修 士 論 文

機械学習を用いた音楽のスタイル変換

広島大学大学院先進理工系科学研究科

瀬野僚太郎

令和６年３月

修 士 論 文

機械学習を用いた音楽のスタイル変換

指導教員　伊藤靖朗 教授

広島大学大学院
先進理工系科学研究科情報科学プログラム

M225760 瀬野僚太郎

提出年月: 令和 6 年 2 月

概 要

本研究では，機械学習を用いてゲーム音楽をクラシック音楽に変換する音楽のスタイル変換を行う．本研究で使用するゲーム音楽は，ファミリーコンピュータに収録されたゲーム BGM である．音楽のスタイル変換は DTM などの専用の機械を使用することで自分好みに音楽を編集することができるが，音楽や DTM に関する専門知識が必要である．また，DTM による音楽のスタイル変換は各楽器，音階を手打ちで入力するため，非常に時間がかかる．そこで，専門知識が無くても自動で音楽のスタイル変換を行う方法として，機械学習を使用した音楽のスタイル変換手法を提案する．

本研究では，ある話者の音声を別の話者の音声に変換する声質変換 (VC:VoiceConversion) をベースとした音楽のスタイル変換手法を提案する．声質変換は，大きく 2 つの部分に分けることができる．1 つ目は音響特徴量のスタイル変換を行うスタイル変換器，2 つ目は音響特徴量から音声を生成する Neural Vocoder である．本研究では，スタイル変換器に MaskCycleGAN-VC，Neural Vocoder に ParallelWaveGAN を使用する声質変換手法をベースとした音楽のスタイル変換器を提案する．MaskCycleGAN-VC と ParallelWaveGAN は，人間のスピーチ音声を分析するように設計されている．しかし，人間のスピーチ音声と本研究で扱うクラシック音楽，ゲーム音楽と比較すると，特徴が大きく異なる．例えば，人間のスピーチ音声に比べてクラシック音楽やゲーム音楽は高周波成分が豊富であり，ロングトーン (長く連続する音) も多く含まれている．そのため，ベースモデルをそのまま使用した場合，高周波成分がうまく再現できず，ノイズが生成される．また，ロングトーンを生成しようとしたときには不安定な音声が生成されてしまう．そこで，本研究ではクラシック音楽，ゲーム音楽を分析するための改良モデルを提案する．

本研究では，ベースモデルに高周波成分と音階情報を分析するようにモデルを改良する．提案モデルは，高周波成分や音階情報を学習するモデルに改良することで音楽スタイル変換の生成音声の品質がベースモデルから向上した．また，アンケート調査の結果から提案手法が既存手法を上回り，スタイル変換の品質を向上させることが分かった．

目 次

図 目 次

第1章　はじめに

1.1　研究の背景

機械学習を用いた音声の生成に関する研究では，文章から音声を生成する音声合成 (TTS:Text-to-Speech)，音声の話者を変換する声質変換 (VC:VoiceConversion) などが盛んに行われている．本研究では，声質変換を応用した音楽のスタイル変換を行う．声質変換は，大きく 2 つの部分に分けることができる．1 つ目は，音響特徴量を変換するスタイル変換器である．スタイル変換器は，変換前の音声をメルスペクトログラムに変換し，メルスペクトログラムのスタイル変換を行う．2 つ目は，音声を生成する音声生成器である．音声生成器は，メルスペクトログラムから音声を生成する．機械学習を用いた声質変換では，この 2 つのモデルを組み合わせて声質変換を行う．

一般に，人間の発話音声の変換を行う場合は，変換前の音声と変換後の音声でメルスペクトログラムの周波数構造が似ているため，スタイル変換器で変換前後の特徴を学習しやすい．例えば，男性の「こんにちわ」という音声と女性の「こんにちわ」という音声は周波数構造が似ているといえる．人間の発話音声を変換する場合は，CycleGAN-VC2[1] や MaskCycleGAN-VC[2] などの既存のスタイル変換器と WaveNet[3] などの既存の音声生成器を組み合わせることで高品質に声質変換が可能である．しかし，変換の前後で特徴が大きく異なる音声の場合，既存の手法では依然として課題がある．例えば，POP 音楽からクラシック音楽など，ジャンルの異なる音楽の場合は変換前後の依存関係を学習することが難しい．

本研究では，ゲーム音楽をクラシック音楽に変換する音楽のスタイル変換を行う．ゲーム音楽とクラシック音楽は使用されている楽器などの音色が異なるため，変換前後の特徴を学習することが難しい．また，多くの既存研究で対象としている人の発話音声に比べて，ゲーム音楽やクラシック音楽は連続した長い音や高周波成分を豊富に含むため，より表現力が高い音声生成器を設計する必要がある．

1.2　研究の目的

本研究では，機械学習を用いた音楽のスタイル変換を行う．本研究の概要を図 1.1 に示す．メルスペクトログラムのスタイル変換を行う MaskCycleGAN-VC[2] とメルスペクトログラムから音声を生成する ParallelWaveGAN[4] を組み合わせた声質変換手法をベースに，音楽のスタイル変換手法を提案する．本研究ではベースモデルに高周波成分を学習す

るための識別機や音階情報を学習するための損失関数を使用し，クラシック音楽，ゲーム音楽に特化した音楽スタイル変換手法を提案する．

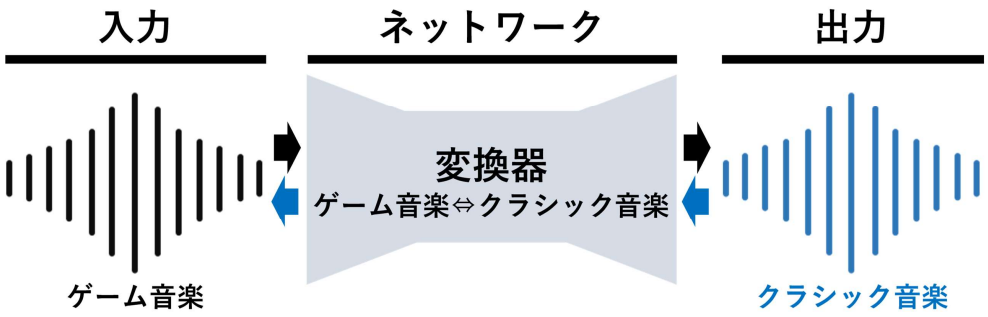


図 1.1: 本研究で行う音楽スタイル変換の概要

1.3 本論文の概要

本論文は，次のように構成されている．2 章では，関連研究を紹介する．3 章では，提案手法について述べる．4 章では，実験の詳細やデータセットについて触れた後，様々な実験と結果を示す．最後に，5 章で本論文のまとめを行う．

第2章　関連研究

2.1　声質変換

声質変換 (VC:Voice Conversion) は，発話音声の話者を変換することをいい，近年，機械学習を用いた研究が盛んに行われている．一般に，機械学習を用いた声質変換は大きく 2 つの部分に分けることができる．1 つ目は音響特徴量のスタイル変換を行うスタイル変換器，2 つ目は音響特徴量から音声を生成する Vocoder である．機械学習をベースとした Vocoder を特に Neural Vocoder という．声質変換の全体図を図 2.1 に示す．本章では，声質変換を構成する音響特徴量スタイル変換器，Neural Vocoder とそれらに関連する重要な既存研究について触れていく．

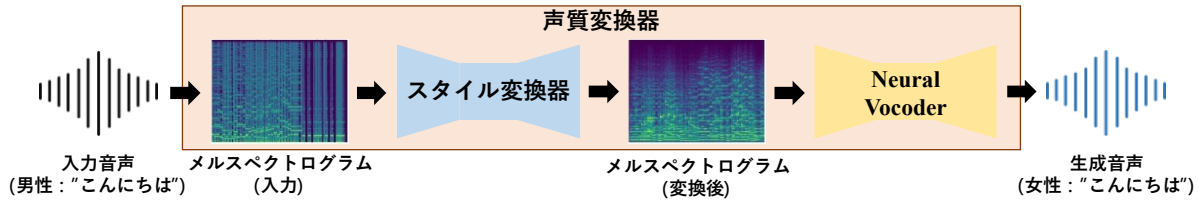


図 2.1: 声質変換の全体図

2.2　音響特徴量のスタイル変換器

機械学習を用いて画像や音声などを生成する分野において，GAN[5] をベースとしたモデルが頻繁に使用されている．GAN は 2014 年に発表された画像生成モデルである．GAN のモデル構造を図 2.2 に示す．GAN は，生成器 (Generator) と識別機 (Discriminator) の二つのモデルで構成されている．生成器は，ノイズを入力して画像を出力する．識別器は，データセットの画像と生成器の生成画像を入力し，0 から 1 の値を一つ出力する．識別器の出力は入力画像が本物か偽物か識別した値を表しており，0 は偽物，1 は本物を表している．識別器は入力画像を正しく識別できるように学習を行なう．生成器は，生成画像を識別器の出力が本物になるように学習を行なう．このように，生成器と識別器は敵対的に学習させることで，生成器は本物に近い画像を生成することができる．

