香川高専

卒業研究論文

論文題目

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

提出年月日	令和 4 年 2 月 25 日
学　　科	電気情報工学科
氏　　名	川口 翔也　　印
指導教員（主査）	北村 大地 講師　　印
副　　査	柿元 健 准教授　　印
学　科　長	辻 正敏 教授　　印

香川高等専門学校

Amplitude Spectrogram Prediction from Acoustic Features Based on Deep Learning

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 learning the latent features of MFCCs of multiple musical instruments with VAE, I can generate a new acoustic signal that has the features of multiple musical instruments. However, the proposed system assumes that the amplitude spectrogram can be predicted from the three features, fundamental frequency, 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 fundamental frequency, MFCC, and time-varying gains inputs. This is a subsystem necessary to realize the above-mentioned proposed system. In particular, three types of DNN are investigated to find the optimal network architecture. The experimental results using piano and guitar sounds show that the bidirectional recurrent DNN achieves the best performance for both instruments.

Keywords: variational auto-encoder, deep learning, mel-frequency cepstrum coefficient, timbre conversion

（和訳）

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，各楽器音の特徴量を併せ持つ新しい音響信号を新たに生成できると考えれる．しかしながら，この提案システムは基本周波数，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，基本周波数，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を実験的に調査する．ピアノ及びギターを用いた実験では，両楽器において双方向再帰型の DNN が高精度であることを示した．

目次

第 1 章

序論

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．これらのスペクトルには音高，音色，及び音量の情報が全て含まれている．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．このような基本的な特徴量の中で，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．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の同定 [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]，音色に基づくクラスタリングによる音源分

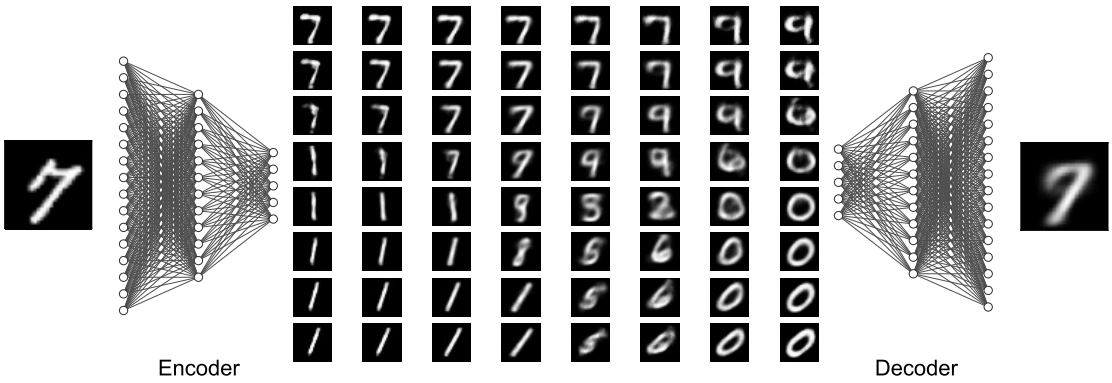


Fig. 1.1. The latent space of VAE trained with images of handwritten numbers.

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．特に VAE を用いた楽器音の解析や生成 [20, 21, 22, 23] 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）が生成可能できる．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することが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が期待される．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

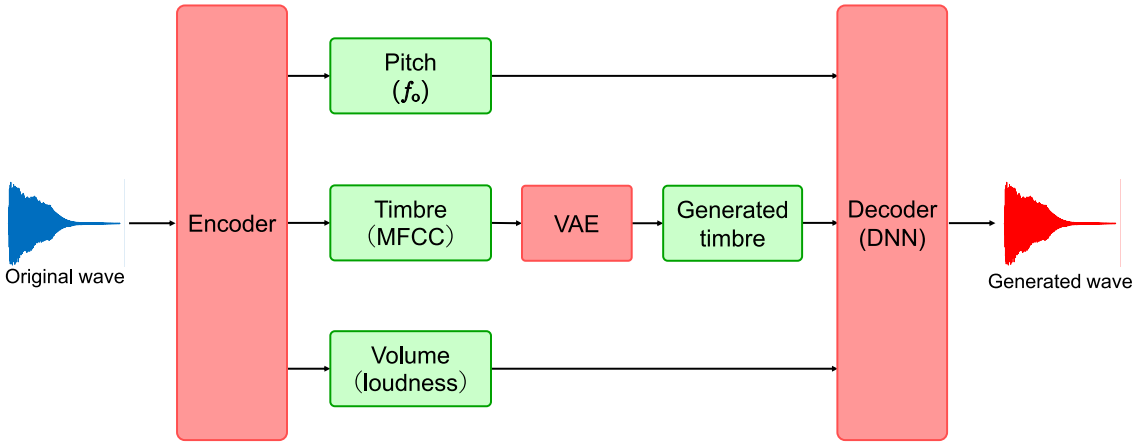


Fig. 1.2. Overview of the proposed sound generation system.

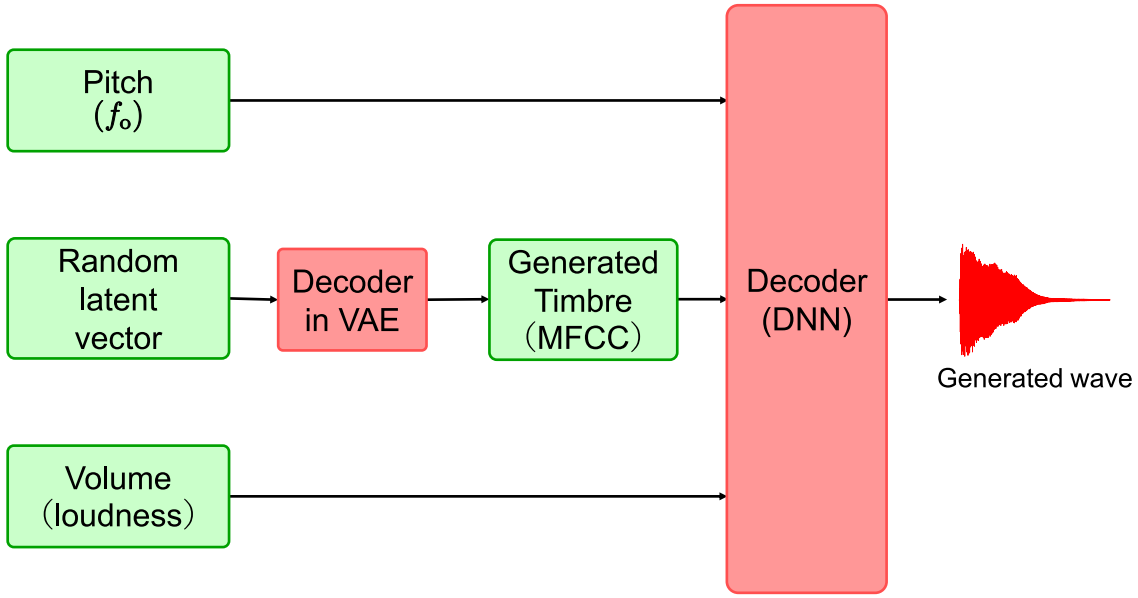


Fig. 1.3. New sound gerneration using the proposed system with a random vector input.

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 [20, 21, 22, 23] 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が提案を目指す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_{o}，音色は MFCC，及び音量はラウドネス（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 [24] 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）[25] 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実験的に調査する．

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．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，客観的な予測精度の評価と考察を行う．最後に 5 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 τ 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 $\boldsymbol{z}^{(j)}$ は次式で表される．

$$\begin{aligned}\boldsymbol{z}^{(j)} &= \left[z\left((j-1)\tau+1\right), z\left((j-1)\tau+2\right), \cdots, z\left((j-1)\tau+Q\right)\right]^{\mathrm{T}} \\ &= \left[z^{(j)}(1), z^{(j)}(2), \cdots, z^{(j)}(q), \cdots, z^{(j)}(Q)\right]^{\mathrm{T}} \in \mathbb{R}^{Q}\end{aligned} \tag{2.1}$$

ここで，$.^{\mathrm{T}}$ は転置を表し，$l = 1, 2, \cdots, L$，$j = 1, 2, \cdots, J$，及び $q = 1, 2, \cdots, Q$ 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．

$$J = \frac{L}{\tau} \tag{2.2}$$

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

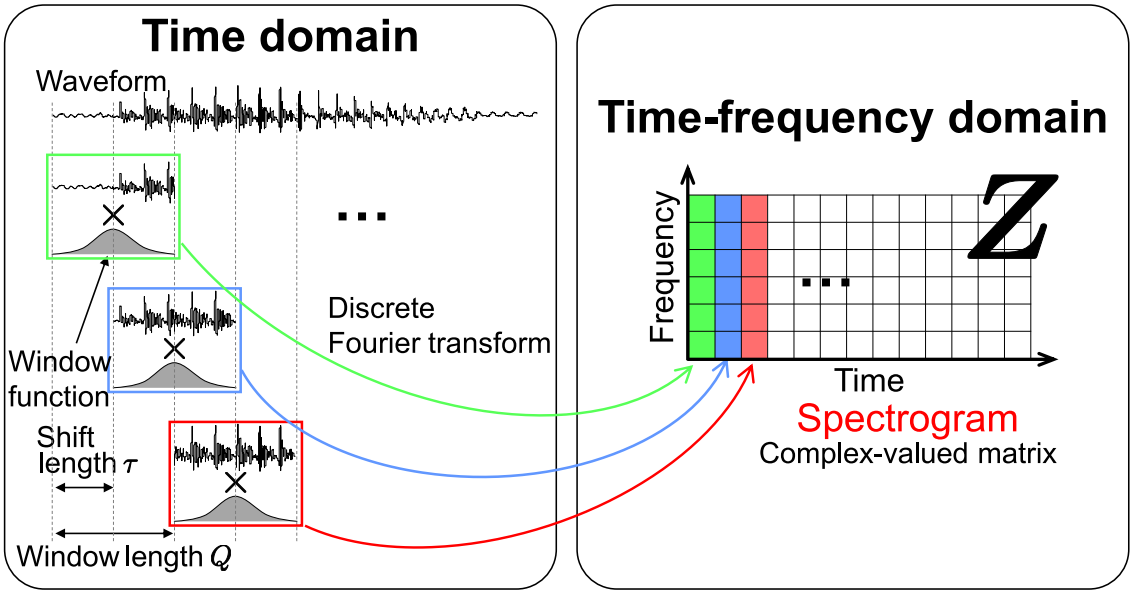


Fig. 2.1. Mechanism of STFT.

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

$$\boldsymbol{Z} = \mathrm{STFT}_{\boldsymbol{\omega}}(\boldsymbol{z})\ \in \mathbb{C}^{I\times J} \tag{2.3}$$

ここで，$\boldsymbol{\omega} = [\omega(1), \omega(2), \cdots, \omega(Q)]^{\mathrm{T}} \in \mathbb{R}^{Q}$ は STFT で用いる窓関数であり，$\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ の (i, j) 番目の要素は次式で表される．

$$z_{ij} = \sum_{q=1}^{Q} \omega(q) z^{(j)}(q) \exp\left\{\frac{-\iota 2\pi (q-1)(i-1)}{F}\right\} \tag{2.4}$$

ここで，z_{ij} は $\boldsymbol{Z}$ 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），$i = 1, 2, \cdots, I$ は周波数ビンのインデクス，ι 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 $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ 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 $|\boldsymbol{Z}|\ \in \mathbb{R}_{\geq 0}^{I\times J}$ や，絶対値の 2 乗を取ったパワースペクトログラム $|\boldsymbol{Z}|^{\cdot 2}\ \in \mathbb{R}_{\geq 0}^{I\times J}$ 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．

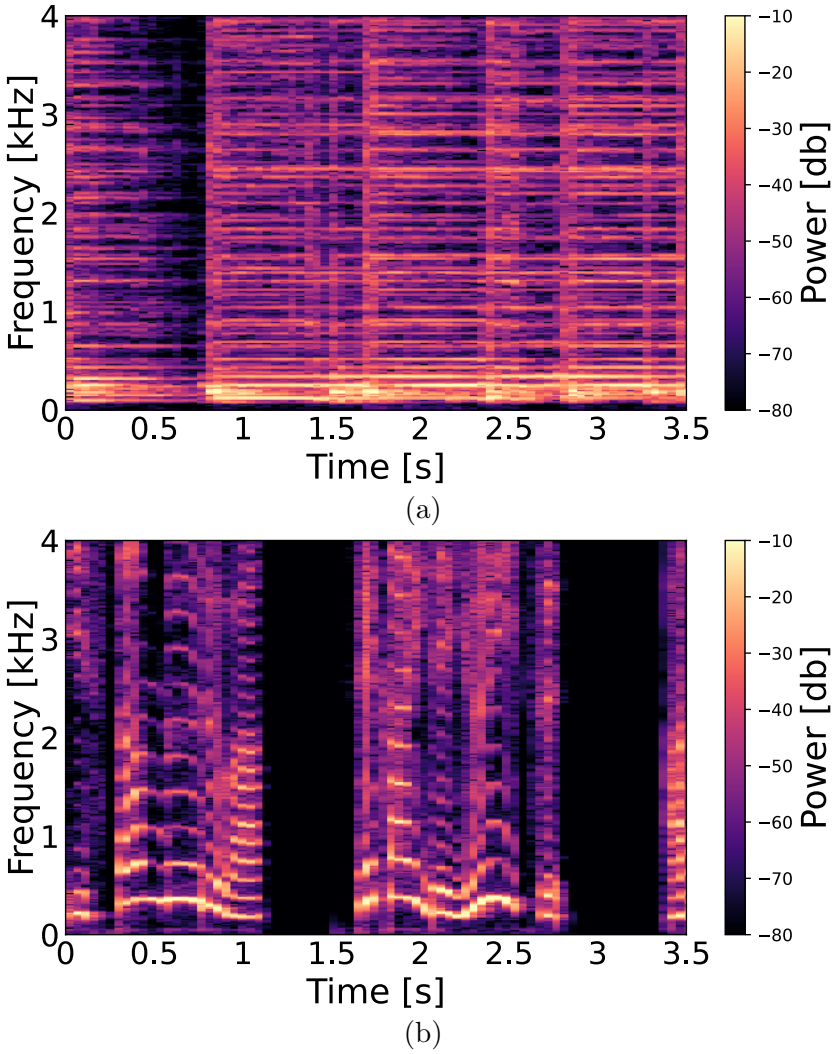


Fig. 2.2. Power spectrograms of (a) music and (b) speech signals.

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

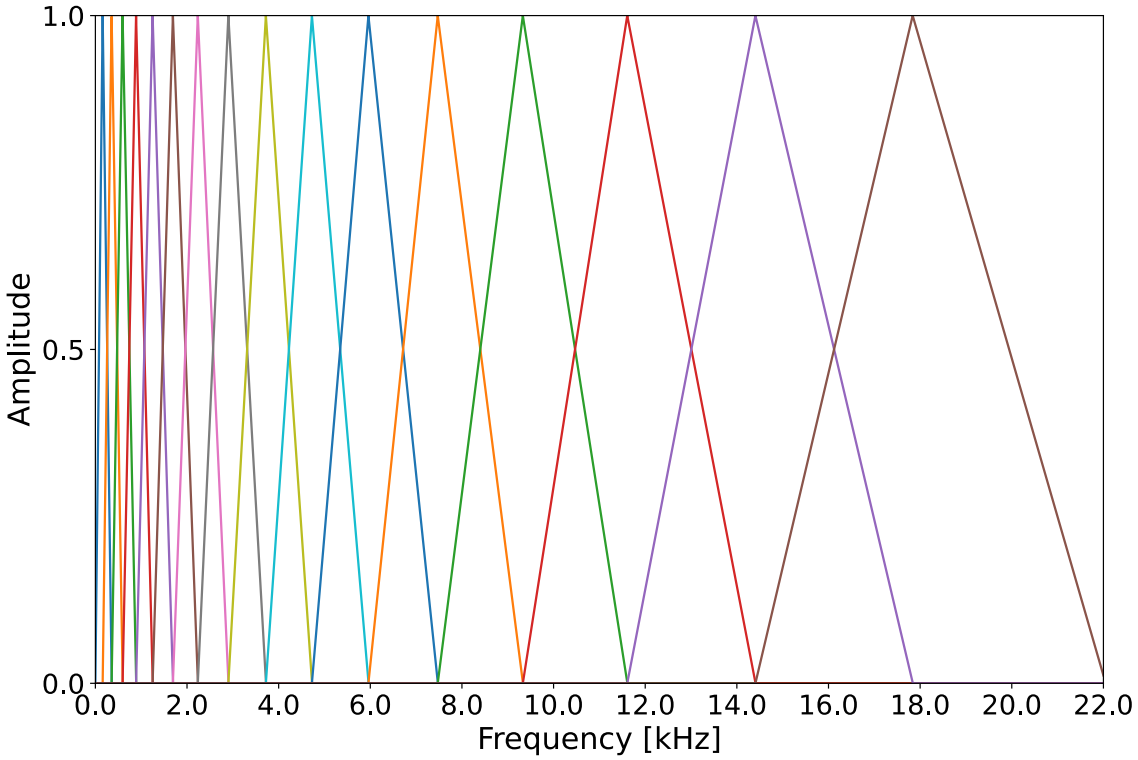


Fig. 2.3. Mel-filter bank ($f_{\min} = 0$ Hz, $f_{\max} = 22050$ Hz, $K = 16$).

関係は次式で定義される [26].

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

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

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 $|\boldsymbol{Z}|^{\cdot 2}$ 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 $|\boldsymbol{Z}|^{\cdot 2}$ 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 $\boldsymbol{P}$ と呼び，この変換を次式で表す．

$$\boldsymbol{P} = \mathrm{MelFiltering}(|\boldsymbol{Z}|^{\cdot 2}) \in \mathbb{R}^{K \times J} \tag{2.7}$$

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 $\boldsymbol{P}$ 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．

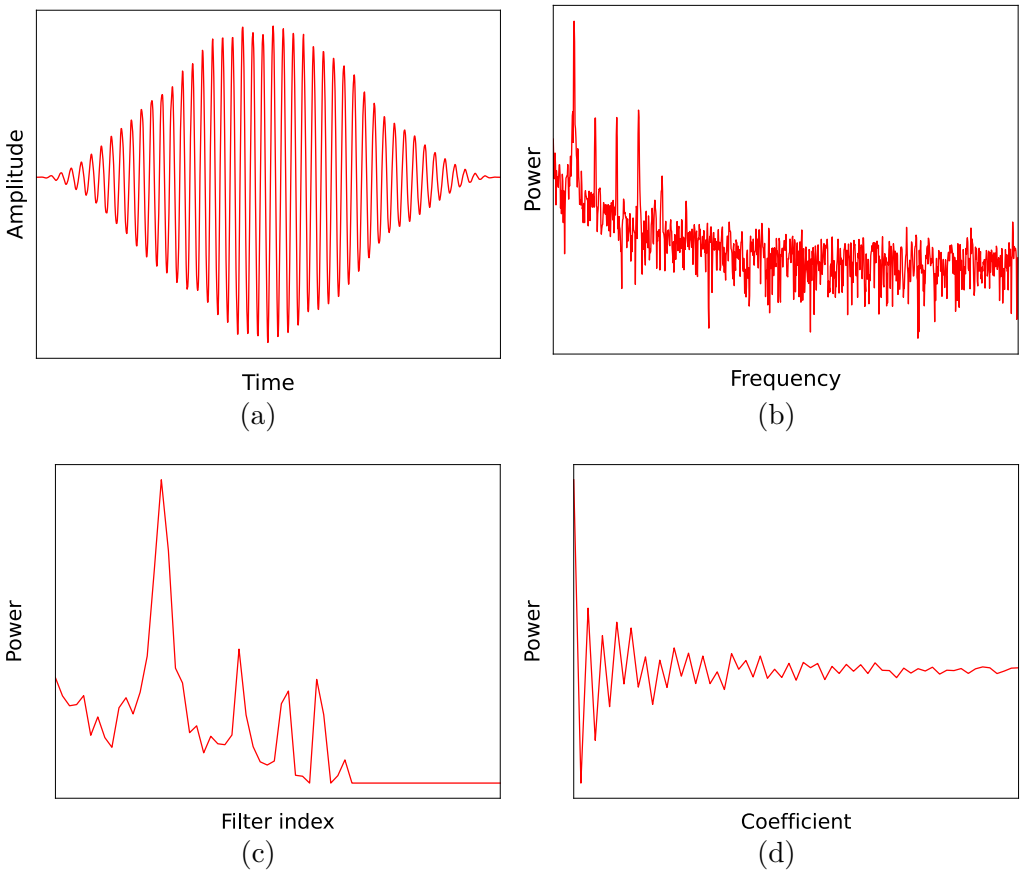


Fig. 2.4. (a) windowed short-time signal of piano B5 sound, (b) power spectrum, (c) mel-frequency spectrum, and (d) MFCC.

