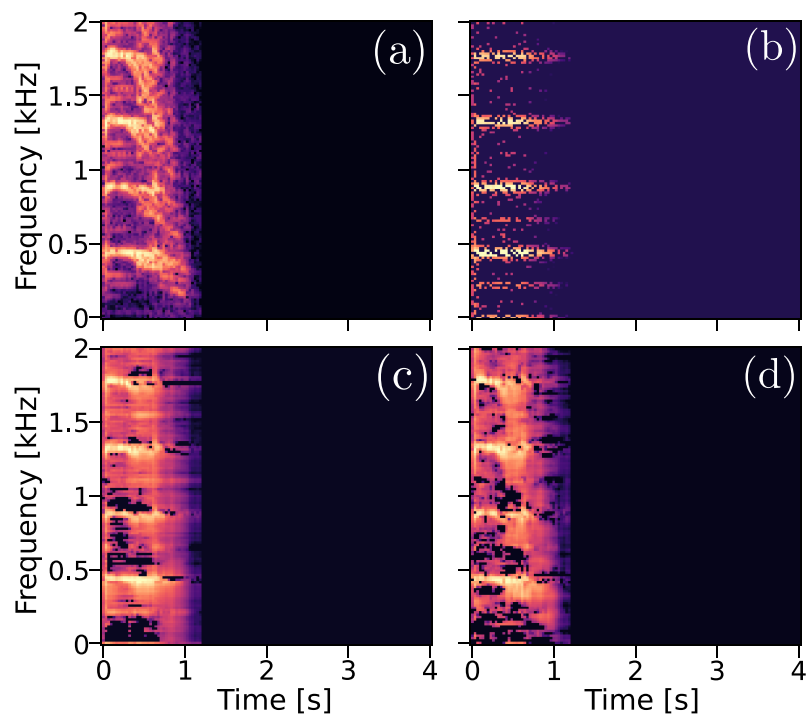


Fig.A.45. Example of spectrograms for test data (acoustic reed5 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

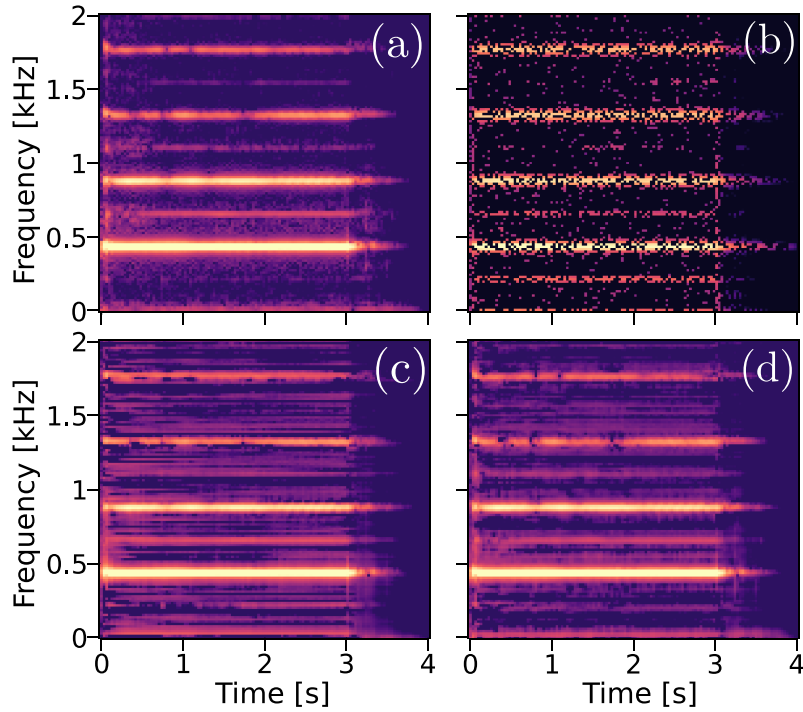


Fig.A.46. Example of spectrograms for test data (acoustic reed6 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