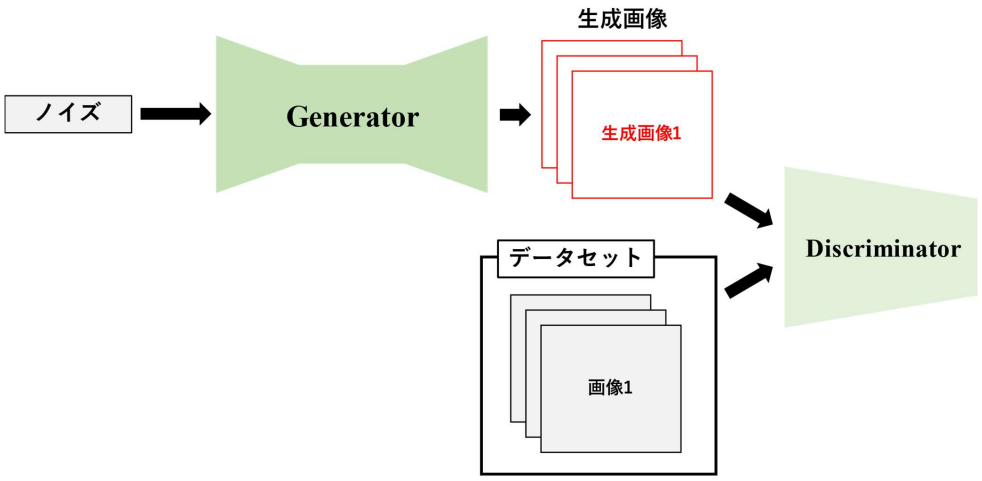


図 2.2: GAN のモデル構造

音響特徴量のスタイル変換を行なうスタイル変換器は，画像変換モデルをベースに設計されている．特に，教師なし画像変換モデルである CycleGAN[6] をベースにメルスペクトログラムや CQT スペクトログラムなどの音響特徴量をスタイル変換するモデルが数多く提案されている．

CycleGAN は，GAN をベースに設計されたモデル構造となっており，2 つの生成器 $G_{X\to Y}$，$G_{Y\to X}$ と，2 つの識別機 D_X，D_Y から構成される．2 つの生成器は，$G_{X\to Y}$ がドメイン X の画像を入力してドメイン Y の画像を生成し，$G_{Y\to X}$ がドメイン Y の画像を入力してドメイン X の画像を生成する．2 つの識別機は，D_X にドメイン X のデータセットの正解画像と $G_{Y\to X}$ が生成した画像を入力し，D_Y にドメイン Y のデータセットの正解画像と $G_{X\to Y}$ が生成した画像を入力してそれぞれ本物か偽物か識別する．また，CycleGAN では，それぞれの生成器が元の画像の特徴を維持しながら画像変換を行なうために Cycle Consistency Loss を使って学習する．Cycle Consistency Loss は，ドメイン X の画像を $G_{X\to Y}$ に入力して生成された画像をさらに $G_{Y\to X}$ で再構成した画像が元の画像と一致するように学習するための損失関数である．このように，CycleGAN はペアのデータセットなしで画像のスタイル変換が可能なモデルである．

声質変換で音響特徴量のスタイル変換を行なうスタイル変換器では，音声を周波数分析して得られるメルスペクトログラムや CQT スペクトログラムをスタイル変換する．メルスペクトログラムや CQT スペクトログラムは音声データから得られる特徴量であり，スタイル変換前とスタイル変換後のペアのデータセットを用意するのは難しい場合が多い．例えば，話者 A と話者 B が”こんにちは”と発話する音声データは，発話内容が同じであっても発音の速度などが異なるため，同じ発音でも時間がずれているものが多い．そのため，音響特徴量のスタイル変換を行なう既存の手法は，ペアのデータセットなしで画像から画像へのスタイル変換が可能なモデルである CycleGAN[6] をベースとしたモデルが提案されている．

MaskCycleGAN-VC[2] は CycleGAN をベースとした声質変換を行なうためのスタイル

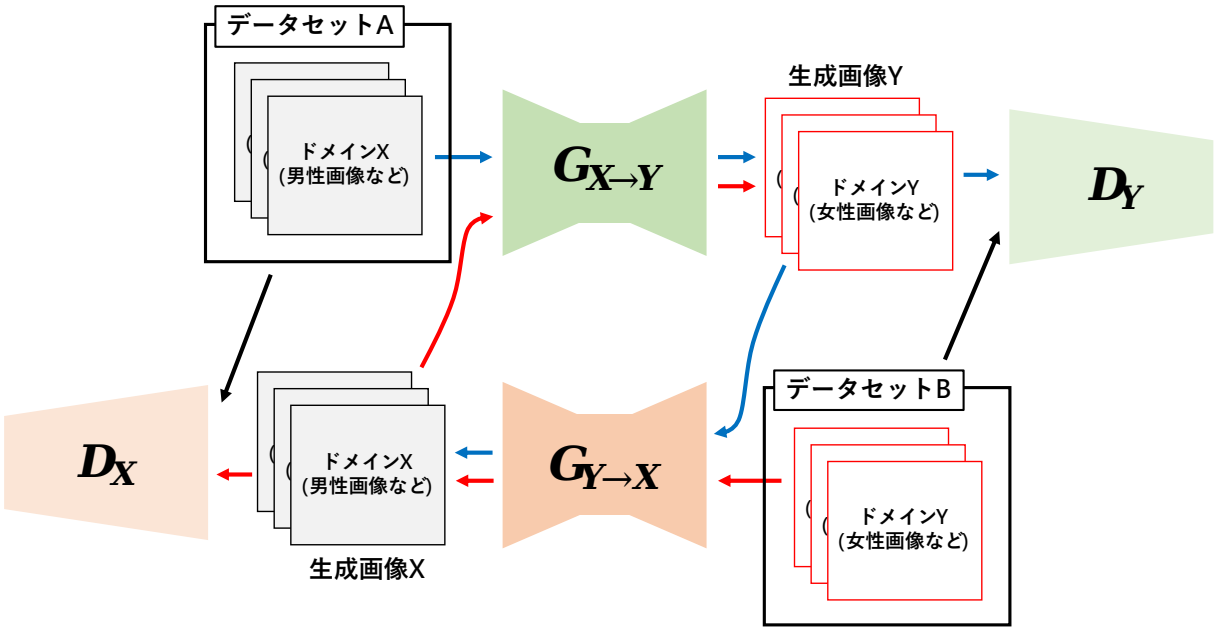


図 2.3: CycleGAN のモデル構造

変換器であり，CycleGAN-VC2[1] を発展させたモデルである．この節では，順を追って CycleGAN-VC2，MaskCycleGAN-VC について述べる．

2.2.1 CycleGAN-VC2

CycleGAN-VC2[1] は CycleGAN をベースとした音響特徴量のスタイル変換モデルで，メルケプストラムという音響特徴量のスタイル変換を行う．CycleGAN-VC2 のモデルの構造を図 2.4 に示す．CycleGAN-VC2 は元の CycleGAN に加えて識別機が 2 つ増えており，生成器が 2 つと識別機が 4 つで構成される．CycleGAN-VC2 の 2 つの生成器はメルケプストラムを入力し，メルケプストラムを生成する．また，4 つの識別機は，メルケプストラムを入力し，データセット音声から作成した正解メルケプストラムか生成器が生成した偽物メルケプストラムかを識別する．メルケプストラムは音声信号を周波数分析して得られる音響特徴量であり，画像と同じ二次元のデータである．そのため，元となっている CycleGAN のモデルが画像を分析する構造と同じように CycleGAN-VC2 はメルケプストラムを分析することができる．CycleGAN-VC2 で新しく追加された 2 つの識別機は，正解メルケプストラムと 2 つの生成器で再構成したメルケプストラムの識別を行なう．

D'_X はドメイン X の正解メルケプストラムと，$G_{X\to Y}$ と $G_{Y\to X}$ で再構成したメルケプストラムの識別を行ない，D'_Y はドメイン Y の正解メルケプストラムと，$G_{Y\to X}$ と $G_{X\to Y}$ で再構成したメルケプストラムの識別を行なう．4 つの識別機 D_X，D_Y，D'_X，D'_Y は入力メルケプストラムが本物か偽物か正しく識別できるように学習し，2 つの生成器 $G_{X\to Y}$，$G_{Y\to X}$ は 4 つの識別機をだますように学習する．このように，CycleGAN-VC2 はドメイ