MFCC を $\boldsymbol{C}$ と表記し，この処理を次式で表す．

$$\boldsymbol{C} = \mathrm{DCT}(\boldsymbol{P}) \in \mathbb{R}^{K \times J} \tag{2.8}$$

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 $|\boldsymbol{Z}|^{\cdot 2}$ と MFCC $\boldsymbol{C}$ 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．

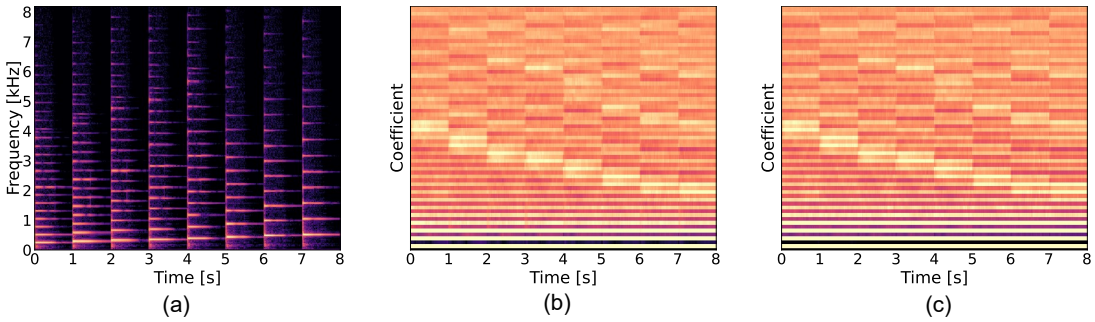


Fig. 2.5. (a) power spectrogram of ascending notes played by piano sound, (b) MFCC, and (c) normalized MFCC.

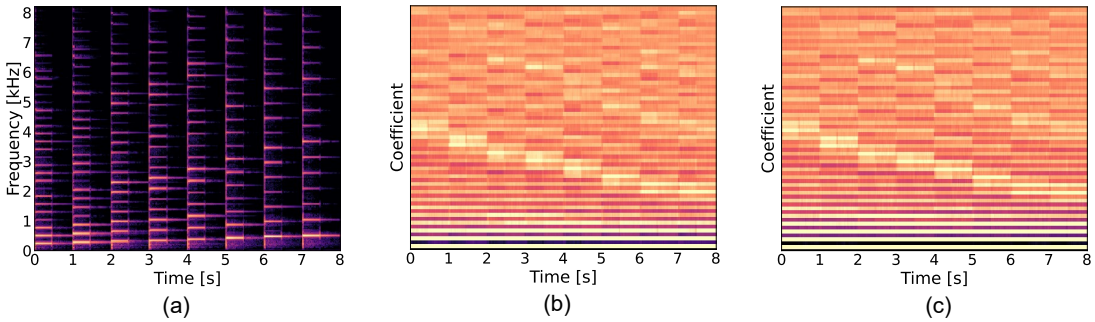


Fig. 2.6. (a) power spectrogram of ascending notes played by guitar sound, (b) MFCC, and (c) normalized MFCC.

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．

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

ここで，$n = 1, 2, \cdots, N$ は入力データの要素インデクス，$h = 1, 2, \cdots, H$ は出力データの要素インデクス，w_{nh} は重み係数を表す．また ϕ 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

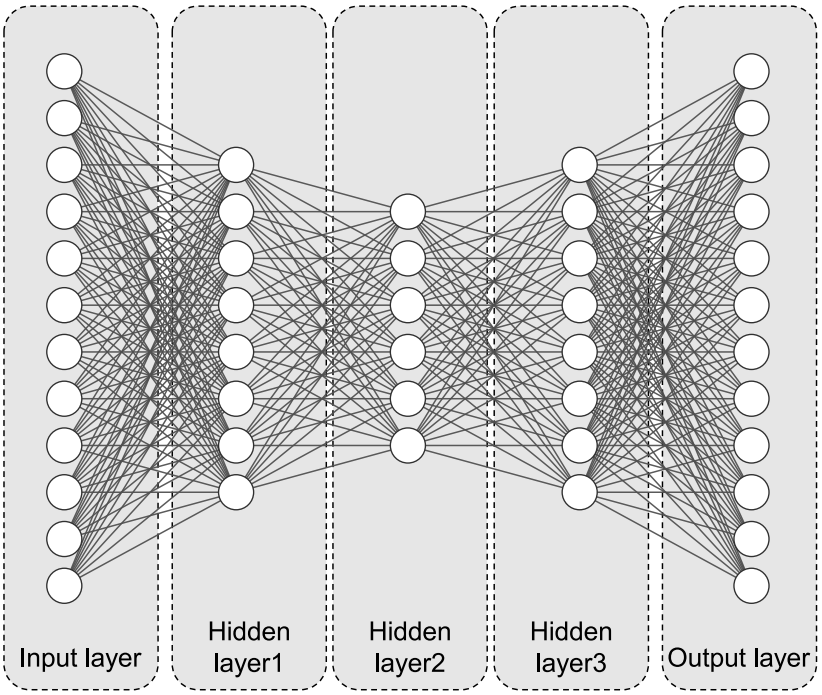


Fig. 2.7. Multilayer neural network.

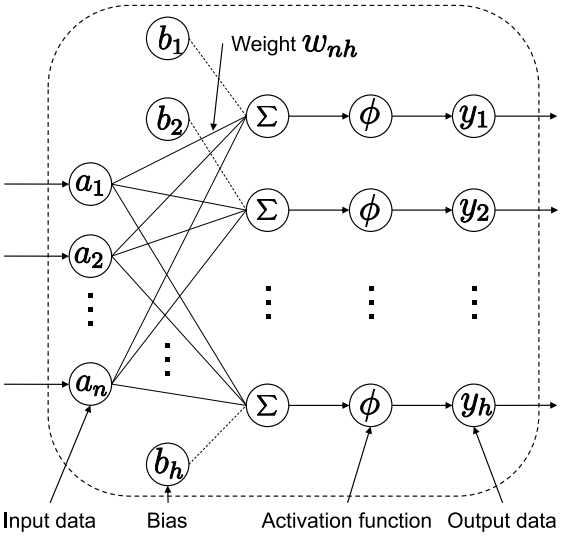


Fig. 2.8. Structure of fully connected layer.

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

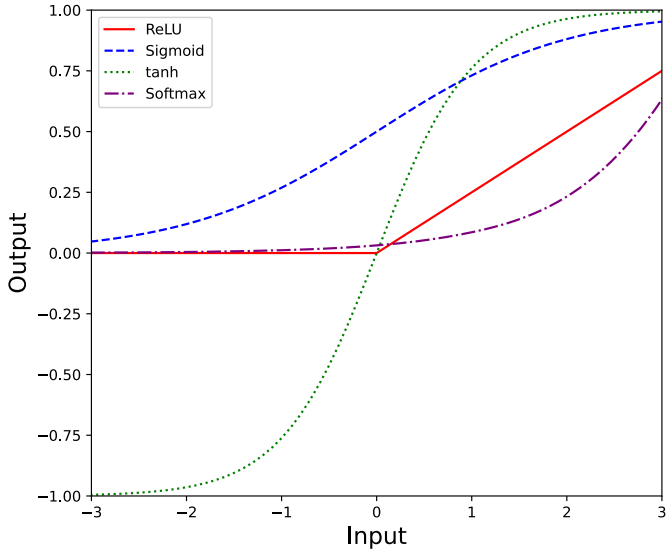


Fig. 2.9. Popular activation functions used in DNN.

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 [20]，及び VAE を用いた音色・音高の分離表現の学習 [21, 22, 23] 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 F0），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 F0 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．

$$L_{\mathrm{MSS}}{}^{(F)}(\boldsymbol{A}^{(F)}, \hat{\boldsymbol{A}}^{(F)}) = \|\boldsymbol{A}^{(F)} - \hat{\boldsymbol{A}}^{(F)}\|_1 + \|\log \boldsymbol{A}^{(F)} - \log \hat{\boldsymbol{A}}^{(F)}\|_1 \tag{2.10}$$

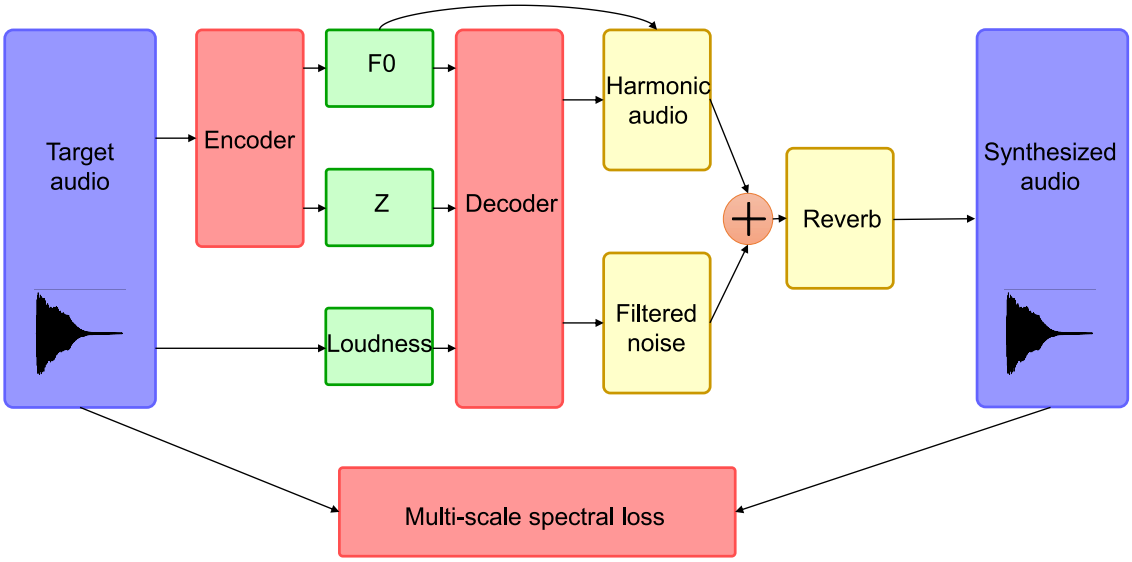


Fig. 2.10. DDSP autoencoder architecture. Red components are part of the neural network architecture, green components are the latent representation, and yellow components are deterministic synthesizers and effects. For the detailed explanation of this figure, see [13].

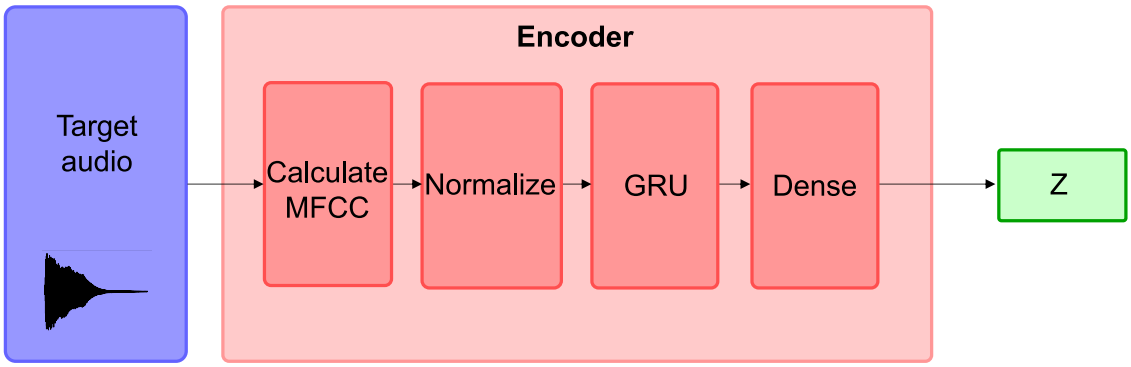


Fig. 2.11. Diagram of the Z-encoder. For the detailed explanation of this figure, see [13].

ここで，$\boldsymbol{A}^{(F)}$ 及び $\hat{\boldsymbol{A}}^{(F)}$ 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 [27] 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

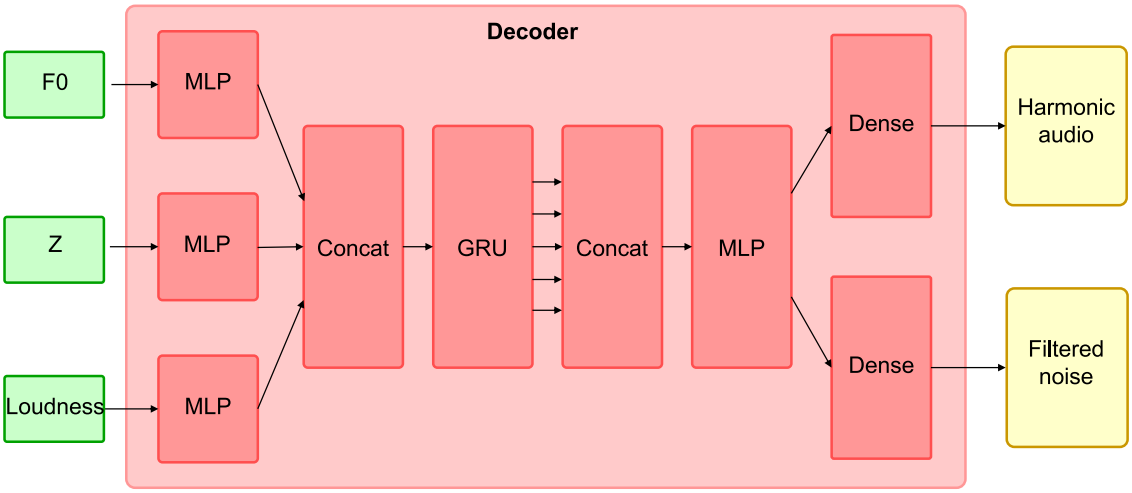


Fig. 2.12. Diagram of the decoder for the harmonic synthesizer and the filtered noise synthesizer. For the detailed explanation of this figure, see [13].

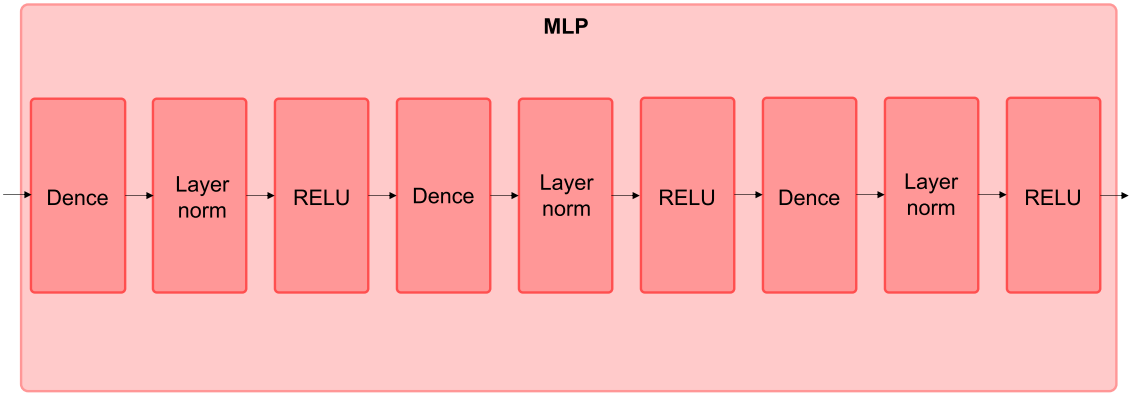


Fig. 2.13. MLP in the decoder of DDSP [13].

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 [28] 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，入力の音響信号から音色，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，これは入力の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

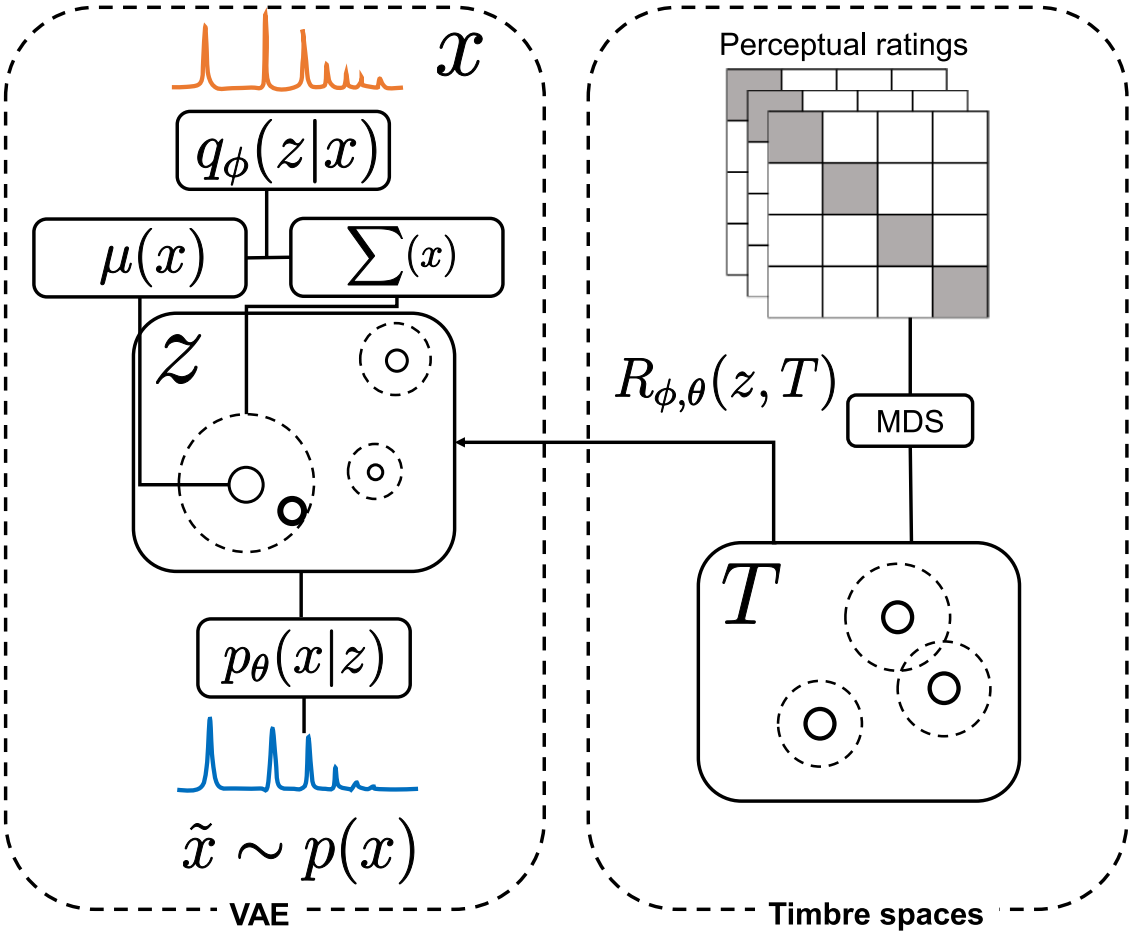


Fig. 2.14. Concept of regularizing VAE with perceptual metrics. Latent spaces and variables in VAE are forced to be matched to the perceptual timbre spaces structured by perceptual ratings. For the details of this figure, see [20].