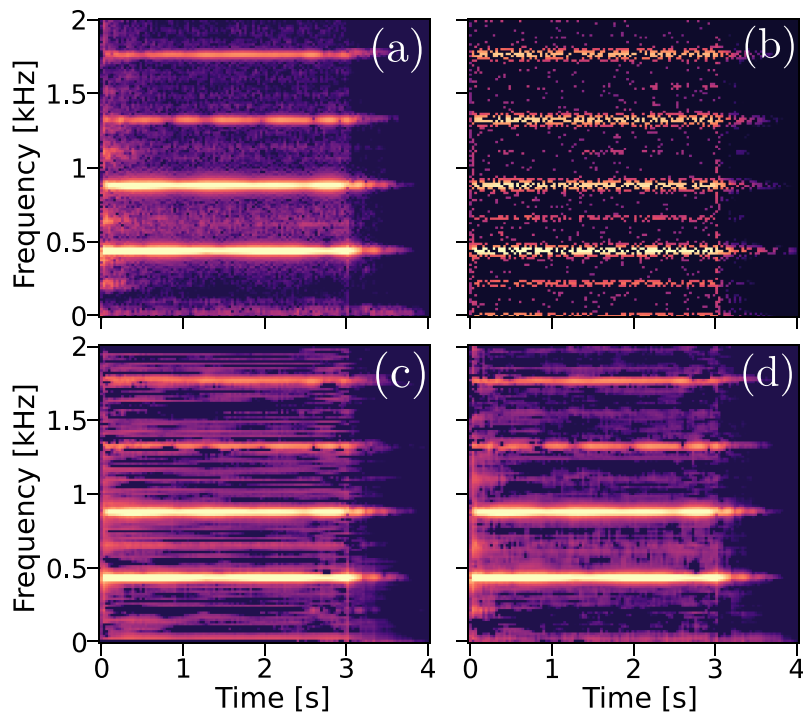


Fig.A.47. Example of spectrograms for test data (acoustic reed7 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

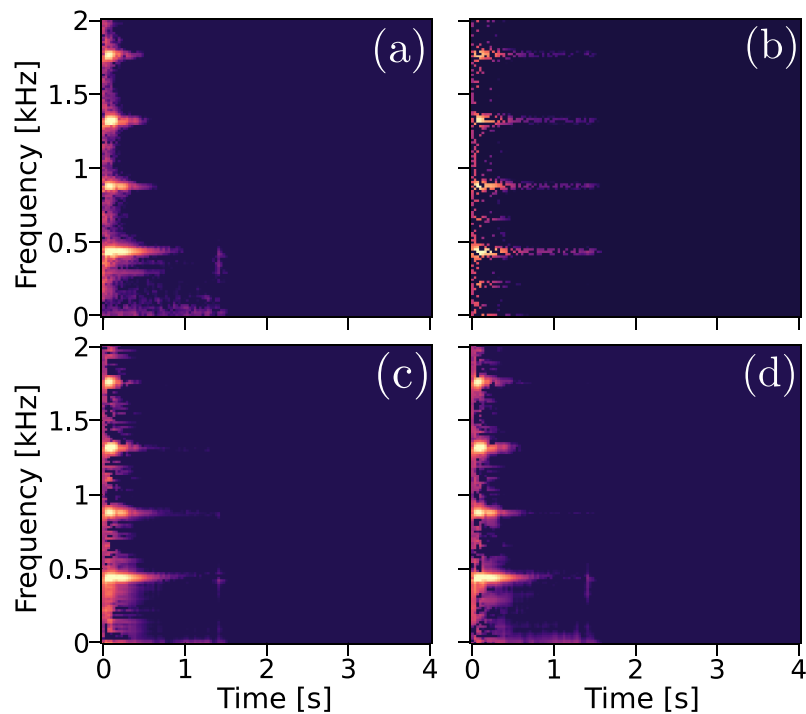


Fig.A.48. Example of spectrograms for test data (acoustic string1 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