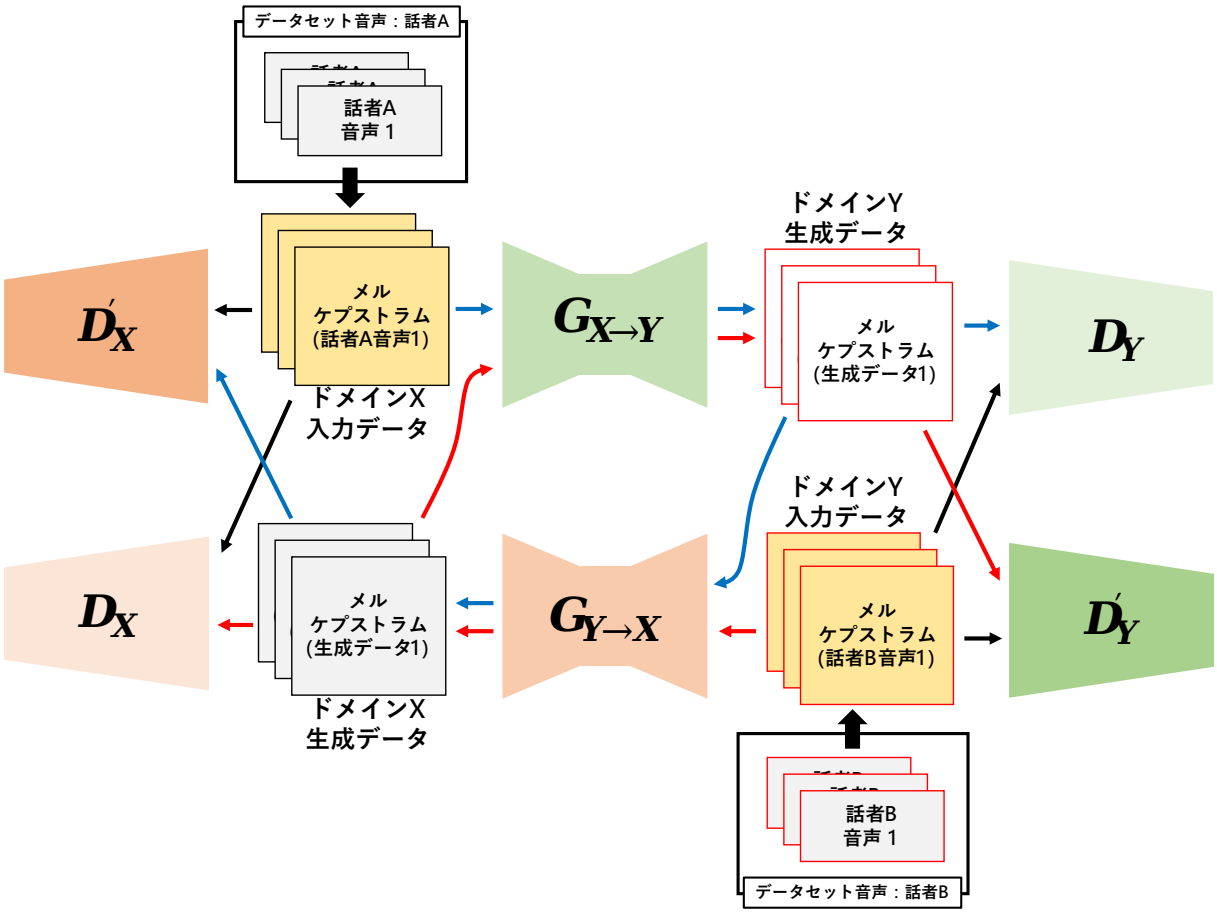


図 2.4: CycleGAN-VC2 のモデル構造

ン変換を重点的に学習する構造によってメルケプストラムのスタイル変換を実現し，メルケプストラムから音声を生成する Neural Vocoder と組み合わせることで声質変換が可能となった．

しかし，CycleGAN-VC2 は既存の Neural Vocoder との相性が悪いという欠点がある．既存の Neural Vocoder は，メルスペクトログラムを音声に変換するように設計されているものが多い．CycleGAN-VC2 はメルケプストラムをスタイル変換することは可能だが，メルスペクトログラムの周波数構造を学習できないため，既存のメルスペクトログラムの Neural Vocoder の発展にもかかわらず，メルケプストラムから音声を生成する Vocoder に限定されるという欠点がある．

2.2.2 MaskCycleGAN-VC

MaskCycleGAN-VC[2] は CycleGAN-VC2 をベースとしたスタイル変換モデルで，メルスペクトログラムのスタイル変換を行う．MaskCycleGAN-VC のモデルの構造を図 2.5 に示す．

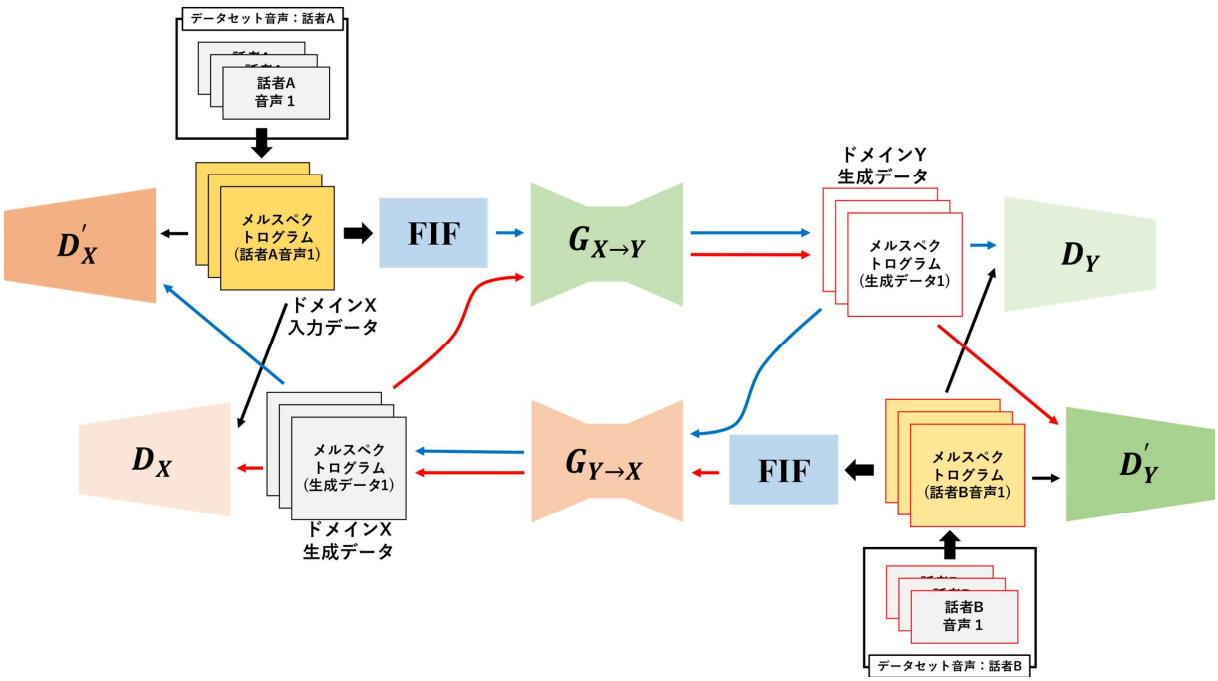


図 2.5: MaskCycleGAN-VC のモデル構造

モデルの構造は CycleGAN-VC2 とほとんど変わらないが，MaskCycleGAN-VC では，CycleGAN-VC2 でメルスペクトログラムの変換が苦手であるという欠点を改善するため FIF(filling in frames) という補助タスクを使用する．FIF の概念図を図 2.6 に示す．FIF では，マスクとメルスペクトログラムを使用する．メルスペクトログラムは音声信号を周波数分析して得られる音響特徴量であり，縦軸が周波数，横軸が時間の 2 次元のデータである．マスクは，縦と横の大きさがメルスペクトログラムと同じデータであり，マスクの各要素は 0 か 1 である．FIF では，メルスペクトログラムとマスクの要素毎の積を計算してマスクの一部を欠落させる．生成器には欠落させたメルスペクトログラムとマスクの 2 つのデータを入力する．MaskCycleGAN-VC の生成器は，マスクの情報から欠落しているメルスペクトログラムの位置を特定し，周囲の周波数構造を参考に欠落した部分の周波数構造を修復するように学習する．CycleGAN-VC2 では，メルスペクトログラムの周波数構造を学習することができなかったが，MaskCycleGAN-VC は周波数構造を学習する FIF によってメルスペクトログラムの生成能力が向上し，メルスペクトログラムを使用する既存の Neural Vocoder と組み合わせることが可能になった．