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，解釈性を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 [20]．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 [29, 30] 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，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 [20] 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 [21, 22, 23]．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

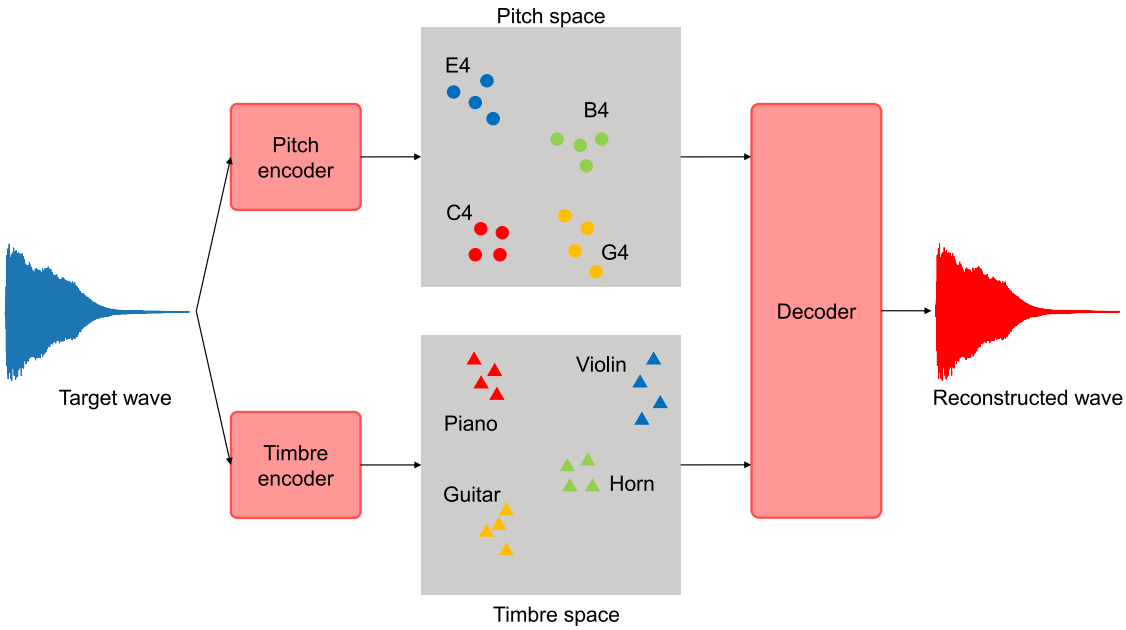


Fig. 2.15. VAE for learning disentangled pitch and timbre representations of music instrument sounds. For the details of this figure, see [23].

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，DDSP で用いられたラウドネス（時間的な音量変化）という特徴量は分離されておらず，ラウドネスに関する特徴量は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，音色と音高の他にラウドネスも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．

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

提案システムのより詳細な全体図を 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 $\boldsymbol{x} = [x(1), x(2), \cdots, x(L)]^{\mathrm{T}} \in \mathbb{R}^{L}$ 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．

$$\boldsymbol{X} = |\mathrm{STFT}_{\boldsymbol{\omega}}(\boldsymbol{x})| \in \mathbb{R}_{\geq 0}^{I \times J} \tag{3.1}$$

この振幅スペクトログラム $\boldsymbol{X}$ をエンコーダに入力し，音高，MFCC，及びラウドネスの 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，ラウドネスは DDSP よりも簡易的な抽出方法として，次式で計算する．

$$v_j = \sum_{i=1}^{I} x_{ij} \tag{3.2}$$

ここで，x_{ij} は振幅スペクトログラム $\boldsymbol{X}$ の要素を表す．式 (3.2) は，振幅スペクトログラム $|\boldsymbol{X}|$ の各列ベクトルの L_1 ノルムに対応する．このようにして得られたラウドネスは，全時間

フレームに関してまとめたベクトル $\boldsymbol{v} = [v_1, v_2, \cdots, v_J]^{\mathrm{T}} \in \mathbb{R}^J_{\geq 0}$ 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 $\boldsymbol{X}$ 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 $\boldsymbol{X}$ を次式で正規化し，正規化後の振幅スペクトログラム $\overline{\boldsymbol{X}} \in \mathbb{R}^{I \times J}_{\geq 0}$ に変換する．

$$\overline{x}_{ij} = \frac{x_{ij}}{v_j} \tag{3.3}$$

ここで，$\overline{x}_{ij}$ は正規化後の振幅スペクトログラム $\overline{\boldsymbol{X}}$ の要素である．従って，正規化後の振幅スペクトログラム $\overline{\boldsymbol{X}}$ から 2.3 節の方法で MFCC $\boldsymbol{C} \in \mathbb{R}^{K \times J}$ 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，及びラウドネスを抽出した後は，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 $\boldsymbol{z} \in \mathbb{R}^D$ の事前分布 $p_\theta(\boldsymbol{z})$ を原点対称多次元 Gauss 分布

$$\boldsymbol{z} \sim p_\theta(\boldsymbol{z}) \tag{3.4}$$

$$= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{3.5}$$

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 $\boldsymbol{\mu} \in \mathbb{R}^D$ 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，$\boldsymbol{0}$ 及び $\boldsymbol{I}$ 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 $\boldsymbol{C}$ から，$\boldsymbol{z}$ の近似事後分布 $q_\phi(\boldsymbol{z}|\boldsymbol{C})$（VAE のエンコーダ）及び尤度関数 $p_\theta(\boldsymbol{C}|\boldsymbol{z})$（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．

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

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

ここで，$\mathrm{diag}(\cdot)$ 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 $\boldsymbol{z}$ の $q_\phi(\boldsymbol{z}|\boldsymbol{C})$ 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 $\boldsymbol{z}$ の潜在空間を多次元 Gauss 分布仮定の下で学習できる．学習後は，潜在変数 $\boldsymbol{z}$ 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 $\hat{\boldsymbol{C}}$，及びラウドネスの 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 $\hat{\boldsymbol{X}}$ 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．

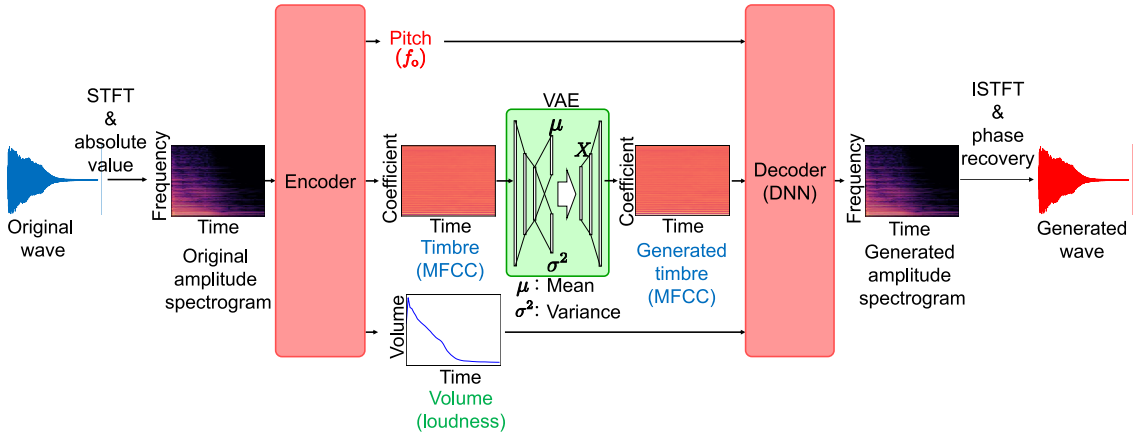


Fig. 3.1. Detailed process flow of the proposed sound generation system. The system consists of inner VAE and outer encoder-decoder. The outer encoder extracts pitch, MFCC, and loudness from an input amplitude spectrogram, Then, MFCC to the inner VAE. The generated MFCC obtained from the inner VAE is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

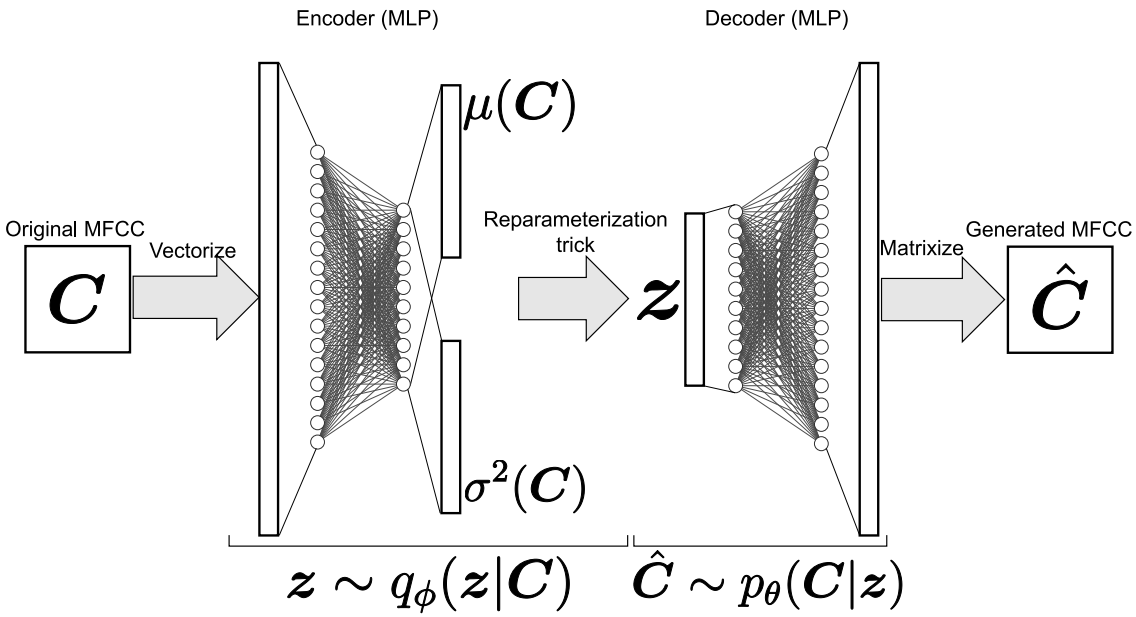


Fig. 3.2. Architecture of vanilla VAE for inferencing a generative model of MFCCs.

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．

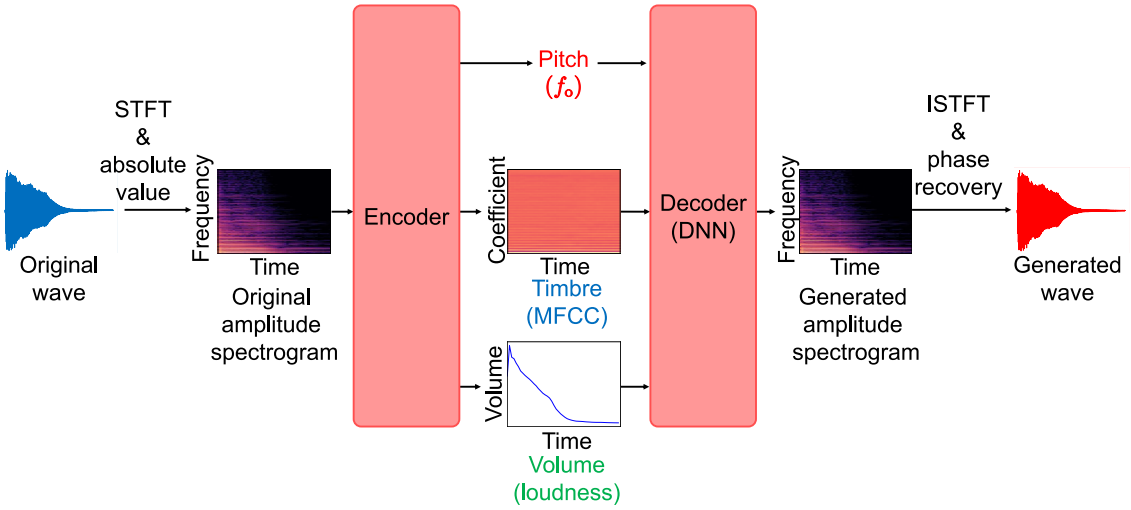


Fig. 3.3. Training process flow of the proposed DNN-based timbre decoder.

3.3　本論文で扱う問題

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，及びラウドネスの 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，及びラウドネスの 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，及びラウドネスの 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 及びラウドネスを入力とする 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，MLP 型 DNN，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 $\hat{\boldsymbol{C}}$ 及びラウドネス $\boldsymbol{v}$ を入力とし，そのような MFCC とラウドネスを持つ振幅スペクトログラム $\hat{\boldsymbol{X}}$ 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 $\hat{\boldsymbol{C}}$ 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 $\boldsymbol{C}$ 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 $\boldsymbol{X}$ 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 $\boldsymbol{X}$ 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 $\mathbf{x}$，$\mathbf{y}$，$\mathbf{p}$ 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 $\mathbf{x}$，$\mathbf{y}$，及び $\mathbf{p}$ と定義する．これまでに説明した通り，DNN デコーダは MFCC とラウドネスを入力特徴量とするため，MFCC $\boldsymbol{C}$ を 1 次元に整形（ベクトル化）し，ラウドネスベクトル $\boldsymbol{v}$ を結合した $KJ + J$ 次元のベクトルを入力ベクトル $\mathbf{x}$ とする．このとき，MFCC $\boldsymbol{C}$ 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 $\boldsymbol{X}$ 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 又は MSS 損失関数のいずれかを用いる．

$$L_{\mathrm{MSE}}(\mathbf{y}, \mathbf{p}) = \|\mathbf{y} - \mathbf{p}\|_2^2 \tag{3.8}$$

$$L_{\mathrm{MSS}}(\mathbf{y}, \mathbf{p}) = \|\mathbf{y} - \mathbf{p}\|_1 + \|\log \mathbf{y} - \log \mathbf{p}\|_1 \tag{3.9}$$

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

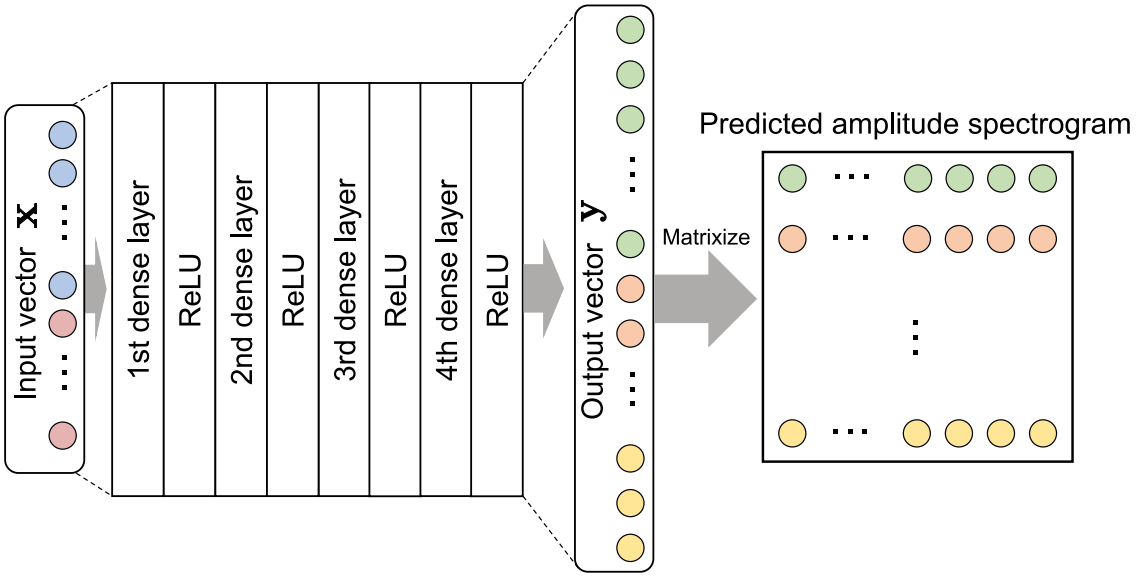


Fig. 3.4. Architecture of MLP used as the DNN decoder.

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．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 $\mathbf{h}_1^{(\mathrm{forward})}, \mathbf{h}_2^{(\mathrm{forward})}, \cdots, \mathbf{h}_J^{(\mathrm{forward})}$，未来から過去の方向（以後，逆方向と呼ぶ）の各 RNN の出力を未来側から順に $\mathbf{h}_1^{(\mathrm{backward})}, \mathbf{h}_2^{(\mathrm{backward})}, \cdots, \mathbf{h}_J^{(\mathrm{backward})}$ 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 及びラウドネス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 とラウ

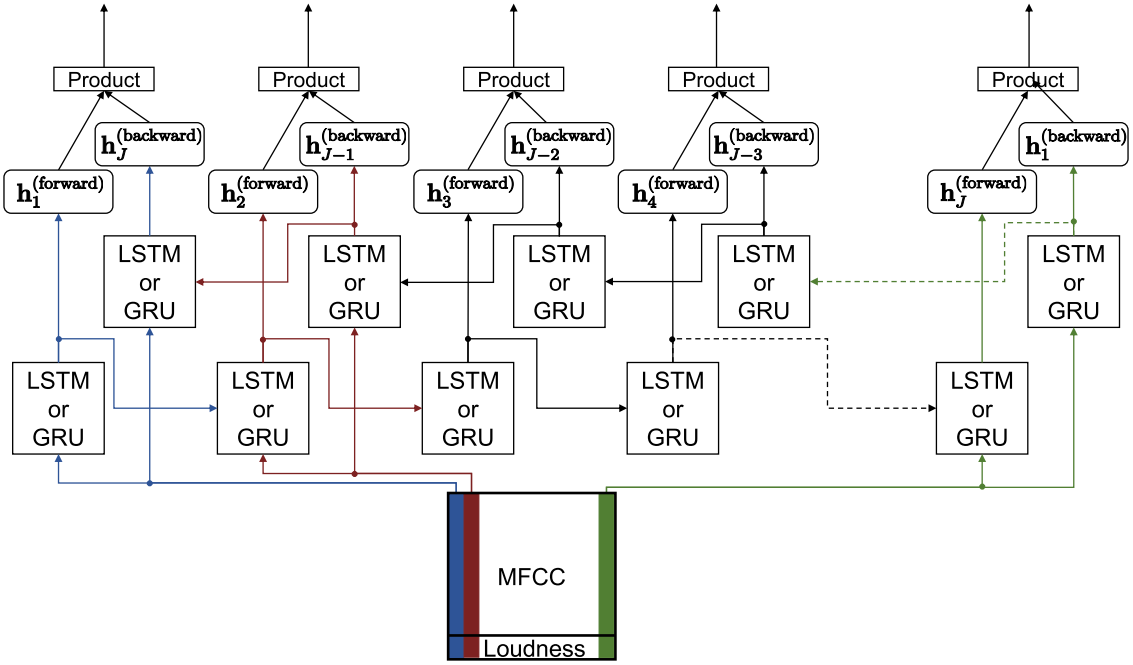


Fig. 3.5. Architecture of BiRNN used as the DNN decoder.