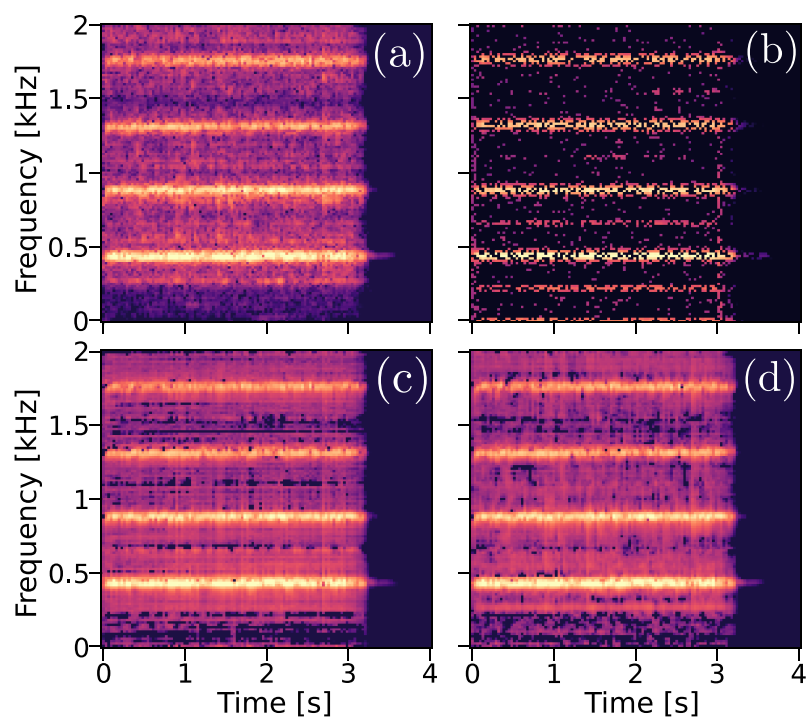


Fig.A.49. Example of spectrograms for test data (acoustic string2 sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

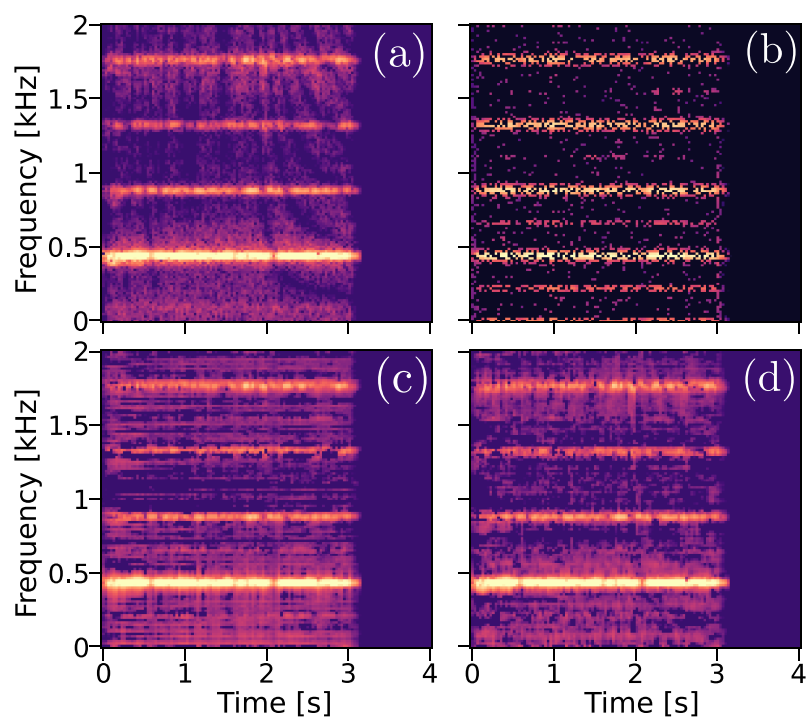


Fig.A.50. Example of spectrograms for test data (acoustic vocal sound): (a) original, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.