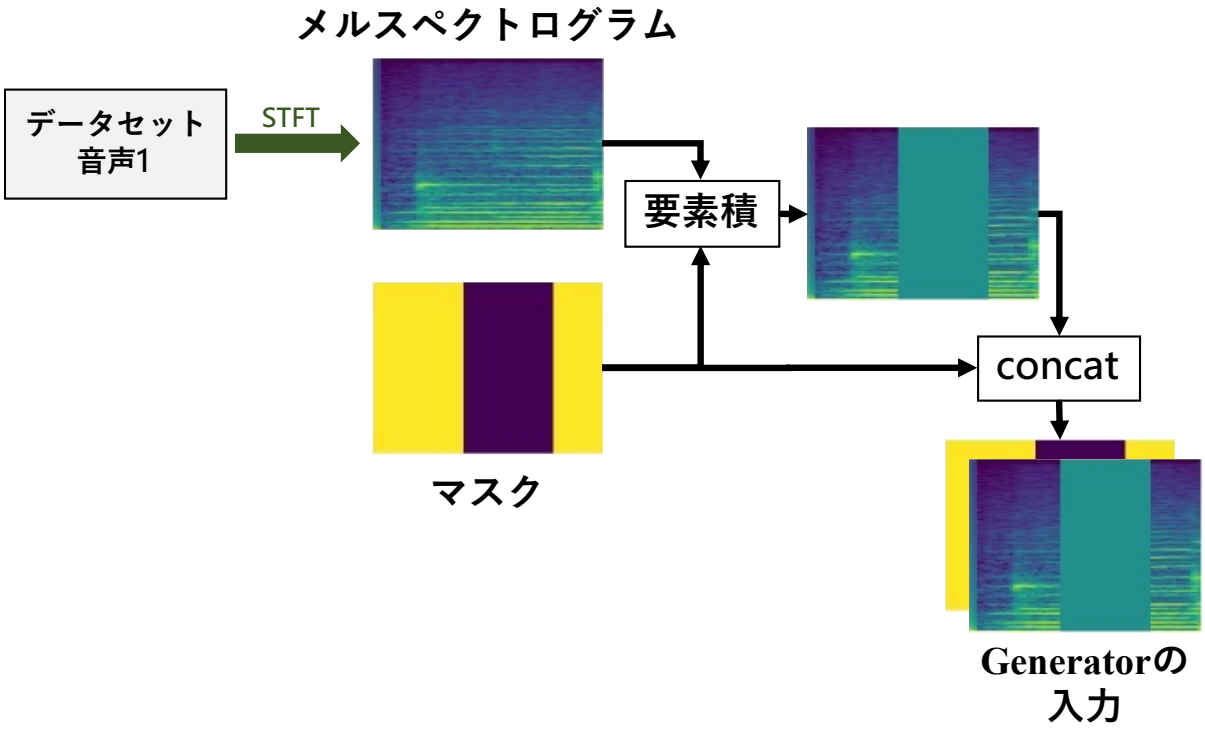


図 2.6: FIF の概念図

2.3 Neural Vocoder

メルスペクトログラムなどの音響特徴量を入力して音声を生成するモデルを Vocoder という．特に，機械学習を用いた Vocoder を Neural Vocoder という．WaveNet[3] は 2016 年に発表された NeuralVocoder であり，機械学習を用いた音声の生成において始めて高品質な音声が生成可能となったモデルである．ParallelWaveGAN[4] は，2020 年に発表された Neural Vocoder であり，音声合成などの分野で頻繁に使用されている．また，生成タスクによって，同時期に発表された MelGAN[7] や HiFiGAN[8] などと比較して生成音声が高品質な場合が多い．SingGAN[9] は，既存の Neural Vocoder をベースに歌声を生成するように改良された Neural Vocoder である．この節では，順を追って WaveNet，ParallelWaveGAN，SingGAN について述べる．

2.3.1 WaveNet

WaveNet は，2016 年に発表された音声生成モデルであり，機械学習を用いた音声の生成で初めて高品質な音声を生成することが可能になった手法である．音声信号は 1 次元の信号であり，1 秒毎のサンプル数は 16000 から 44100 サンプルになるものもあるため，短

い音声でもデータが多いという特徴がある．また，音声信号は時間とともに波形が変化する時系列信号である．従来の方法では，数理モデルを用いた音声生成手法が一般的であったが，自然な音声に含まれるゆらぎというランダム性が再現できず，生成音声の品質が悪いという欠点があった．機械学習を用いる場合，時系列データを扱うモデルとして RNN があるが，サンプル数が非常に多い音声信号を RNN で学習，推論するのは非常に時間がかかるため，RNN は音声生成タスクに適していない．WaveNet では，RNN を使用せずに，音声を自己回帰に生成する過程を高速に計算するためのモデル構造として，１次元畳み込みを応用した Dilated Causal Convolution を提案した．Dilated Causal Convolution の概念図を図 2.7 に示す．

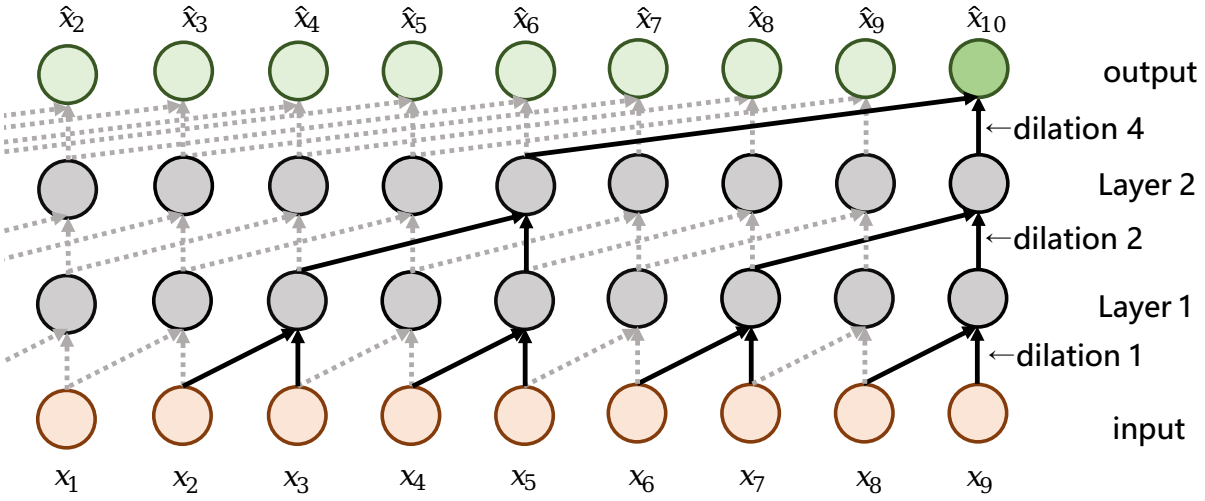


図 2.7: Dilated Causal Convolution

Dilated Causal Convolution では，音声信号の過去のサンプルをもとに次の 1 サンプルを生成する構造を畳み込みによって実装している．また，dilation という畳み込みのサンプルの間隔をあけるパラメータを調整することで受容野を広くする工夫がなされている．受容野とは，出力の 1 サンプルを生成をするために畳み込み演算される入力のサンプル数のことである．例えば，図 2.7 の場合，出力サンプルの $\hat{x}_{10}$ は，入力信号の x_2 から x_9 の８サンプルを畳み込んで生成されているため，受容野は 8 である．dilation を適用しない畳み込みの場合は層の数に比例して受容野が広くなるが，dilation を適用する場合は層の数に応じて指数的に広くなり，効率的に受容野を広くすることができる．

WaveNet のモデル全体構造を図 2.8 に示す．WaveNet は，Dilated Causal Convolution を含む ResBrock を深く積み重ねた構造となっている．WaveNet は深い畳み込みのネットワークのため，並列計算が得意な GPU などを使用することで効率的に学習が可能になり，従来の音声生成手法と比較して生成音声の品質が大幅に向上し，発話音声の生成では人間に近い音声が生成可能になった．一方で，音声の生成時には１サンプルずつ音声を生成する必要があり，生成に非常に時間がかかるという欠点がある．2017 年には，WaveNet の生成速度を改善したモデルとして ParallelWaveNet[10] が発表され，リアルタイムでの音

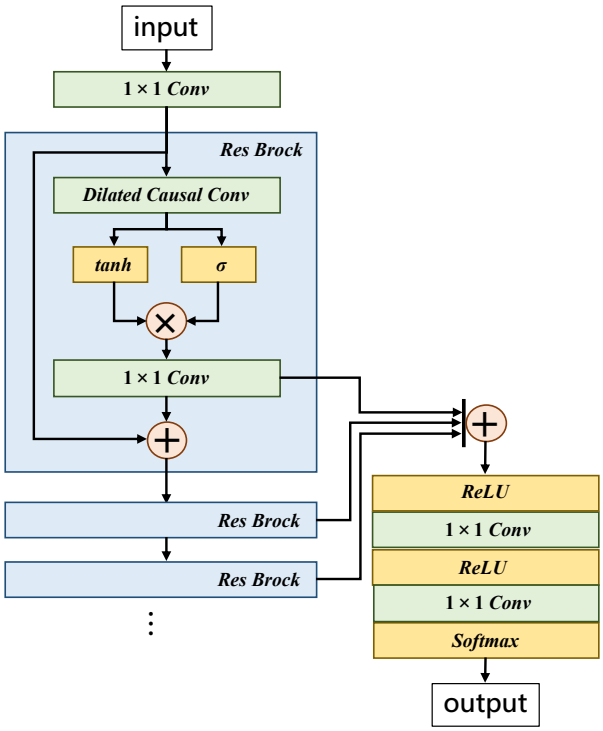


図 2.8: WaveNet のモデル全体構造

声生成が可能になった．しかし，学習済みの WaveNet モデルを使った 2 段階の学習が必要なため，学習に 2 倍以上時間がかかるという欠点がある．

2.3.2 ParallelWaveGAN

ParallelWaveGAN[4] はメルスペクトログラムを入力して音声波形を生成する Neural Vocoder である．ParallelWaveGAN のモデル構造を図 2.9 に示す．ParallelWaveGAN は GAN をベースとしたモデルの構造となっており，生成器と識別機から構成される．ParallelWaveGAN の生成器は，WaveNet のモデル構造をベースにしたモデル構造となっている．WaveNet は音声サンプルを入力して音声サンプルを一つずつ生成する構造になっているのに対し，ParallelWaveGAN はメルスペクトログラムとホワイトノイズを入力して音声波形を出力する．入力ホワイトノイズは出力音声と同じ長さの 1 次元信号であり，生成器は入力メルスペクトログラムの情報を分析し，ホワイトノイズの信号をメルスペクトログラムの特徴を反映した音声信号に変換させるタスクを学習する．WaveNet は音声サンプルを 1 つずつ生成する構造のため，生成に非常に時間がかかる．ParallelWaveGAN の生成器は音声とホワイトノイズを入力して並列計算によって音声波形を出力するため，WaveNet と比較して高速に生成できる利点がある．ParallelWaveGAN の識別機はデータセットの本物音声と生成器の生成音声を入力して本物か偽物か判別する．このように GAN をベースとした構造となっており，生成器は識別機をだませるような本物に近い音声を生