ドネスを次のように結合した行列を入力 $\mathbf{X}$ と定義する．

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{v}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{(K+1)\times J} \tag{3.10}$$

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 [24] 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 $\overline{\mathbf{X}} \in \mathbb{R}^{D\times J}$ 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 $\overline{\mathbf{x}}_j$ 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 $\overline{\mathrm{X}}$ 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 $\mathbf{H}$ 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．

$$\mathbf{f}_j = \sigma(\mathbf{W}^{\mathbf{f}}\overline{\mathbf{x}}_j + \mathbf{R}^{\mathbf{f}}\mathbf{h}_{j-1} + \mathbf{b}^{\mathbf{f}}) \tag{3.11}$$
$$\mathbf{i}_j = \sigma(\mathbf{W}^{\mathbf{i}}\overline{\mathbf{x}}_j + \mathbf{R}^{\mathbf{i}}\mathbf{h}_{j-1} + \mathbf{b}^{\mathbf{i}}) \tag{3.12}$$
$$\tilde{\mathbf{c}}_j = \tanh(\mathbf{W}^{\tilde{\mathbf{c}}}\overline{\mathbf{x}}_j + \mathbf{R}^{\tilde{\mathbf{c}}}\mathbf{h}_{j-1} + \mathbf{b}^{\tilde{\mathbf{c}}}) \tag{3.13}$$
$$\mathbf{c}_j = \mathbf{f}_j \circ \mathbf{c}_{j-1} + \mathbf{i}_j \circ \tilde{\mathbf{c}}_j \tag{3.14}$$
$$\mathbf{o}_j = \sigma(\mathbf{W}^{\mathbf{o}}\overline{\mathbf{x}}_j + \mathbf{R}^{\mathbf{o}}\mathbf{h}_{j-1} + \mathbf{b}^{\mathbf{o}}) \tag{3.15}$$
$$\mathbf{h}_j = \mathbf{o}_j \circ \tanh(\mathbf{c}_j) \tag{3.16}$$

ここで，$\sigma(\cdot)$ はベクトルに対するシグモイド関数，$\mathbf{W}^{\mathbf{f}}$，$\mathbf{W}^{\mathbf{i}}$，$\mathbf{W}^{\tilde{\mathbf{c}}}$，及び $\mathbf{W}^{\mathbf{o}}$ は時間 j における入力ベクトル $\overline{\mathbf{x}}_j$ に対する重み係数行列，$\mathbf{R}^{\mathbf{f}}$，$\mathbf{R}^{\mathbf{i}}$，$\mathbf{R}^{\tilde{\mathbf{c}}}$，及び $\mathbf{R}^{\mathbf{o}}$ 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，$\mathbf{b}^{\mathbf{f}}$，$\mathbf{b}^{\mathbf{i}}$，$\mathbf{b}^{\tilde{\mathbf{c}}}$，及び $\mathbf{b}^{\mathbf{o}}$ は，それぞれの係数に対する

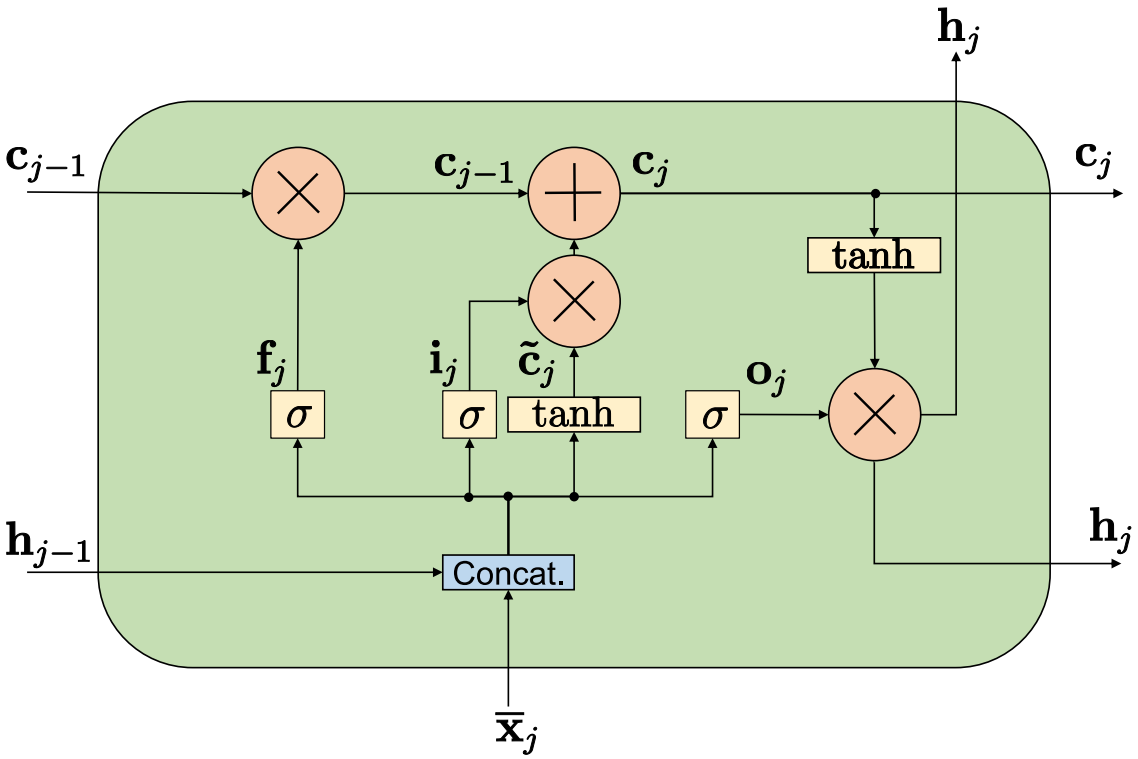


Fig. 3.6. Structure of LSTM unit.

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 $\mathbf{c}_{j-1}$ からど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 $\overline{\mathbf{x}}_j$ 及び時間 $j-1$ における出力ベクトル $\mathbf{h}_{j-1}$ からど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 $\mathbf{h}_j$ 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 $\overline{\mathbf{x}}_j$ 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 $\mathbf{h}_j^{(\mathrm{forward})}$ 及び逆方向の出力ベクトル $\mathbf{h}_{J-j}^{(\mathrm{backward})}$ が出力され，その要素毎の積を取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．

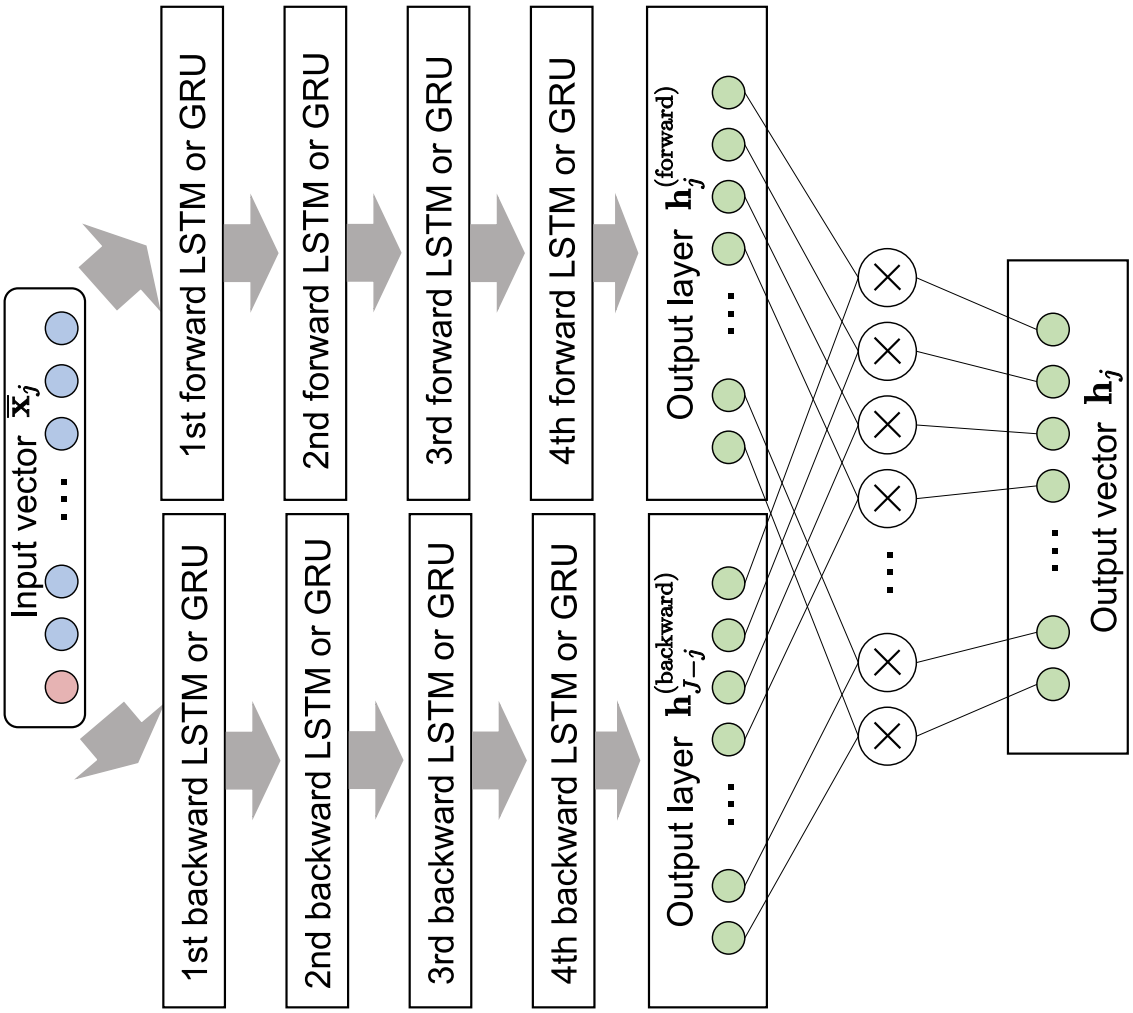


Fig. 3.7. Data flow in the multi-layer BiRNN at time frame j.

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 における双方向の RNN に GRU[25] 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 $\overline{\mathbf{X}} \in \mathbb{R}^{D \times J}$ 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 $\overline{\mathbf{x}}_j$ 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 $\overline{\mathrm{X}}$ の特徴量の次元数及びこの GRU から

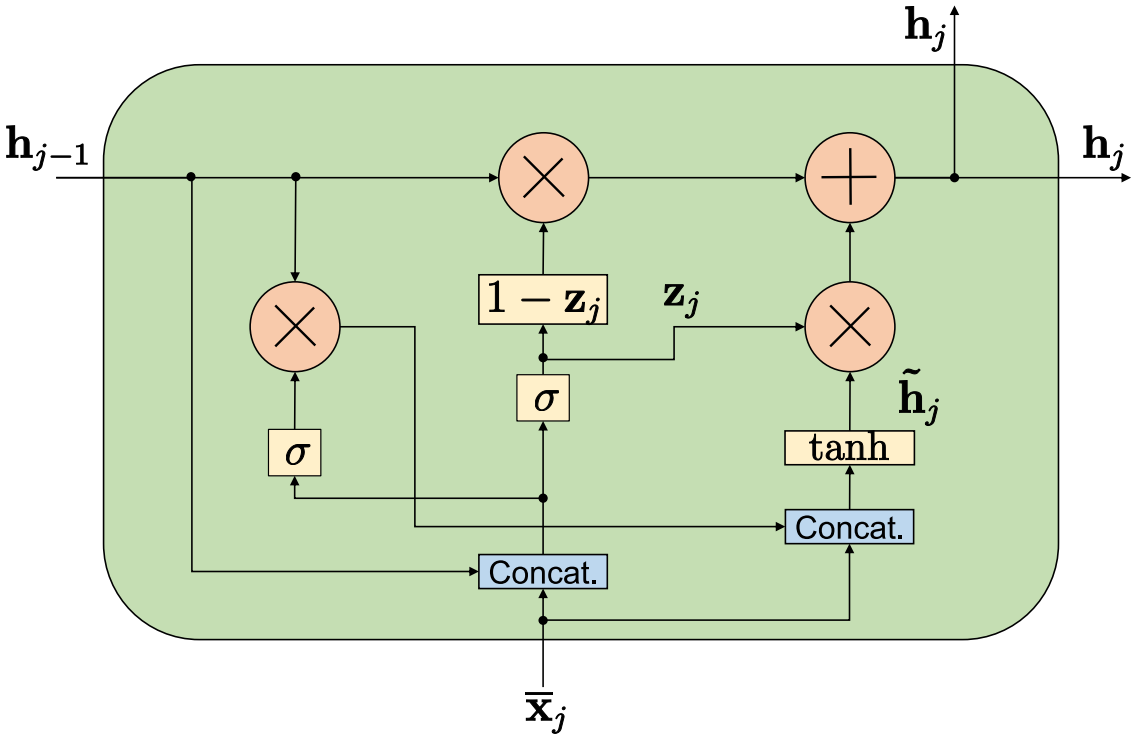


Fig. 3.8. Structure of GRU.

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．

$$\mathbf{r}_j = \sigma(\mathbf{W}^{\mathbf{r}}\mathbf{x}_j + \mathbf{R}^{\mathbf{r}}\mathbf{h}_{j-1} + \mathbf{b}^{\mathbf{r}}) \tag{3.17}$$

$$\tilde{\mathbf{h}}_j = \tanh(\mathbf{W}^{\tilde{\mathbf{h}}}\mathbf{x}_j + \mathbf{R}^{\tilde{\mathbf{h}}}(\mathbf{r}_j \circ \mathbf{h}_{j-1}) + \mathbf{b}^{\tilde{\mathbf{h}}}) \tag{3.18}$$

$$\mathbf{z}_j = \sigma(\mathbf{W}^{\mathbf{z}}\mathbf{x}_j + \mathbf{R}^{\mathbf{z}}\mathbf{h}_{j-1} + \mathbf{b}^{\mathbf{z}}) \tag{3.19}$$

$$\mathbf{h}_j = (1 - \mathbf{z}_j) \circ \mathbf{h}_{j-1} + \mathbf{z}_j \circ \tilde{\mathbf{h}}_j \tag{3.20}$$

ここで，$\mathbf{W}^{\mathbf{z}}$，$\mathbf{W}^{\mathbf{r}}$，及び $\mathbf{W}^{\tilde{\mathbf{h}}}$ は時間 j における入力ベクトル $\overline{\mathbf{x}}_j$ に対する重み係数行列，$\mathbf{R}^{\mathbf{z}}$，$\mathbf{R}^{\mathbf{r}}$，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，$\mathbf{b}^{\mathbf{z}}$，$\mathbf{b}^{\mathbf{r}}$，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 $\overline{\mathbf{x}}_j$ 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 での説明と同様に，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 の相対誤差を用いて，各 DNN の予測精度の比較評価を行う．4.5 節では，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，musical instrument digital interface（MIDI）音源の Roland SVC に含まれる楽器音のうち，Table 4.1 に示すピアノ 4 種類及びギター 4 種類の

Table 4.1. Instrument type and name of MIDI signals used in this experiment

Instrument type	Instrument name
Piano	Piano1
Piano	Piano2
Piano	Piano3
Piano	Honky-tonk
Guitar	Nylon-str.Gt
Guitar	Steel-str.Gt
Guitar	Jazz Gt
Guitar	Muted GT

Table 4.2. 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

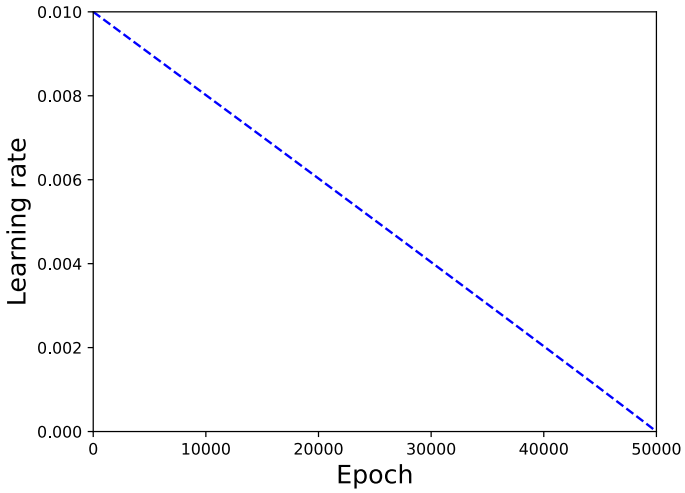


Fig. 4.1. Relationship between the learning rate and the number of epochs.

計 8 種類の楽器音を用いた．この MIDI 音源 8 種類と MIDI 音源 8 種類のそれぞれに対して 4 種類のイコライザで音色を変化させたデータを加算して合計でピアノ 20 種類及びギター 20 種類の合計 40 種類の楽器音信号を生成した．イコライザの種類としては，低音域を強調したもの，高周波を強調したもの，コーラスをかけたもの，及びリバーブ（残響）をかけたものの 4 種類である．イコライザはあくまで楽器音データを増やすことが目的であるため，僅かに音色が変化する程度とした．用意した 40 種類の音源の内，ピアノ 18 種類及びギター 18 種類の計 36 種類を学習データとして各 DNN の最適化に用いた．残りのピアノ 2 種類及びギター 2 種類の計 4 種類を検証データとして用いた．最終的な評価は検証データをテストデータとして代用して行った．音源は C3 から B5 までの 36 音（3 オクターブ）を MIDI 音源から作成したが，本論文では C4 音，E4 音，G4 音，及び C5 音の計 4 種類の音についてのみ結果を記載する．MIDI 音源から作成した音はサンプリング周波数が 44.1 kHz であり，テンポを 120 bpm に設定した際の 4 分音符 1 つで構成された 1.18 s の単音の音響信号である．なお，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 4.2 に示す．

各 DNN の学習条件を説明する．各 DNN において，損失関数は MSE ロス及び MSS ロスの 2 種類を用いた．また，学習のエポック数は全ての DNN において 50000 回に設定した．

さらに，学習率は初期値を 0.01 と設定し，Fig. 4.1 に示すようにエポックス数 50000 回で 0 となるように線形なスケジューリングを設定した．本論文の実験では，過学習を防ぐために，MLP 型 DNN は 10000 回，BiLSTM 型 DNN 及び BiGRU 型 DNN は 1000 回を超えて，検証データに対する誤差の減少が見られない場合は繰り返し回数が 50000 回に到達する前でも学習を早期終了させた．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の結果を MLP 型，BiLSTM 型，及び BiGRU 型をそれぞれ 4.3.1 項から 4.3.3 項に示し，それぞれ主観的に評価を行う．4.2 節で述べた通り，本論文では C4 音，E4 音，G4 音，及び C5 音の計 4 種類の音に対して実験を行ったが，本節では G4 音のテストデータに含まれるピアノ 1 種類及びギター 1 種類の計 2 種類のみの予測結果を示す．その他の結果は付録 A に示す．

4.3.1　MLP 型 DNN

Fig. 4.2(a) 及び (b) は，ピアノの G4 音を MSE ロスで学習した MLP 型 DNN の入力と予測結果のパワースペクトログラムをそれぞれ示している．また，Fig. 4.3(a) 及び (b) は，ピアノの G4 音を MSS ロスで学習した MLP 型 DNN の入力と予測結果のパワースペクトログラムをそれぞれ示している．同様に，Figs. 4.4 及び 4.5 には，それぞれギターの G4 音の入力と予測結果のパワースペクトログラムを，MSE ロス及び MSS ロスの MLP 型 DNN のそれぞれで示している．

まず，ピアノの音を対象としている Figs. 4.2 及び 4.3 を比較すると，MSE ロスではピアノ音の持つ調波構造の骨組みのようなものが一部に確認されるが，正確な予測は全くできていないことが分かる．また，MSS ロスの予測結果については，MSE ロスの場合よりもさらに精度の低い予測となっている．信号の後半（約 0.5 s 以降）にパワーの大きい成分がいくつか散見されるが，調波構造が再現できておらず，MLP 型 DNN での振幅スペクトログラムの予測は非常に困難であることが分かる．

次に，ギターの音を対象としている Figs. 4.4 及び 4.5 を比較すると，高精度な振幅スペクトログラムの予測が全く行えていないことがわかる．MSS ロスの予測結果では，ピアノの音の場合と同様に信号の後半（約 0.5 s 以降）にパワーの大きい成分が散見されるが，ラベルの振幅スペクトログラムとはかけ離れたものとなっており，MLP 型 DNN での予測は困難である．

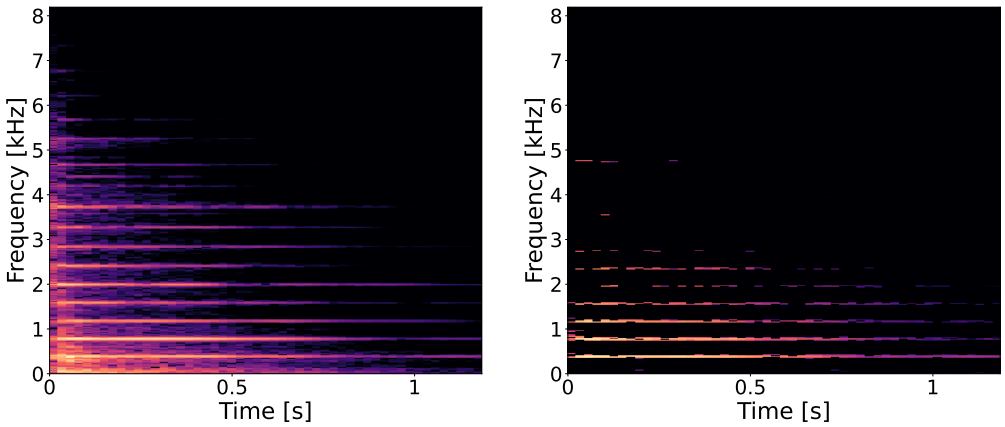


Fig. 4.2. The power spectrograms of (a) the input piano G4 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

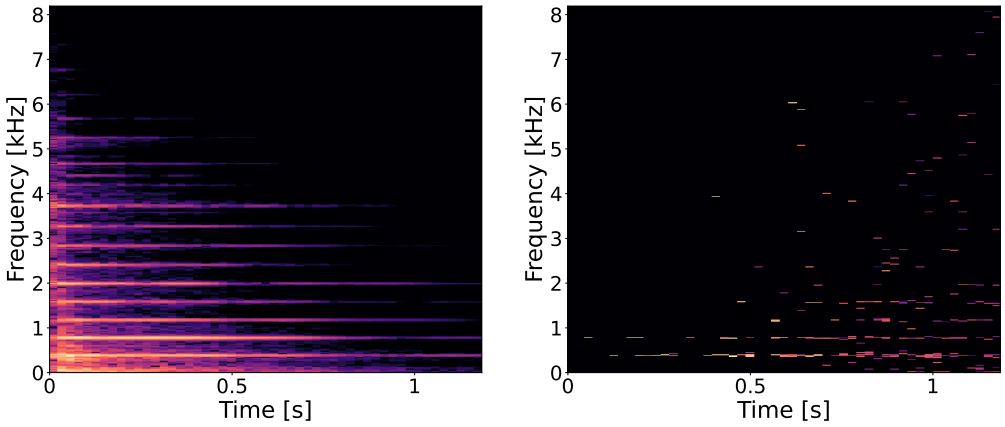


Fig. 4.3. The power spectrograms of (a) the input piano G4 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.

4.3.2 BiLSTM 型 DNN

Fig. 4.6(a) 及び (b) は，ピアノの G4 音を MSE ロスで学習した BiLSTM 型 DNN の入力と予測結果のパワースペクトログラムをそれぞれ示している．また，Fig. 4.7(a) 及び (b) は，ピアノの G4 音を MSS ロスで学習した BiLSTM 型 DNN の入力と予測結果のパワースペクトログラムをそれぞれ示している．同様に，Figs. 4.8 及び 4.9 には，それぞれギターの G4 音の入力と予測結果のパワースペクトログラムを，MSE ロス及び MSS ロスの BiLSTM 型 DNN のそれぞれで示している．

まず，ピアノの音を対象としている Figs. 4.6 及び 4.7 を比較すると，ピアノ音の持つ調波構造やその時間変化等がどちらもある程度の精度で予測されていることが分かる．但し，

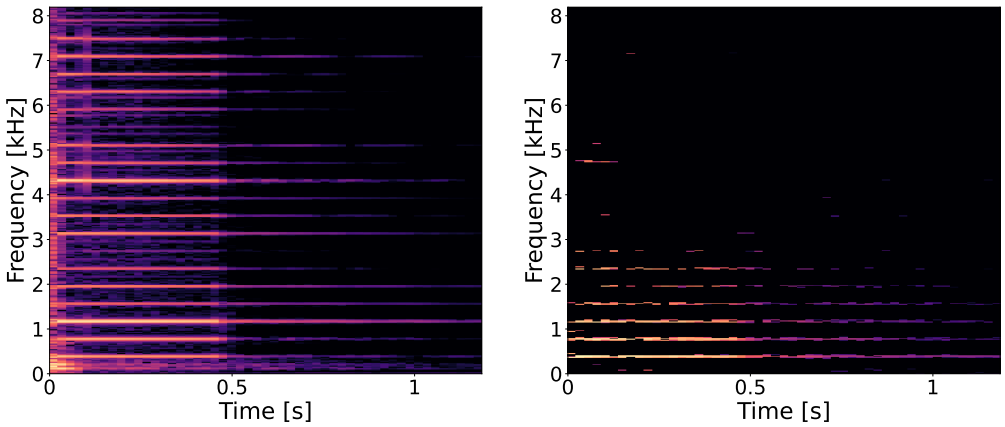


Fig. 4.4. The power spectrograms of (a) the input guitar G4 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

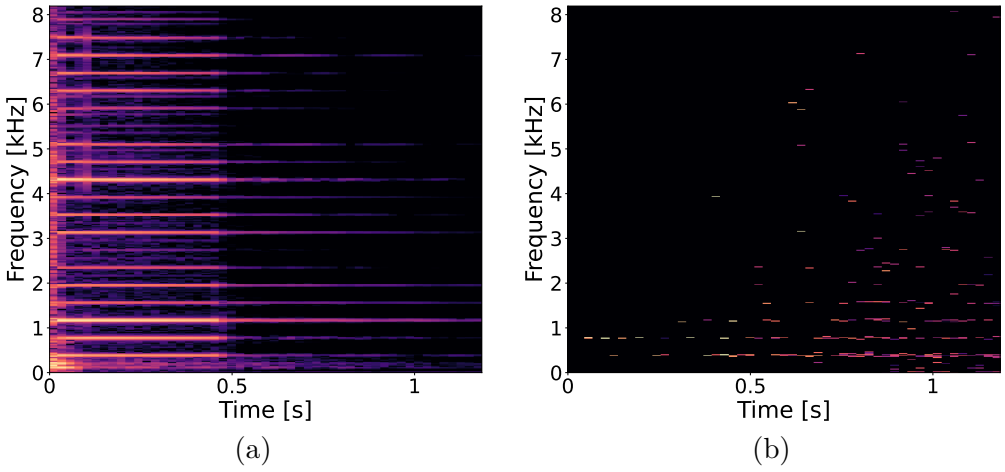


Fig. 4.5. The power spectrograms of (a) the input guitar G4 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.

MSE ロスを学習に用いた Fig. 4.6(b) を見ると，高調波成分（4 kHz 以上）の調波構造が入力のそれとは異なる結果となってしまっていることが分かる．一方，MSS ロスを学習に用いた Fig. 4.7(b) では，調波構造の予測精度は MSE ロスより高いが，調波構造の間の歪みの予測精度は低いことが分かる．これらの違いから，BiLSTM 型において MSE ロスは MSS ロスより各周波数の時間変化におけるパワーの強弱，打撃音，及び低周波数帯域の予測精度が高いことが分かる．対して，MSS ロスは，調波構造の予測精度は高く予測ができることが分かる．

次に，ギターの音を対象としている Figs. 4.8 及び 4.9 を比較すると，ピアノの場合と同じく調波構造はある程度予測され，ピアノにはない約 0.5 s で見られるパワーの減衰がある程度予測できた．MSE ロスは 0 s で見られる打撃音をある程度予測できたが，MSS ロスはあまり予測できていない．

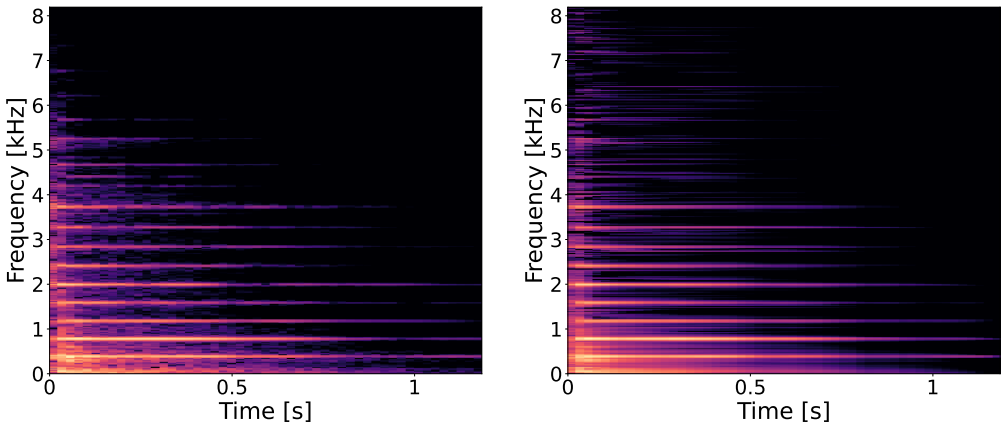


Fig. 4.6. The power spectrograms of (a) the input piano G4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

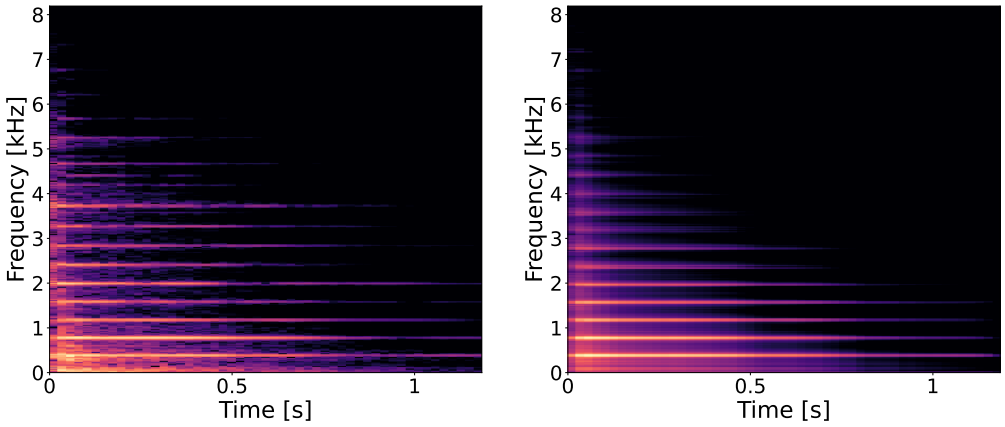


Fig. 4.7. The power spectrograms of (a) the input piano G4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.

4.3.3 BiGRU 型 DNN

Fig. 4.10(a) 及び (b) は，ピアノの G4 音を MSE ロスで学習した BiGRU 型 DNN の入力と予測結果のパワースペクトログラムをそれぞれ示している．また，Fig. 4.11(a) 及び (b) は，ピアノの G4 音を MSS ロスで学習した BiGRU 型 DNN の入力と予測結果のパワースペクトログラムをそれぞれ示している．同様に，Figs. 4.12 及び 4.13 には，それぞれギターの G4 音の入力と予測結果のパワースペクトログラムを，MSE ロス及び MSS ロスの BiGRU 型 DNN のそれぞれで示している．

まず，ピアノの音を対象としている Figs. 4.10 及び 4.11 を比較すると，MSS ロスを学習に用いた Fig. 4.11(b) はピアノ音の持つ調波構造やその時間変化等がどちらもある程度の精度で

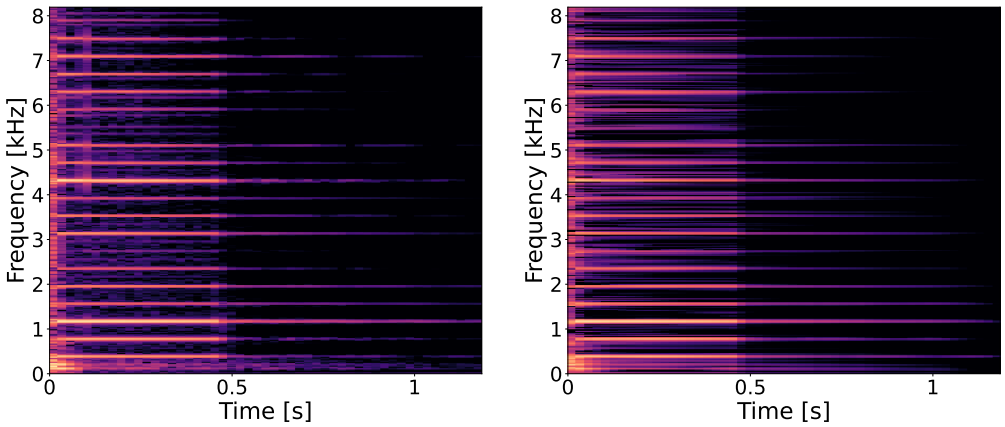


Fig. 4.8. The power spectrograms of (a) the input guitar G4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

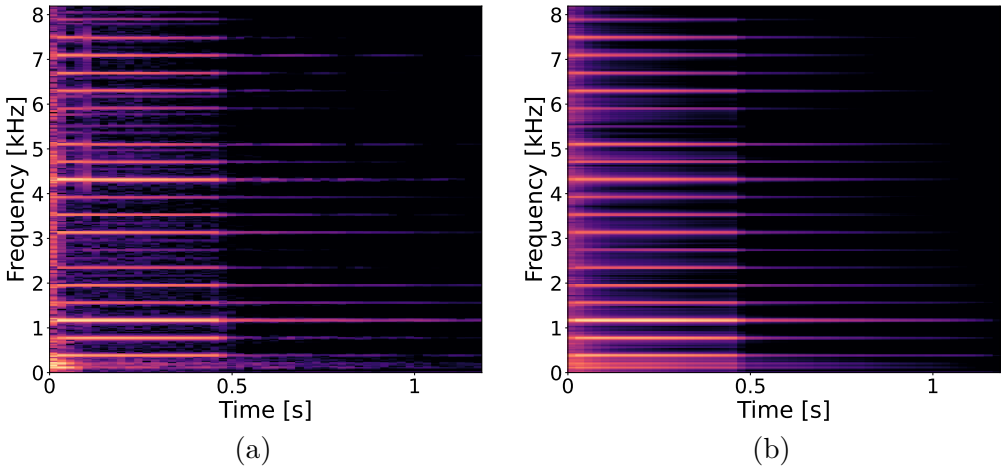


Fig. 4.9. The power spectrograms of (a) the input guitar G4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.

予測されていることが分かる．一方，MSE ロスを学習に用いた Fig. 4.10(b) を見ると，高調波成分（3 kHz 以上）の調波構造が入力のそれとは異なる結果となってしまっているが，低周波帯域の予測精度は MSS ロスより精度が高いことが分かる．これらの違いから，BiGRU 型において MSE ロスは MSS ロスより各周波数の時間変化におけるパワーの強弱，打撃音，及び低周波数帯域の予測精度が高いことが分かる．対して，MSS ロスは，調波構造の予測精度は高いことが分かる．

次に，ギターの音を対象としている Figs. 4.12 及び 4.13 を比較すると，ギター音の持つ調波構造の予測精度はあまり高くないが，ピアノにはない約 0.5 s で見られるパワーの減衰がある程度予測できた．MSE ロスは調波構造に対するの歪みが大きい．一方，MSS ロスは歪みは少ないが，調波構造において周波数が高くなるにつれてパワーが弱くなるという，入力のそれとは異なる結果となってしまっていることが分かる．

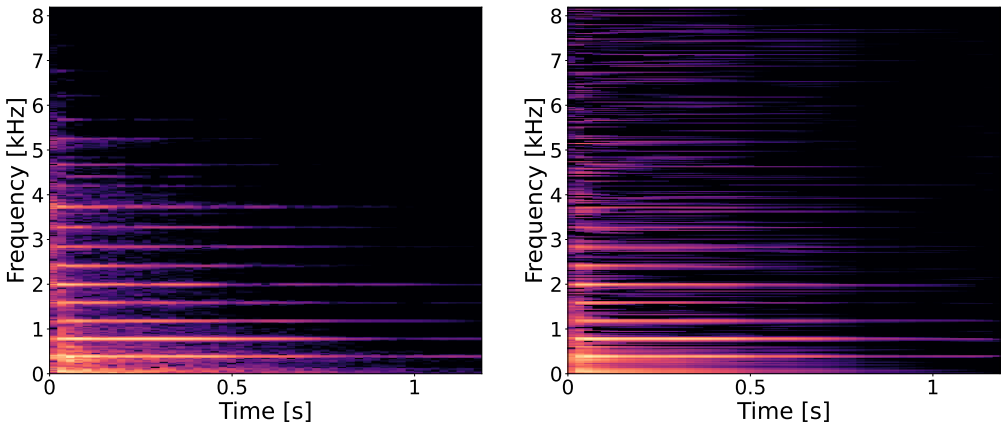


Fig. 4.10. The power spectrograms of (a) the input piano G4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

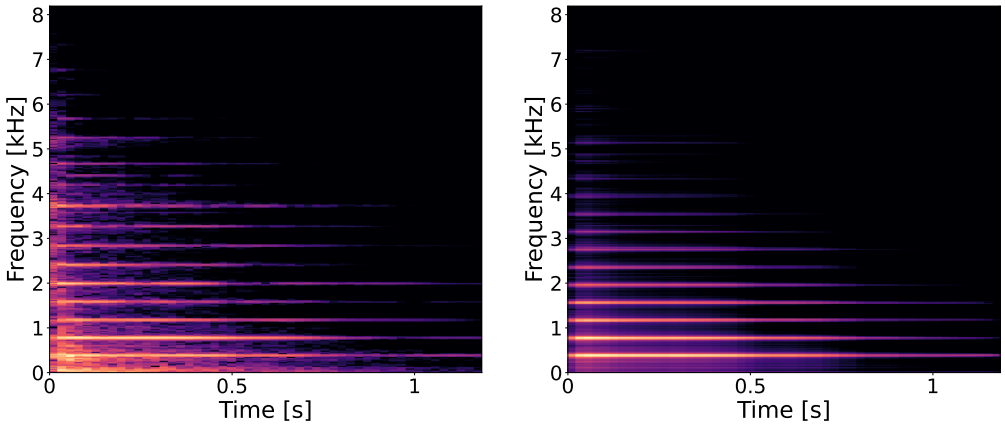


Fig. 4.11. The power spectrograms of (a) the input piano G4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.

4.4 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主観的に評価を述べた．本節では，MFCC に対する予測精度の客観的評価を行う．本論文では，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．MFCC 相対二乗誤差 (MFCC relative squared error: MRSE)[31] は次式で表される．

$$\mathrm{MRSE} = 10 \log \frac{\sum_{j=1}^{J} \sum_{k=2}^{14} (c_{kj} - \hat{c}_{kj})^2}{\sum_{j=1}^{J} \sum_{k=2}^{14} (c_{kj})^2} \ [\mathrm{dB}] \tag{4.1}$$

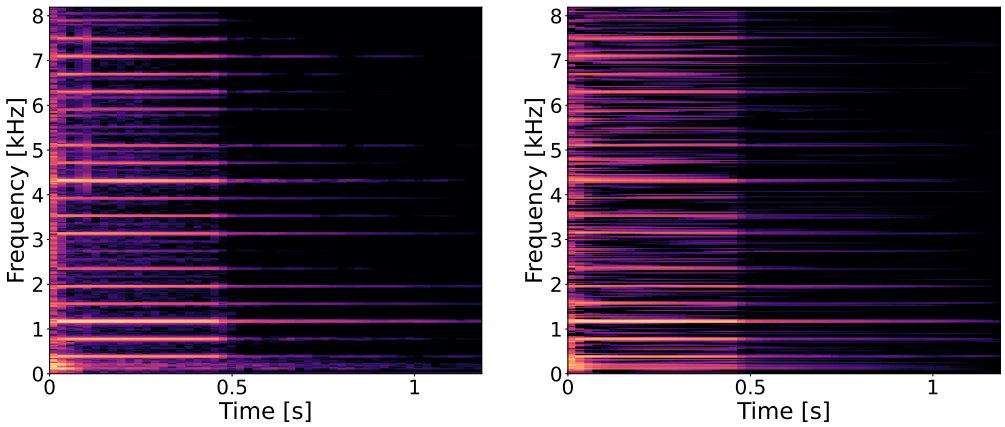


Fig. 4.12. The power spectrograms of (a) the input guitar G4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

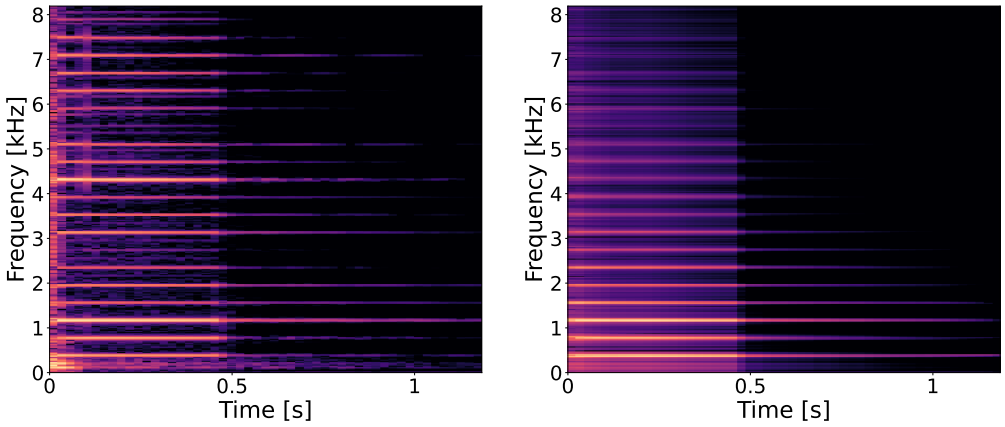


Fig. 4.13. The power spectrograms of (a) the input guitar G4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.