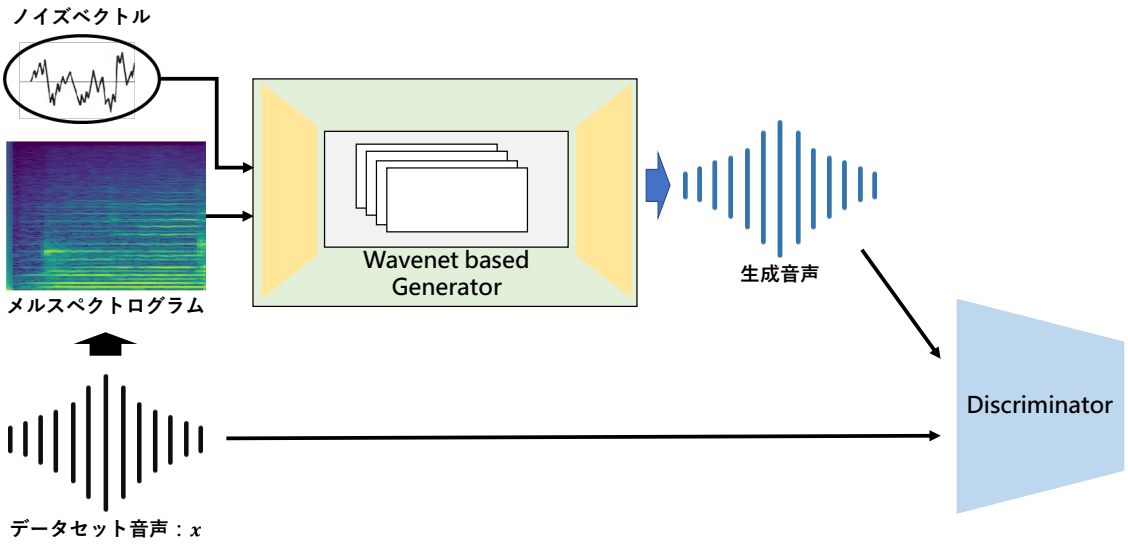


図 2.9: ParallelWaveGAN のモデル構造

成できるように学習し，識別機は本物と偽物を識別できるように学習する敵対的な構造となっている．ParallelWaveGAN は，生成器の生成音声の品質を向上するために，GAN の学習に加えて STFT Loss を使って学習を行う．STFT Loss の概念図を図 2.10 に示す．

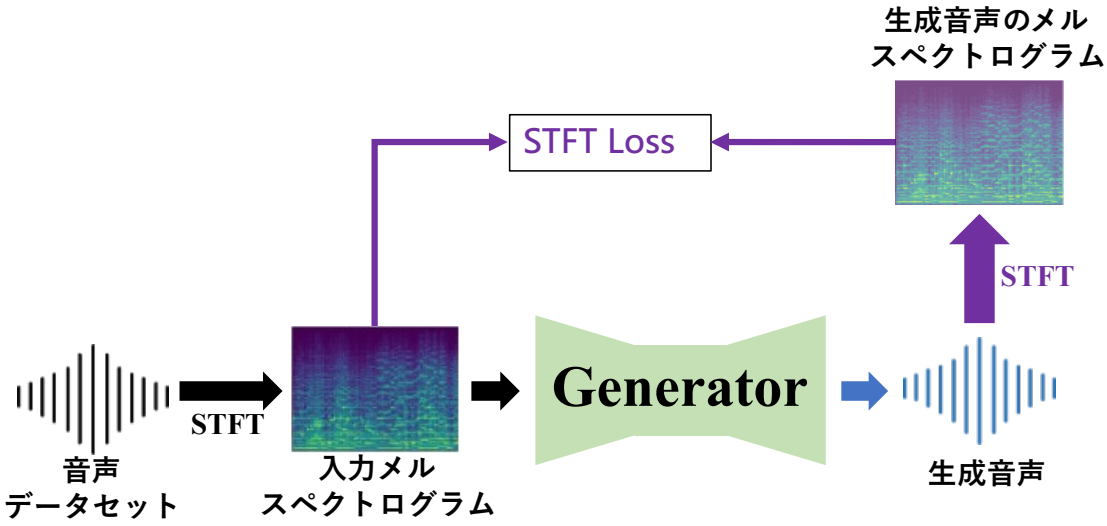


図 2.10: STFT Loss の概念図

STFT Loss は，入力のメルスペクトログラムと生成器の生成音声のメルスペクトログラムが一致するように学習するための損失関数である．GAN の学習と STFT Loss を組み合わせることで，生成器は識別機をだますように学習するだけでなく，生成音声の周波数構造が一致するように学習するため，より生成音声が高品質になる．STFT Loss を計算

するための損失関数 $L_s(G)$ の式を示す.

$$L_s(G) = E_{z\sim p(z),x\sim p_{data}}[L_{sc}(x,\hat{x}) + L_{mag}(x,\hat{x})]$$

損失関数 $L_s(G)$ は，振幅スペクトログラムの損失関数 L_{sc} と，対数スペクトログラムの損失関数 L_{mag} の 2 つから構成される．L_{sc} と L_{mag} の式を示す.

$$L_{sc}(x,\hat{x}) = \frac{\parallel |\mathrm{STFT}(x)| - |\mathrm{STFT}(\hat{x})| \parallel_F}{\parallel |\mathrm{STFT}(x)| \parallel_F}$$

$$L_{mag}(x,\hat{x}) = \frac{1}{N} \parallel \log|\mathrm{STFT}(x)| - \log|\mathrm{STFT}(\hat{x})| \parallel_1$$

ここで，$\|\cdot\|_F$ はフロベニウスノルム，$\|\cdot\|_1$ は L_1 ノルムを表す．また，$|\mathrm{STFT}(\cdot)|$ はスペクトログラムへの変換，N はスペクトログラムの要素数を表す．STFT Loss は，話者固有の音声パターンを分析するために，異なる STFT パラメータで Loss を計算する形となる．最終的な STFT Loss の式 L_{aux} を示す.

$$L_{aux}(G) = \frac{1}{M}\sum_{m=1}^{M} L_s^{(m)}(G)$$

ParallelWaveGAN の学習時には STFT Loss と GAN の Loss を組み合わせて学習を行なう．生成器の損失関数 L_G の式を示す.

$$L_G(G,D) = L_{aux} + \lambda L_{adv}(G,D)$$

2.3.3 SingGAN

SingGAN[9] は，メルスペクトログラムと励起信号を入力して音声波形を生成する Neural Vocoder であり，歌声音声を生成できるように改良されたモデルある．MelGAN[7] や ParallelWaveGAN などの既存の Neural Vocoder は，人間のスピーチ音声を生成するために設計されているため，音声変換や音声合成などの人間のスピーチ音声を生成するタスクにおいて，高品質な音声が生成可能となっている．一方で，これらの Neural Vocoder は歌声を生成するように設計されていないため，生成音声が不安定になるという欠点がある [9]. ところが，歌声を生成するために設計された Neural Vocoder は少ないため，HiFiSinger[11] などの歌声生成に関する既存の研究では，Neural Vocoder に人間のスピーチを生成するためのモデルを採用しているものが多い．SingGAN では，既存の Neural Vocoder を歌声生成タスクに使用した場合の問題点を分析し，既存のモデルを改良した歌声生成用の Neural Vocoder を提案した．SingGAN のモデル構造を図 2.11 に示す.