ここで, $k = 1, 2, \cdots, K$ は MFCC の次元のインデクスを示す．また, c_{kj} 及び $\hat{c}_{kj}$ はそれぞれ $\boldsymbol{C}_{kj}$ 及び $\hat{\boldsymbol{C}}_{kj}$ の要素である．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．

4.3 節で示した結果を含め，4.2 節で説明したテストデータのピアノ 2 種類及びギター 2 種類の計 4 種類に対して，各 DNN 及び各損失関数で MRSE を算出した結果を Table 4.5–4.4 に示す．いずれも値が小さいものほど高い精度で音色を予測できていることを表す．

Table 4.3 は 4.3 節に示したピアノの 1 番目のテストデータの予測結果の評価を示す．ピアノの 1 番目のテストデータの予測は BiLSTM 型 DNN が両誤差関数で 4 音に対して音色を保った振幅スペクトラムの予測が行えた．それに対し，MSE ロスを用いた BiGRU 型 DNN 及び MLP 型 DNN は低い精度となった．Table 4.4 はピアノの 2 番目のテストデータの予測結果の評価を示す．ピアノの 2 番目のテストデータの予測もピアノの 1 番目のテストデータ

Table 4.3. MRSE [dB] of the predicted amplitude spectrogram of the first piano test signal for each DNN

Note	MLP-type DNN		BiLSTM-type DNN		BiGRU-type DNN	
	MSE loss	MSS loss	MSE loss	MSS loss	MSE loss	MSS loss
C4	-36.2507	-31.1617	-51.32951	**-56.5120**	-38.18392	-43.847
E4	-36.5973	-27.5703	-46.10653	**-51.2356**	-30.49936	-47.2455
G4	-33.7590	-29.2400	-49.00426	**-52.2066**	-32.0031	-47.3381
C5	-31.9831	-29.6216	-47.92678	**-50.4268**	-40.02535	-47.1641

Table 4.4. MRSE [dB] of the predicted amplitude spectrogram of the second piano test signal for each DNN

Note	MLP-type DNN		BiLSTM-type DNN		BiGRU-type DNN	
	MSE loss	MSS loss	MSE loss	MSS loss	MSE loss	MSS loss
C4	-35.2115	-30.2584	-53.28281	**-54.4526**	-35.59197	-42.8832
E4	-36.0495	-28.8597	-47.36151	**-52.5985**	-32.38396	-46.6054
G4	-33.9487	-27.6613	-49.45064	**-51.9225**	-30.81845	-46.4259
C5	-33.9500	-30.0948	-52.49461	**-52.8652**	-39.18496	-48.7106

同様に BiLSTM 型 DNN が両誤差関数で 4 音に対して音色を保った振幅スペクトラムの予測が行えた．Table 4.5 は 4.3 節に示したギターの 1 番目のテストデータの予測結果の評価を示す．ギターの 1 番目のテストデータの予測は BiLSTM 型 DNN は MSE 及び MSS の両誤差関数で 4 音に対して音色を保った振幅スペクトラムの予測が行えた．Table 4.6 はギターの 2 番目のテストデータの予測結果の評価を示す．ギターの 2 番目のテストデータの予測は誤差関数に MSS ロスを用いた BiLSTM 型 DNN 及び BiGRU 型 DNN が 4 音に対して音色を保った振幅スペクトラムの予測が行えた．Table 4.3–4.6 の結果をまとめると，BiLSTM 型 DNN が 4 種類の楽器音信号に対する振幅スペクトラムの予測の精度が高い．さらに，誤差関数は BiLSTM 型 DNN 及び BiGRU 型 DNN において，MSS ロスの方が高精度に予測できるモデルの学習を実現している．それに対して，MLP 型 DNN は全楽器及び全音高に対して精度は低い．これは，BiLSTM 型 DNN 及び BiGRU 型 DNN は時間フレーム方向の連続性を考慮しながらモデルの学習ができることに起因していると思われる．

4.5 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

Table 4.5. MRSE [dB] of the predicted amplitude spectrogram of the first guitar test signal for each DNN

Note	MLP-type DNN		BiLSTM-type DNN		BiGRU-type DNN	
	MSE loss	MSS loss	MSE loss	MSS loss	MSE loss	MSS loss
C4	-21.5834	-30.7006	**-48.5011**	-48.2526	-42.6628	-24.9285
E4	-23.9157	-33.1786	**-47.0733**	-43.8630	-40.8191	-31.5421
G4	-22.6440	-34.5355	**-50.3177**	-48.2616	-42.6868	-30.3510
C5	-32.0672	-32.4831	**-45.7863**	-45.3116	-41.4711	-32.7398

Table 4.6. MRSE [dB] of the predicted amplitude spectrogram of the second guitar test signal for each DNN

Note	MLP-type DNN		BiLSTM-type DNN		BiGRU-type DNN	
	MSE loss	MSS loss	MSE loss	MSS loss	MSE loss	MSS loss
C4	-39.2519	-32.1298	-42.0985	**-50.3707**	-31.1742	-47.7453
E4	-40.0790	-25.0016	-40.5874	**-51.2291**	-29.88234	-45.8482
G4	-41.8984	-29.5030	-39.7046	**-50.7956**	-29.6234	-43.046
C5	-43.4077	-27.3533	-39.9455	**-48.1906**	-30.60489	-45.8786

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，テストデータの内ピアノ 1 種類及びギター 1 種類の計 2 種類について，誤差関数に MSE ロス及び MSS ロスを用いた MLP 型 DNN，BiLSTM 型 DNN，及び BiGRU 型 DNN のそれぞれの結果を示し，予測されたパワースペクトログラムの比較及び評価を行った．4.4 節では，テストデータのピアノ 2 種類及びギター 2 種類の計 4 種類を用いて，C4 音，E4 音，G4 音，及び C5 音における MFCC に対する予測精度の客観的評価を行った．客観的評価の指標として，MFCC 相対二乗誤差誤差を用い，結果として，DNN3 種類と損失関数 2 種類の計 6 種類の学習済み DNN で全楽器音信号及び全音高に対して高い精度で振幅スペクトラムの予測が行えたのは MSS を損失関数に用いた BiLSTM 型 DNN であることが分かった．次章では，本論文をまとめる．

第 5 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，誤差関数としてそれぞれ MSE ロス及び MSS ロスを適用して学習した．その学習した各 DNN で，テストデータを用いて行った予測実験の結果としては，双方向再帰型 DNN である BiLSTM 型 DNN 及び BiGRU 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．

最後に今後の課題を述べる．提案音生成システムにおいて必要不可欠である音高，音色，及び音量から振幅スペクトログラムの予測するデコーダが作成でき，提案音生成システムの問題が解決された．そのため，提案音生成システムの核となる VAE を用いた楽器音の音色特徴量の抽出及び変換の実装が最も優先される課題である．

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．結果が出るまで時間がかかり悩んでいましたが，親身になってご指導いただいたおかげでよい結果を得ることができました．

本論の副査である柿元健准教授には，論文の構成や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．

北村研究室の先輩である専攻科 2 年の渡辺瑠伊氏，岩瀬裕太氏，大藪宗一郎氏，梶谷奈未氏には，Python に関するアドバイス，サーバに関する知識等をはじめ，数々のご支援をいただきました．また，北村研究室同期の蓮池郁也氏，溝渕悠朔氏，村田佳斗市，細谷泰稚氏には深層学習 Python に関するアドバイスを沢山いただきました．1 年に亘る研究室生活を様々な面で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を金銭的に支え，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. Gnann, "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] P. Esling, A. Chemla-RomenuSantos, and A. Bitton, “Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics,” *Proc. Internatinal Conference on Digital Audio Effects*, 2018.

[21] 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.

[22] 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.

[23] 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.

[24] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” *IEEE Trans. Neural Networks and Learning Sys-*

tems, vol. 28, no. 10, pp. 2222–2232, 2016.

[25] 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.

[26] D. O'Shaughnessy, *Speech Communication: Human and Machine.* Addison-Wesley, 1987.

[27] 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.

[28] 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.

[29] 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.

[30] S. Lakatos, "A common perceptual space for harmonic and percussive timbres," *Perception & psychophysic*, vol. 62, no. 7, pp. 1426-1439, 2000.

[31] 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.

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.12 に MLP 型 DNN の結果を示す．Figs. A.13–A.24 に BiLSTM 型 DNN の結果を示す．Figs. A.25–A.36 に BiGRU 型 DNN の結果を示す．

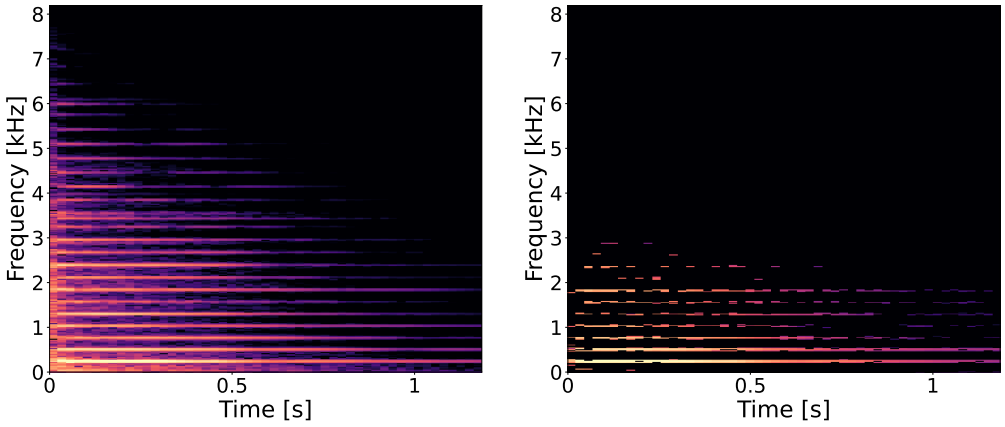


Fig. A.1. The power spectrograms of (a) the input piano C4 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

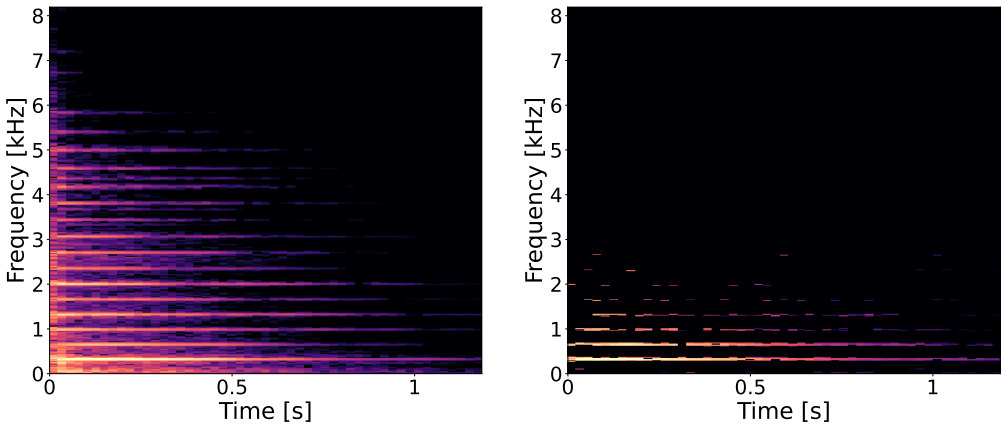


Fig. A.2. The power spectrograms of (a) the input piano E4 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

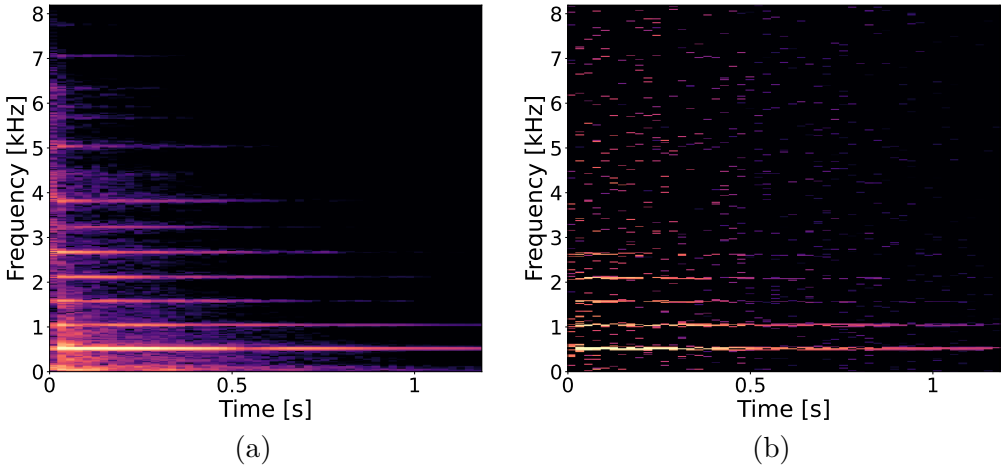


Fig. A.3. The power spectrograms of (a) the input piano C5 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

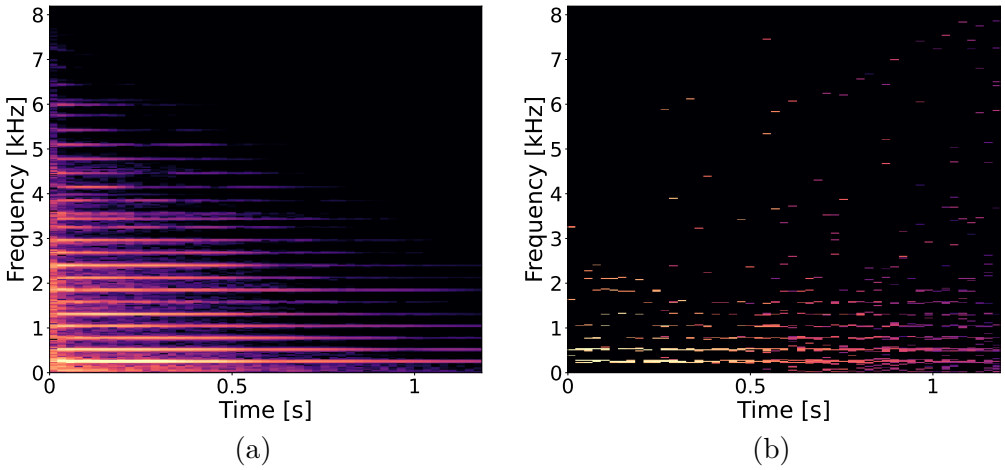
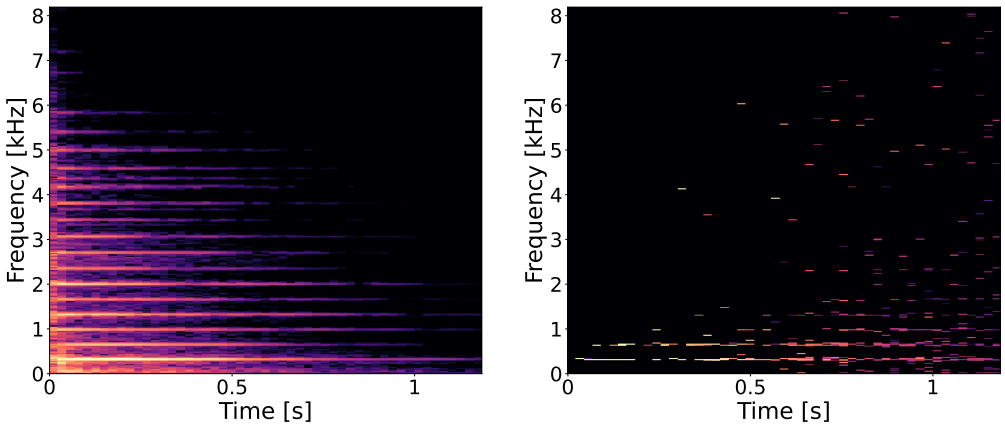
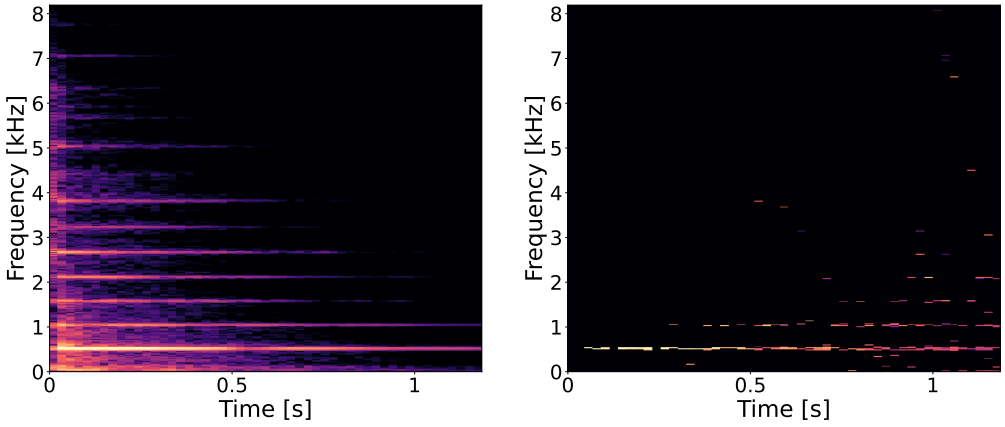


Fig. A.4. The power spectrograms of (a) the input piano C4 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.



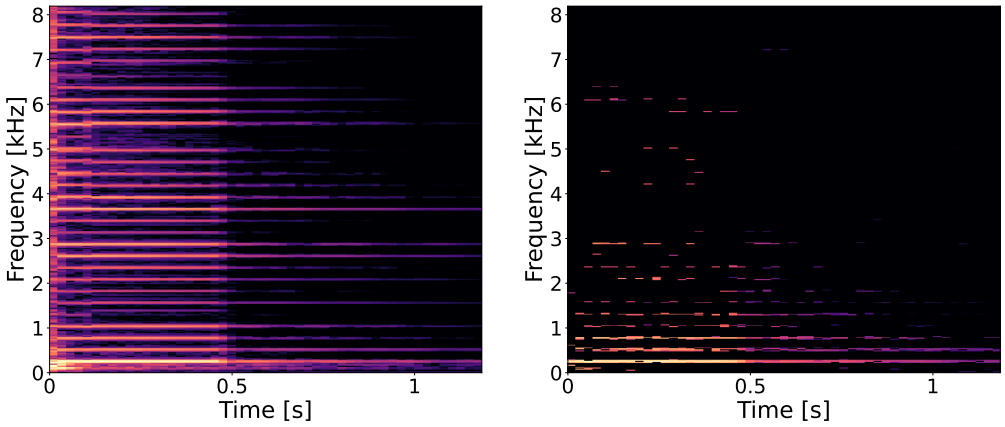
(a)　(b)

Fig. A.5. The power spectrograms of (a) the input piano E4 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.



(a)　(b)

Fig. A.6. The power spectrograms of (a) the input piano C5 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.



(a)　(b)

Fig. A.7. The power spectrograms of (a) the input guitar C4 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

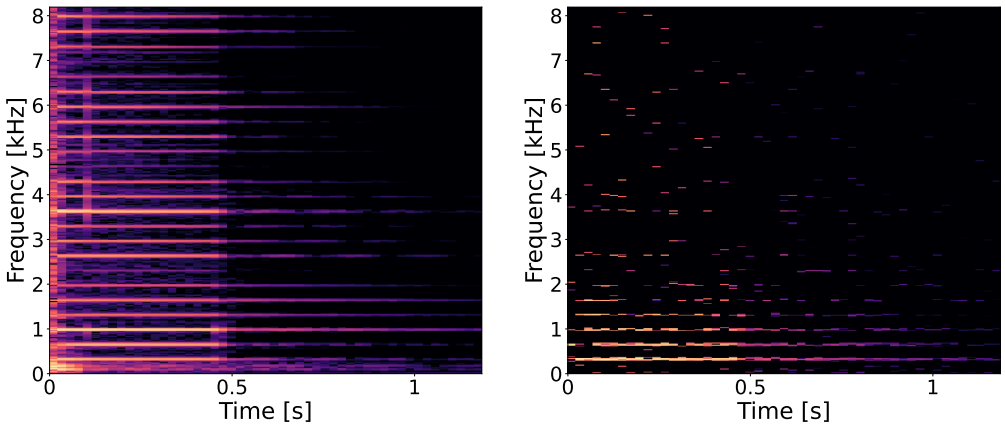


Fig. A.8. The power spectrograms of (a) the input guitar E4 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

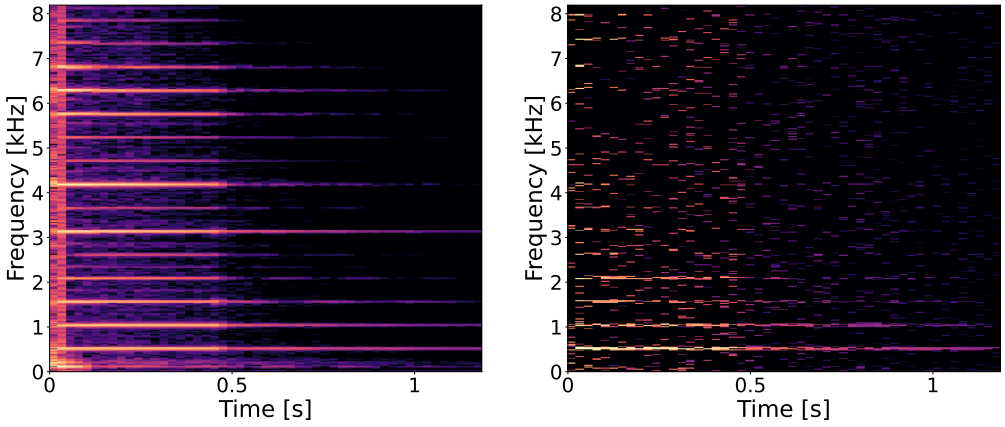


Fig. A.9. The power spectrograms of (a) the input guitar C5 note signal and (b) its predicted signal obtained by the trained MLP with the MSE loss function.

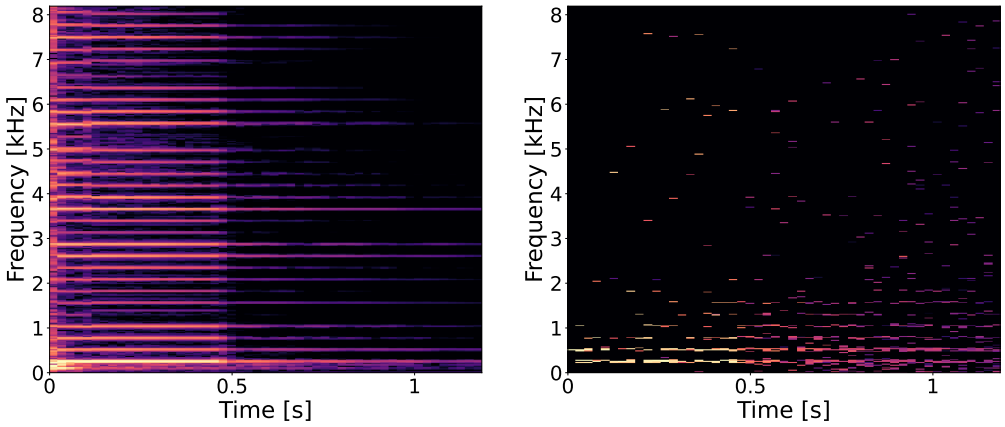
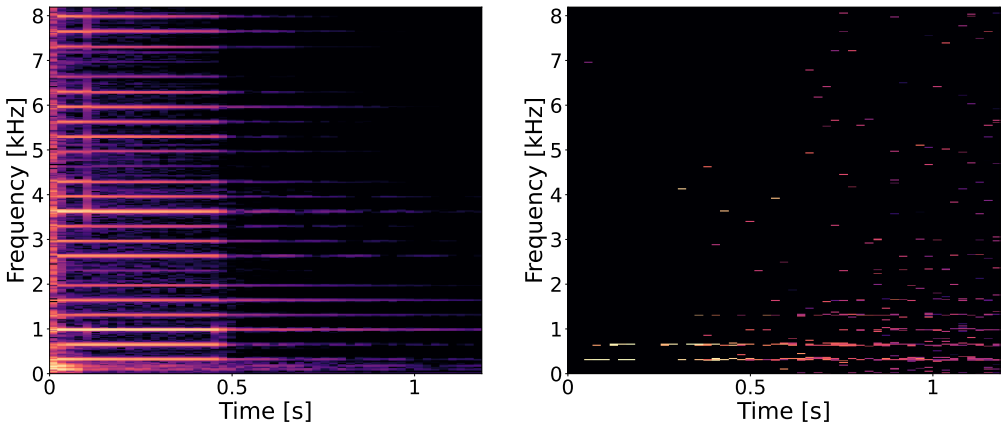
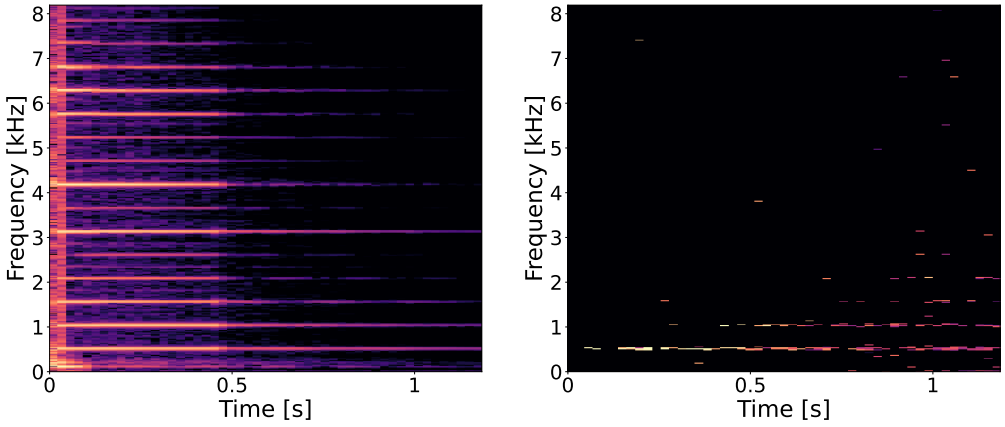


Fig. A.10. The power spectrograms of (a) the input guitar C4 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.



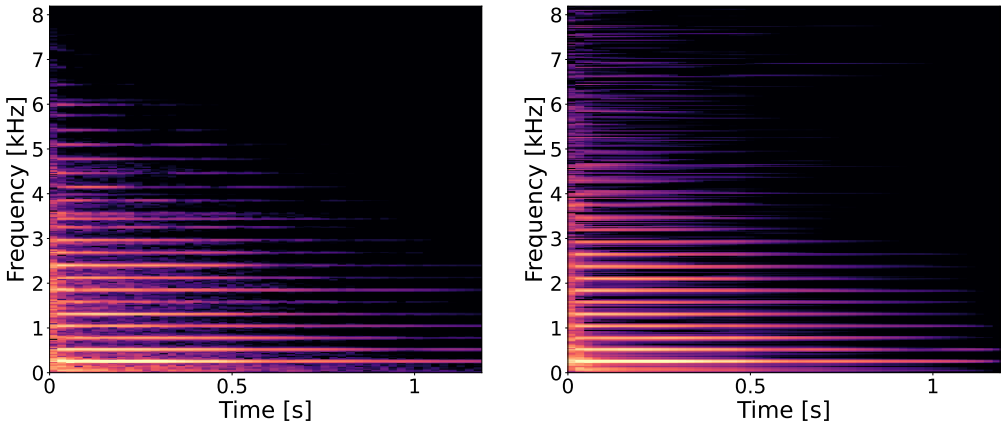
(a)　(b)

Fig. A.11. The power spectrograms of (a) the input guitar E4 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.



(a)　(b)

Fig. A.12. The power spectrograms of (a) the input guitar C5 note signal and (b) its predicted signal obtained by the trained MLP with the MSS loss function.



(a)　(b)

Fig. A.13. The power spectrograms of (a) the input piano C4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

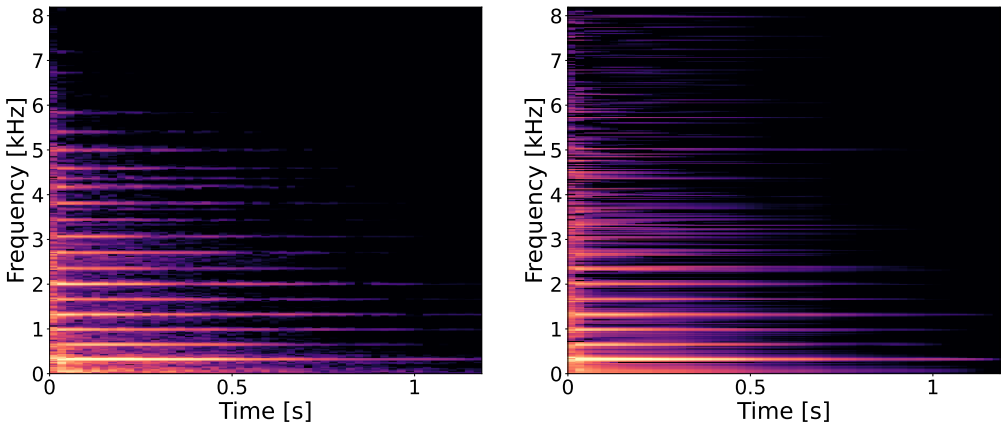


Fig. A.14. The power spectrograms of (a) the input piano E4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

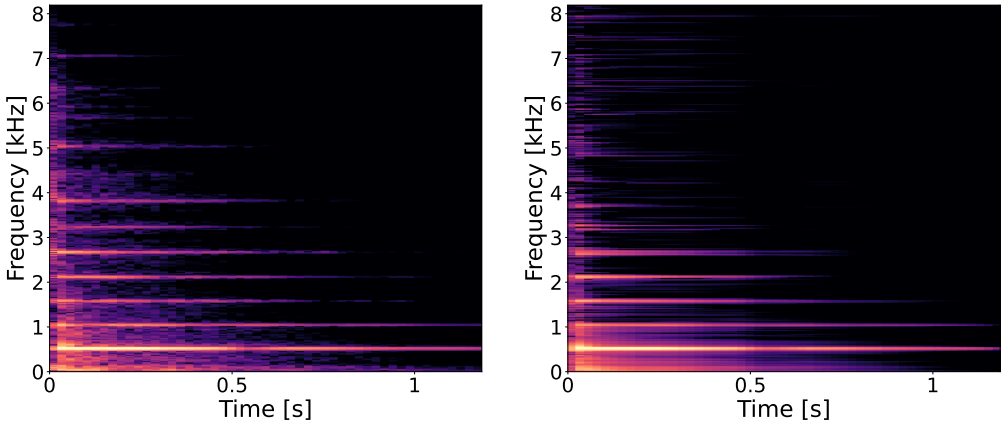


Fig. A.15. The power spectrograms of (a) the input piano C5 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

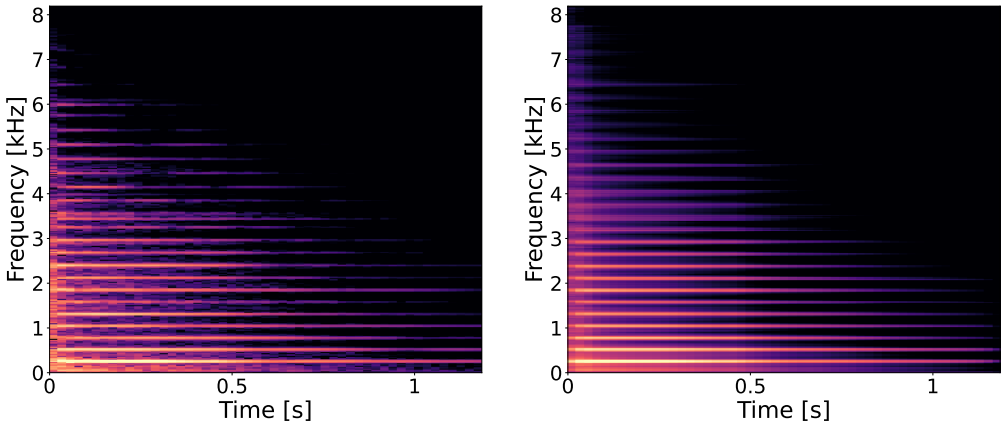


Fig. A.16. The power spectrograms of (a) the input piano C4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.

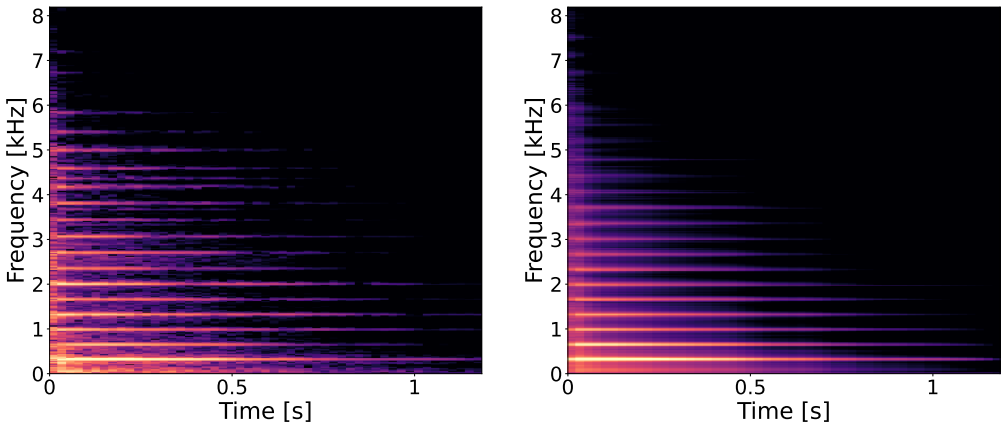


Fig. A.17. The power spectrograms of (a) the input piano E4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.

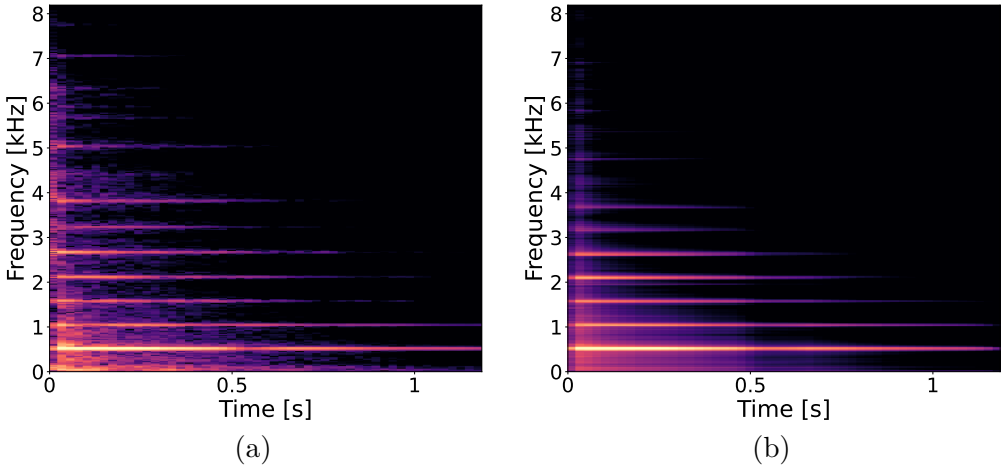


Fig. A.18. The power spectrograms of (a) the input piano C5 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.

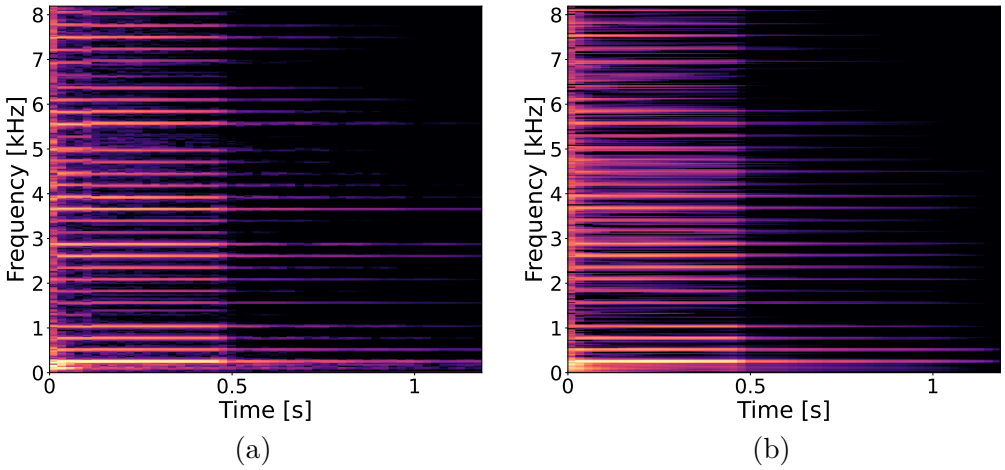


Fig. A.19. The power spectrograms of (a) the input guitar C4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

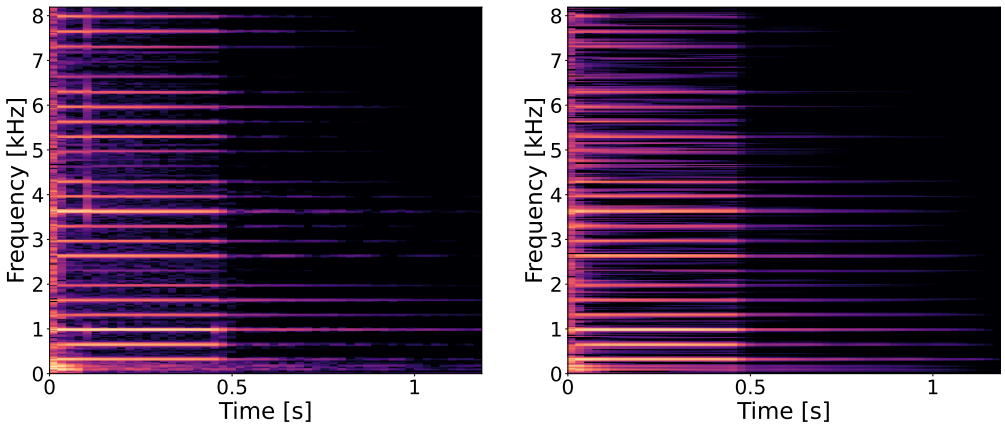


Fig. A.20. The power spectrograms of (a) the input guitar E4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

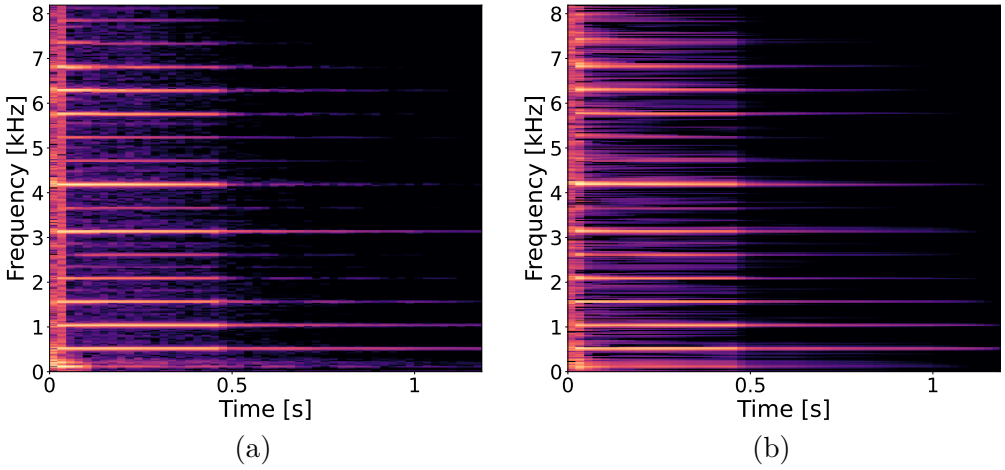


Fig. A.21. The power spectrograms of (a) the input guitar C5 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSE loss function.