SingGAN の基本的な構造は既存の Neural Vocoder と似ており，GAN をベースとしたモデルの構造となっている．SingGAN と既存の Neural Vocoder の違いは大きく 2 つあ

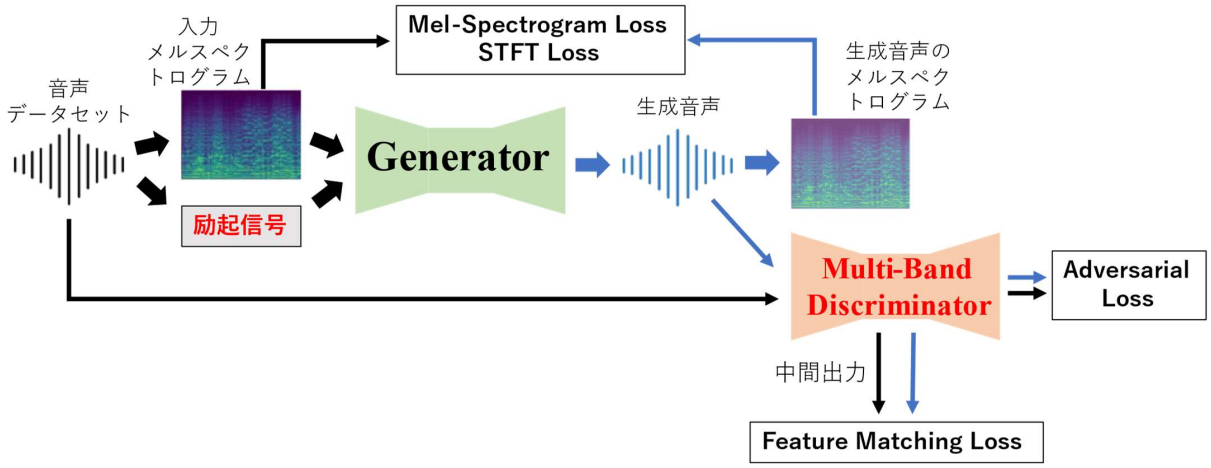


図 2.11: SingGAN のモデル構造

る．1 つ目は，生成器の入力に励起信号を使用している点である．歌声音声はスピーチ音声と比較してメロディの情報，ロングトーン (長く連続する音) を豊富に含むという特徴がある．歌声音声の生成タスクを既存の Neural Vocoder で行った場合，生成器の Dilated Causal Convolution の受容野の長さを超えるロングトーンを生成しようとしたときに生成音声が不安定になるという問題がある．SingGAN の生成器ではロングトーンを安定させるために広い受容野で励起信号の特徴を混ぜ合わせながら学習を行う AFL(Adaptive feature learning) を使用する．AFL の概念図を図 2.12 に示す．

AFL は，Dilated Causal Convolution を深く積み重ねた層で励起信号の特徴量 h_0 を分析し，メルスペクトログラムから抽出した特徴量 S を混ぜ合わせて少しづつ音声波形に変換していく．AFL の Dilated Causal Convolution は 10 層で構成されているため，広い受容野で音声信号を分析することができる．また，AFL の出力層では GAU(Gated Activation Unit) という特殊な活性化関数を使用している．GAU は，分析した特徴量をピクセルごとの重要度に応じて強度を変える．SingGAN の GAU では，メルスペクトログラムと励起信号の特徴量を分析して混ぜ合わせた特徴量 $\tanh(\mu + \alpha)$ とピクセル毎の重要度を表す特徴量 $\mathrm{sigmoid}(\mu + \beta)$ の要素積を計算して，ピクセル毎の強度の強度に応じて調整した特徴量 h を出力する．このように，広い受容野を使ってメルスペクトログラムと励起信号を混ぜ合わせながら分析する AFL を使って学習することで，SingGAN の生成器は歌声音声のロングトーンを重点的に学習することができる．2 つ目の SingGAN と既存の Neural Vocoder の違いは，複数の周波数帯域を分析する Multi-Band Discriminator を使っている点である．歌声音声はスピーチ音声と比較して高周波成分を豊富に含むという特徴がある．高周波成分を学習するように設計されている既存の Neural Vocoder は少なく，歌声音声などの高周波成分を豊富に含む音声を生成しようとしたときに，高周波部分うまく再現できずにノイズのような音声が生成されるという問題点がある．SingGAN では，高周波成分を重点的に学習するための Multi-Band Discriminator を使用する．Multi-Band

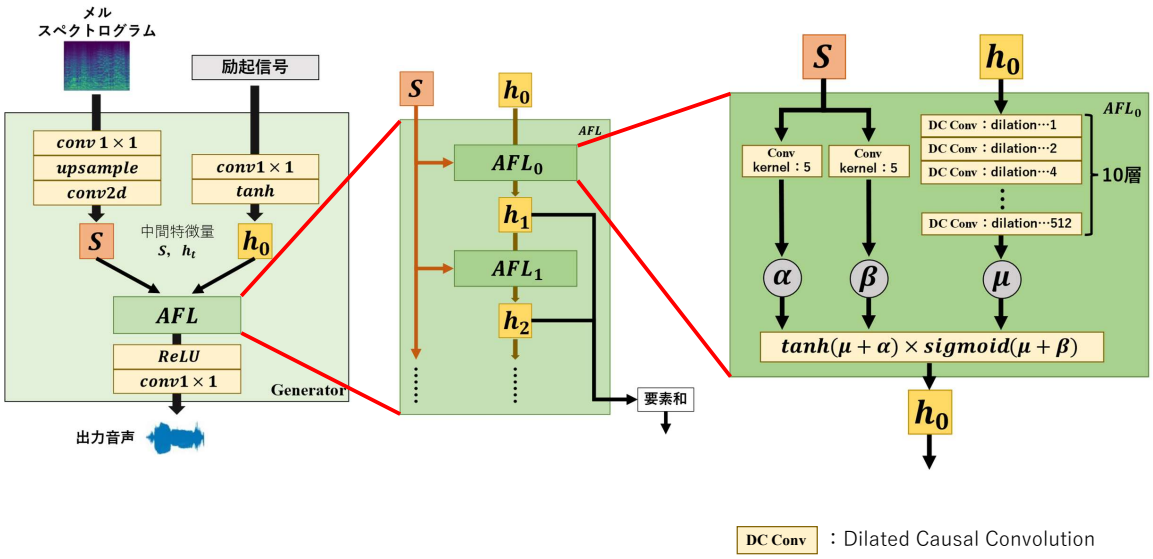


図 2.12: AFL の概念図

Discriminator の概念図を図 2.13 に示す.

図 2.13 の Global-level Discriminator は，既存の Neural Vocoder に使用されている Discriminator と同じであり，データセットの本物音声と生成器の生成音声の識別を担当する．Multi-Band Discriminator は，音声を 4 つの周波数帯域に分割し，それぞれの周波数帯域ごとに本物音声と生成音声を識別する 4 つの識別機を使用する．Global-level Discriminator に加えて Multi-Band Discriminator を使って学習することで，生成器はデータセットの本物に近い音声を生成するように学習するだけでなく，各周波数帯域ごとに本物音声に近い音声を忠実に再現するように学習することができる．

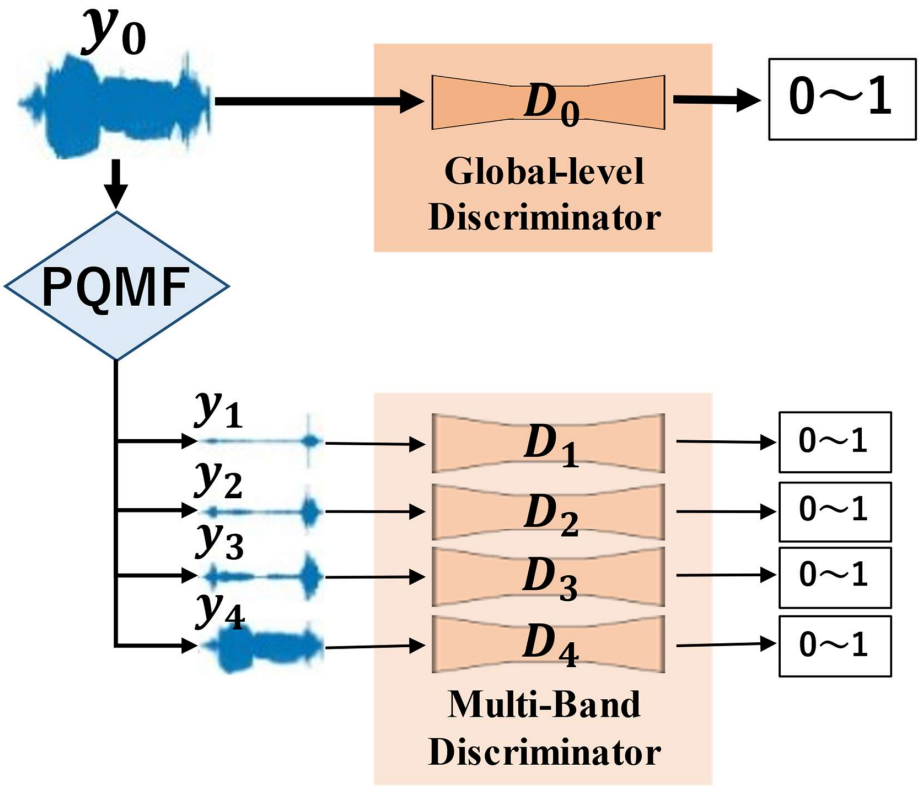


図 2.13: Multi-Band Discriminator の概念図

第3章　提案手法

本研究では，スタイル変換器に MaskCycleGAN-VC，Neural Vocoder に ParallelWaveGAN を使用する声質変換手法をベースに，高周波数成分の分析能力を向上させるための Multi-Band Discriminator，メロディ情報を学習するための励起信号と CQT スペクトログラムを使った学習方法を加えた音楽のスタイル変換手法を提案する．