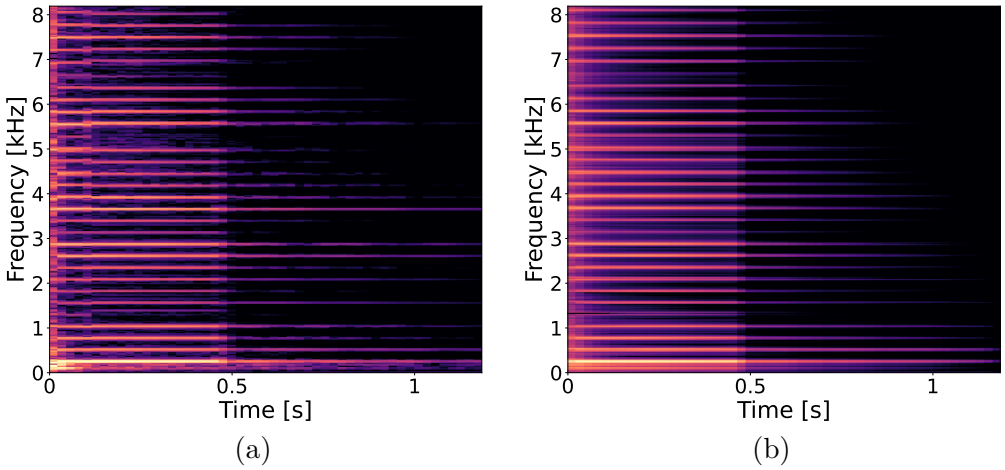
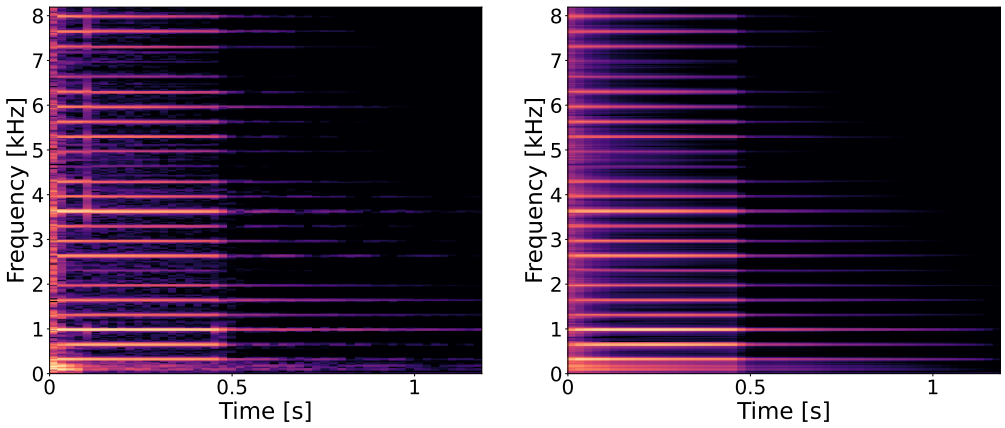
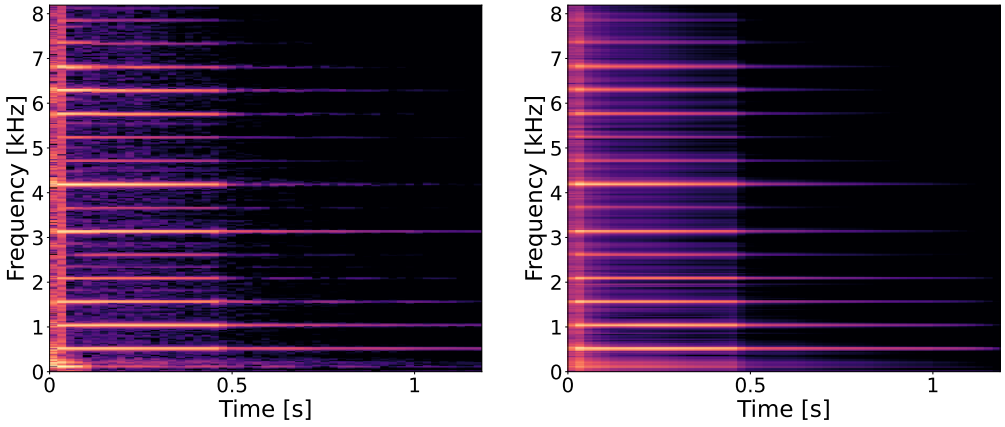


Fig. A.22. The power spectrograms of (a) the input guitar C4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.



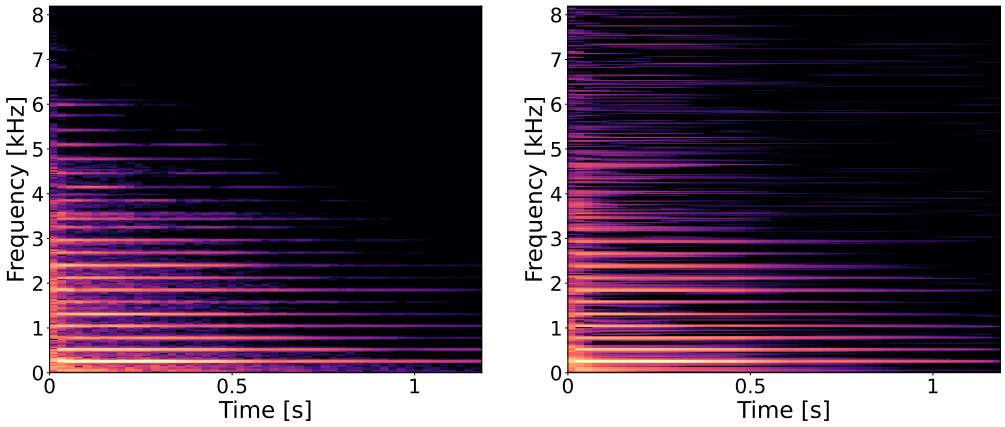
(a)　(b)

Fig. A.23. The power spectrograms of (a) the input guitar E4 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.



(a)　(b)

Fig. A.24. The power spectrograms of (a) the input guitar C5 note signal and (b) its predicted signal obtained by the trained BiLSTM with the MSS loss function.



(a)　(b)

Fig. A.25. The power spectrograms of (a) the input piano C4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

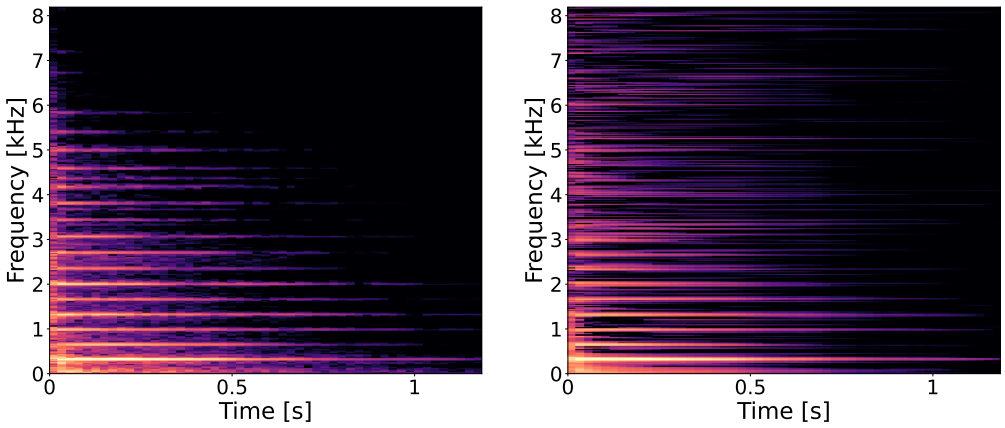


Fig. A.26. The power spectrograms of (a) the input piano E4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

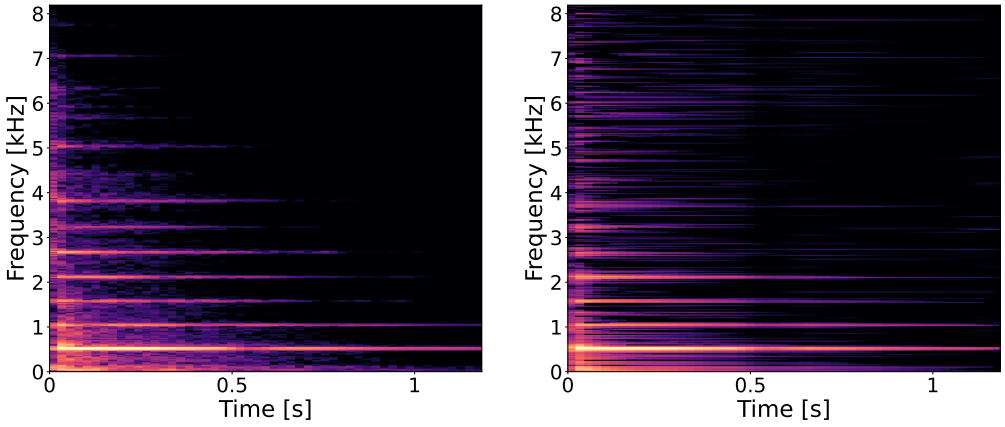


Fig. A.27. The power spectrograms of (a) the input piano C5 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

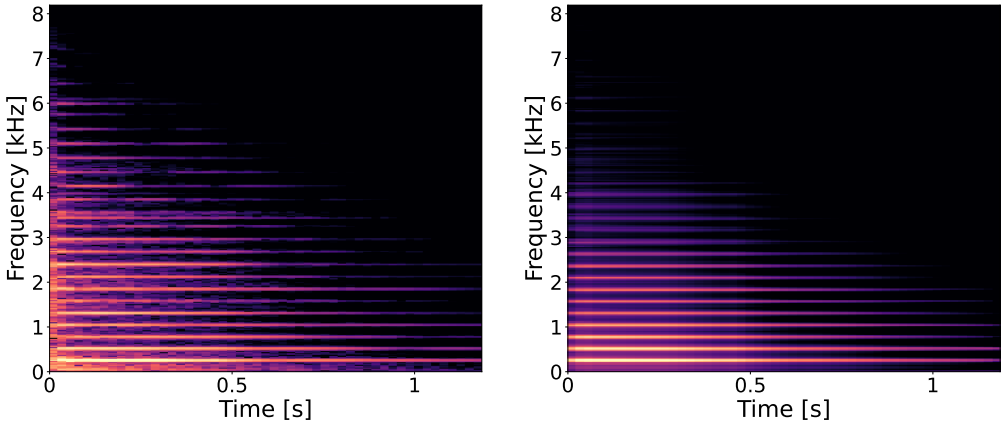
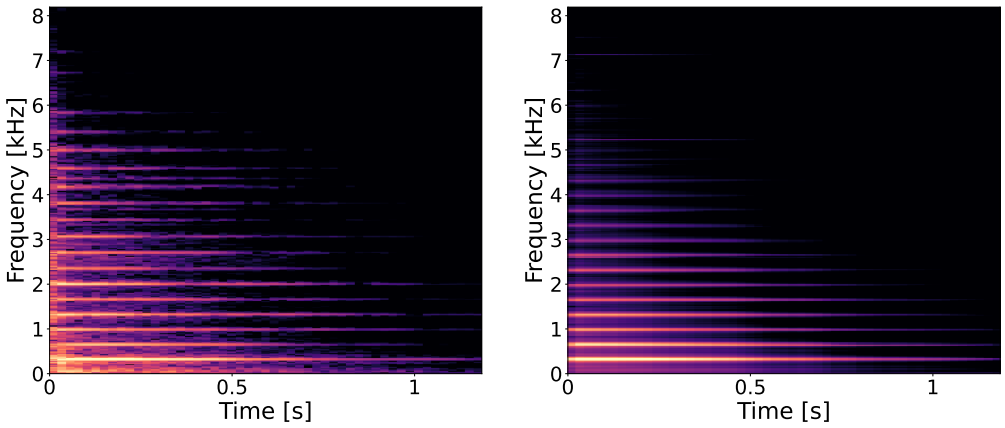
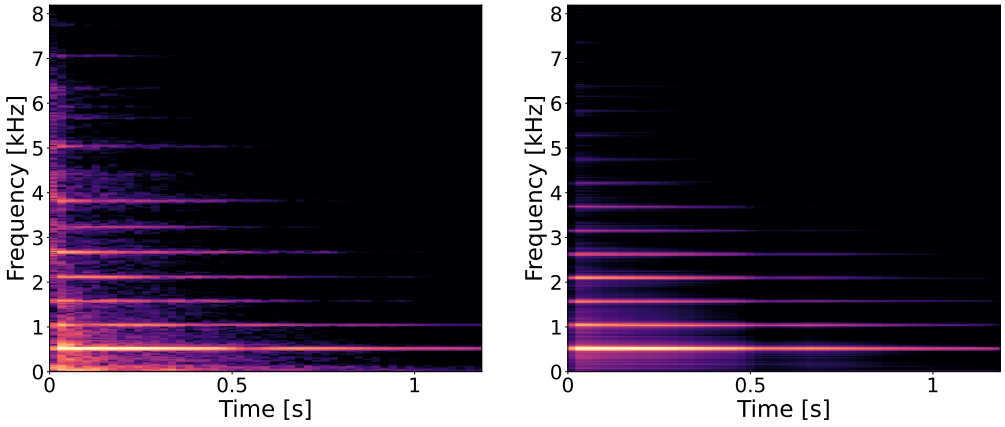


Fig. A.28. The power spectrograms of (a) the input piano C4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.



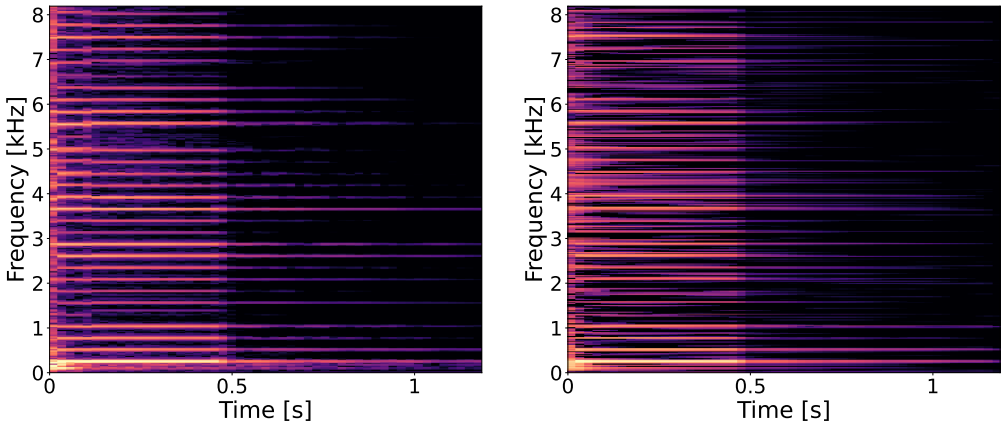
(a)　(b)

Fig. A.29. The power spectrograms of (a) the input piano E4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.



(a)　(b)

Fig. A.30. The power spectrograms of (a) the input piano C5 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.



(a)　(b)

Fig. A.31. The power spectrograms of (a) the input guitar C4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

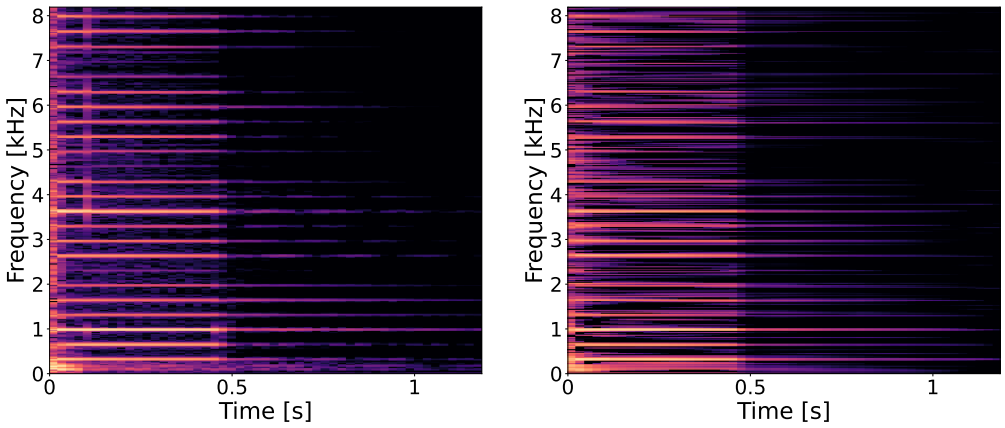


Fig. A.32. The power spectrograms of (a) the input guitar E4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

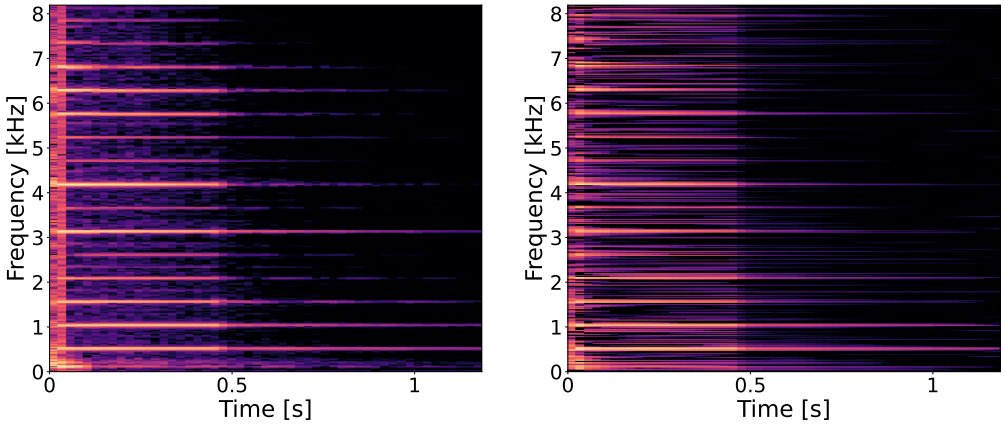


Fig. A.33. The power spectrograms of (a) the input guitar C5 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSE loss function.

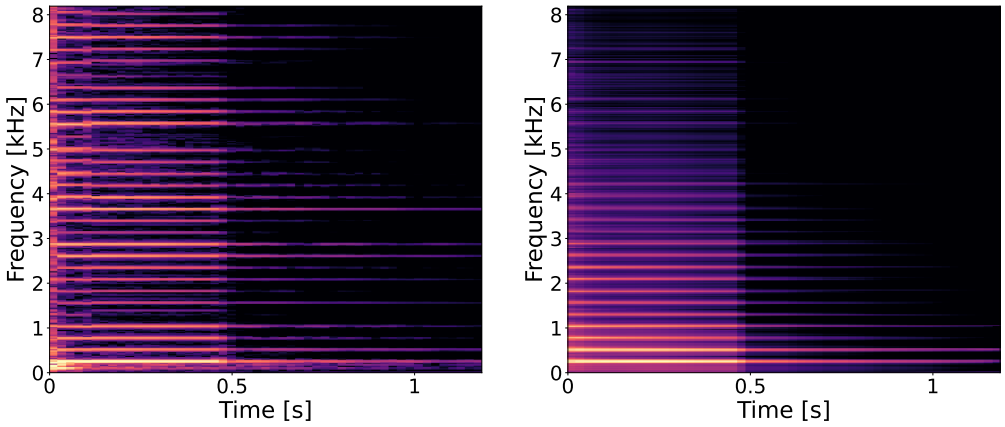


Fig. A.34. The power spectrograms of (a) the input guitar C4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.

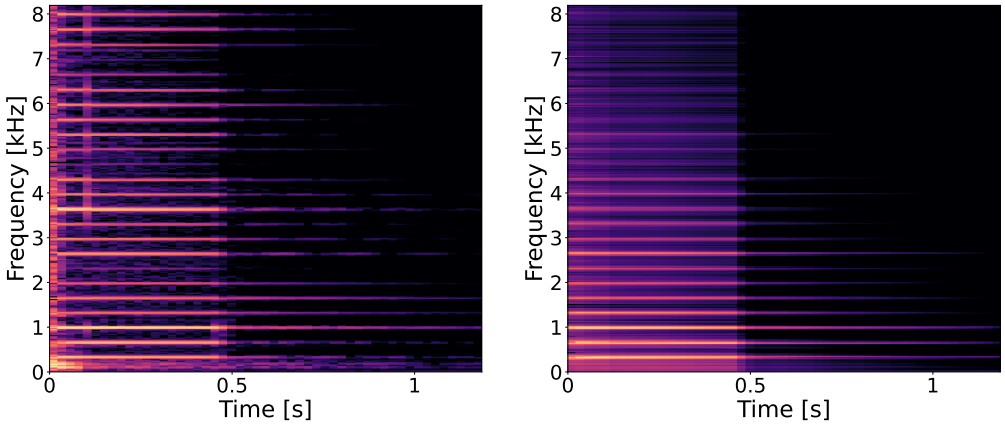


Fig. A.35. The power spectrograms of (a) the input guitar E4 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.

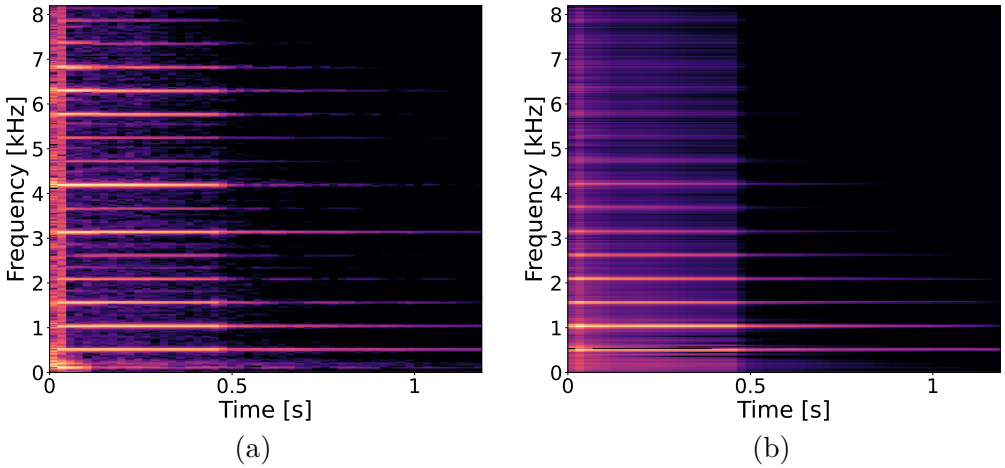


Fig. A.36. The power spectrograms of (a) the input guitar C5 note signal and (b) its predicted signal obtained by the trained BiGRU with the MSS loss function.