3.1　Neural Vocoder

本研究では，ParallelWaveGAN をベースに，ゲーム音楽とクラシック音楽を生成するための Neural Vocoder を提案する．ParallelWaveGAN は人間の発話音声を生成するようにために設計されているが，本研究で使用するクラシック音楽，ゲーム音楽とは特徴が大きく異なるため，ベースモデルをそのまま適用すると生成音声が不安定になる．そこで，本研究ではベースモデルに 2 つの改良を加える．１つ目は，識別機に Multi-Band Discriminator を加えた点，2 つ目は CQT スペクトログラムを使った学習を行う点である．本研究で提案する Vocoder の全体構造を図 3.1 に示す．

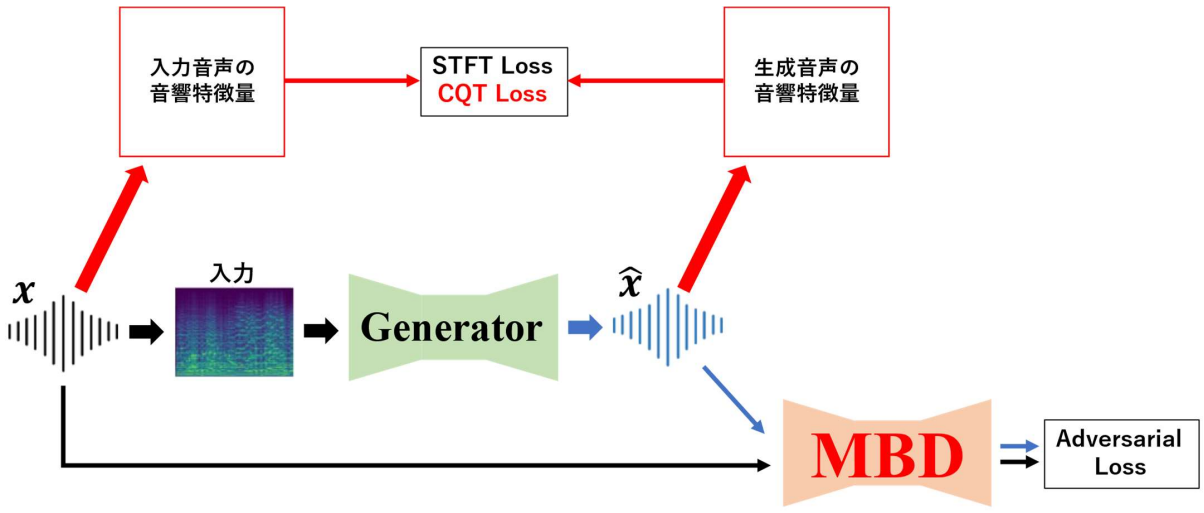


図 3.1: 提案手法の全体構造 (Vocoder)

本研究で扱うゲーム音楽，クラシック音楽は，人間のスピーチ音声と比較して周波数構造が大きく異なる．クラシック音楽は楽器音で構成されており，楽器音は人間のスピーチ

音声と比べて周波数帯域が広い．特に，高周波数成分を豊富に含む点が人間のスピーチとの大きな違いである．また，ゲーム音楽も高周波数成分を豊富に含んでおり，メルスペクトログラムからクラシック音楽よりゲーム音楽が高周波数成分を多く含んでいることがわかる．本研究では，高周波数成分を学習するために Multi-Band Discriminator を使った学習を行う．本研究で使用する Multi-Band Discriminator の全体構造を図 3.2 に示す．

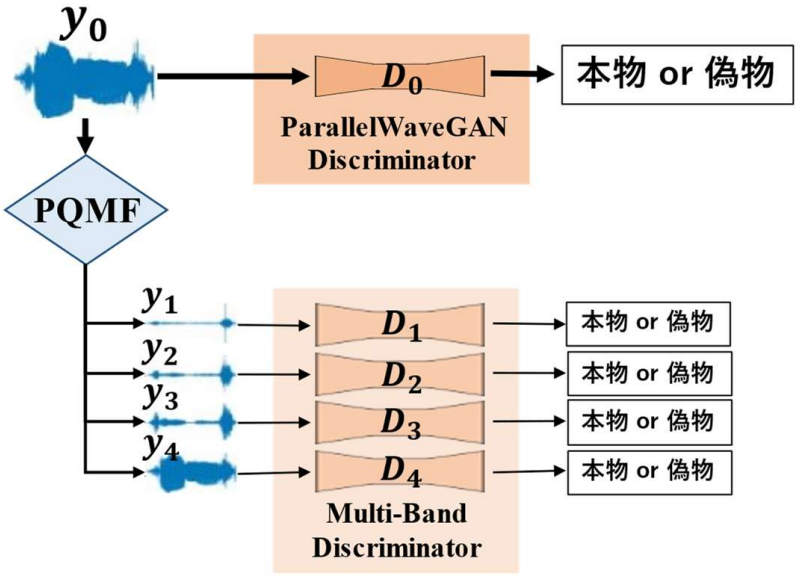


図 3.2: 本研究で使用する Multi-Band Discriminator

本研究では，SingGAN を参考に，データセットの本物音声と生成器の生成音声を識別する 1 つの識別機と，音声を 4 つの周波数帯域に分割して分割した音声をそれぞれ分析する 4 つの識別機から構成される Multi-Band Discriminator を使用する．5 つの識別機には，ParallelWaveGAN の識別機の構造をそのまま利用している．SingGAN の識別機と ParallelWaveGAN の識別機はどちらも入力された音声が本物か偽物か識別するタスクを担っているが，出力のサイズが異なる異なる．SingGAN の識別機は出力が単一の値であり，入力音声が本物か偽物かを表している．ParallelWaveGAN の識別機は 1 次元のベクトルであり，入力音声を Patch と呼ばれる複数の領域に分割して，Patch ごとに本物か偽物か識別する (Patch GAN[12])．本研究では，5 つの識別機すべてに ParallelWaveGAN の識別機をそのまま利用して学習を行う．

2 つ目の改良点は，CQT スペクトログラムを使った損失関数を使って学習を行った点である．ParallelWaveGAN やそのほかの既存手法では音声を STFT(短時間フーリエ変換)して得られるスペクトログラムを使って学習を行う手法が一般的である．STFT は窓関数を使って音声信号の一部分を切り出してフーリエ変換する操作を，窓関数をずらしながら繰り返し行うことで周波数構造の時間変化を分析することができる．STFT は低周波から高周波まで，すべての周波数帯域で同じ長さの窓関数を使用するため，周期の短い高周

波成分は十分なサンプルが得られるが，周期の長い低周波成分は十分なサンプルが得られず，高周波と低周波でサンプルに偏りが生じる．また，ドレミなどの音階は 1 オクターブ音が高くなるごとに周波数は倍になる．例えば，440Hz のラの音を 1 オクターブ上げたラの音は 880Hz である．そのため，音階の変化を STFT で得られるスペクトログラムの周波数軸で見ると，等間隔になっていない．CQT スペクトログラムは，これらの 2 つの問題点を改善した音響特徴量である．CQT スペクトログラムは，音声を CQT(定 Q 変換) して得られる特徴量である．CQT では，窓関数で音声の一部を切り出してフーリエ変換する操作について，切り出す長さが周波数によって異なる．低周波数成分を分析するときは窓関数を長くして切り出す長さが長くなり，高周波数成分を分析するときは窓関数を短くして切り出す長さが短くなる．この時，分析する周波数ごとに周期の数が同じになるように窓関数の長さを調整する．また，CQT では周波数軸が対数スケールになっている．対数スケールにすることで，音階の変化と CQT スペクトログラムの周波数軸の変化が等間隔になる．本研究で扱う音声はクラシック音楽やゲーム音楽であり，ベースモデルに CQT スペクトログラムを使った音楽の音階情報を学習するための損失関数 CQT Loss を追加する．CQT Loss の概念図を図 3.3 に示す．

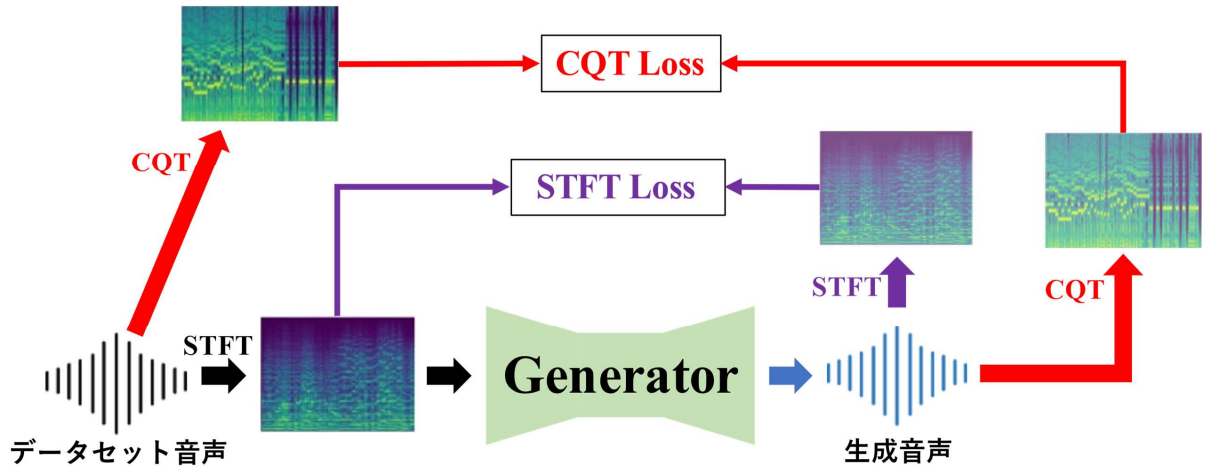


図 3.3: CQT Loss の概念図

CQT Loss は STFT Loss を参考に設計した損失関数であり，基本的な構造は STFT Loss に似ている．CQT Loss は，生成器の CQT スペクトログラムと入力音声の CQT スペクトログラムのピクセル差によって計算する．これによって，生成器は生成音声の音階情報と入力音声の音階情報が一致するように学習することができる．CQT Loss の損失関数 L_{CQT} の式を示す．

$$L_{CQT}(G) = E_{z\sim p(z),x\sim p_{data}}[L_{cqtsc}(x,\hat{x}) + L_{cqtmag}(x,\hat{x})]$$

損失関数 $L_{CQT}(G)$ は，STFT Loss を参考に設計しており，STFT Loss と同様に CQT

スペクトログラムの損失関数 L_{cqtsc} と，対数 CQT スペクトログラムの損失関数 L_{cqtmag} の 2 つから構成される．L_{cqtsc} と L_{cqtmag} の式を示す．

$$L_{cqtsc}(x, \hat{x}) = \frac{\| \, |\mathrm{CQT}(x)| - |\mathrm{CQT}(\hat{x})| \, \|_F}{\| \, |\mathrm{CQT}(x)| \, \|_F}$$

$$L_{cqtmag}(x, \hat{x}) = \frac{1}{N} \| \log|\mathrm{CQT}(x)| - \log|\mathrm{CQT}(\hat{x})| \, \|_1$$

ここで，$|\mathrm{CQT}(\cdot)|$ は CQT スペクトログラムへの変換を表す．ベースモデルの STFT Loss と CQT Loss を組み合わせた音響特徴量の損失関数 L_{aux} の式を示す．

$$L_{aux}(G) = L_{CQT}(G) + \frac{1}{M} \sum_{m=1}^{M} L_s^{(m)}(G)$$

L_{aux} と GAN の Loss を組み合わせた最終的な生成器の損失関数 L_G の式を示す．

$$L_G(G, D_k) = L_{aux} + \sum_{k=1}^{K} \lambda L_{adv}(G, D_k)$$

ここで，K は識別器の個数を表している．本研究では 5 個で構成される Multi-Band Discriminator を使用するため，$K = 5$ である．

3.2 音響特徴量のスタイル変換器

本研究では MaskCycleGAN-VC をベースに，ゲーム音楽のメルスペクトログラムとクラシック音楽のメルスペクトログラムのスタイル変換を行うためのスタイル変換器を提案する．MaskCycleGAN-VC も，ParallelWaveGAN と同様に人間のスピーチ音声を対象としたスタイル変換器である．本研究では，クラシック音楽とゲーム音楽に特化したスタイル変換を行うために，入力する音響特徴量を増やして学習を行うように改良を加える．また，クラシック音楽とゲーム音楽は人間の話者変換と比べて変換前後の特徴の差が大きいため，学習を簡単にするために音声に前処理を施す．本研究で提案するスタイル変換器の全体構造を図 3.4 に示す．

ベースモデルでは，2 つの生成器はメルスペクトログラムを入力し，メルスペクトログラムを出力する．本研究では，音声の基本周波数から生成した励起信号と CQT スペクトログラム追加で入力する．本研究で行う音楽のスタイル変換において，スタイル変換器の目標は変換前後で音階情報はそのままで，クラシック音楽からゲーム音楽，ゲーム音楽からクラシック音楽に音色を変換することが目標である．メルスペクトログラムは音声の周波数構造を表した特徴量であるが，メルスペクトログラムから音階情報を分析するのは難しい．そこで，変換前の音階情報を信号で表現した励起信号と，変換前の周波数構造を音階に特化した形で表現した CQT スペクトログラムの 2 つをを補助の特徴量として生成器

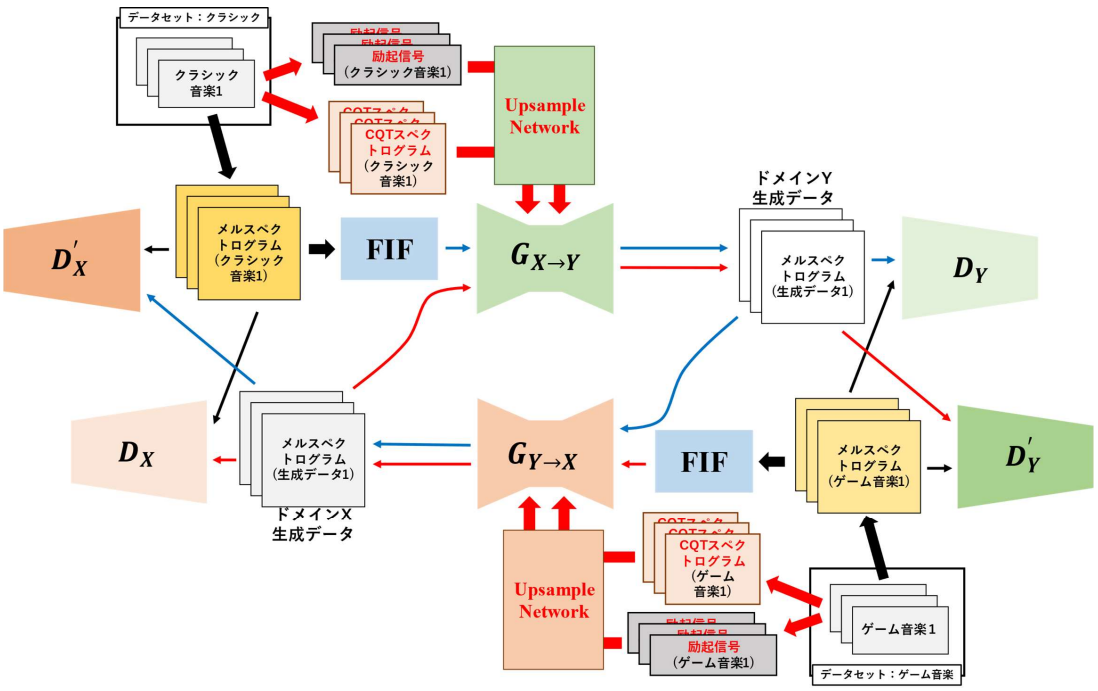


図 3.4: 提案手法の全体構造 (スタイル変換器)

に入力する．励起信号と CQT スペクトログラムは Upsample Network で分析，拡張された後，メルスペクトログラムの特徴量と混ぜ合わせて分析を行う．スタイル変換器の生成器全体構造を図 3.5 に示す．

CQT スペクトログラムは Upsample Network で分析，サイズ調整され，入力メルスペクトログラムを FIF 処理したものと結合する．FIF は，マスクを欠落した部分の周波数構造を周辺の周波数構造を参考に修復するタスクを行っており，CQT スペクトログラムの特徴を合わせて学習することで生成器は CQT スペクトログラムの音階情報とマスクの情報を参考にメルスペクトログラムの周波数構造を生成するように学習を行う．このように，生成器がメルスペクトログラムを生成するタスクに CQT スペクトログラムを補助特徴量として追加することで，音階情報を強調したメルスペクトログラムを生成することができる．

励起信号は，Upsample Network で分析，サイズ調整され，生成器の Res Brock で分析する．Res Brock では，メルスペクトログラムの特徴量と励起信号の特徴量を 1 次元畳み込みで分析し，GLU で二つの特徴量を混ぜ合わせる．GLU は，SingGAN の AFL を参考に設計した．このように，メルスペクトログラムの特徴を詳しく分析する Res Brock に励起信号の特徴量を加えることで，生成器は励起信号の音階情報を付加したメルスペクトログラムを生成できるようになる．

本研究では，ゲーム音楽とクラシック音楽の特徴を分析しやすくするために，データセットの音声に前処理を施す．ベースモデルは人間のスピーチ音声について，話者の変換

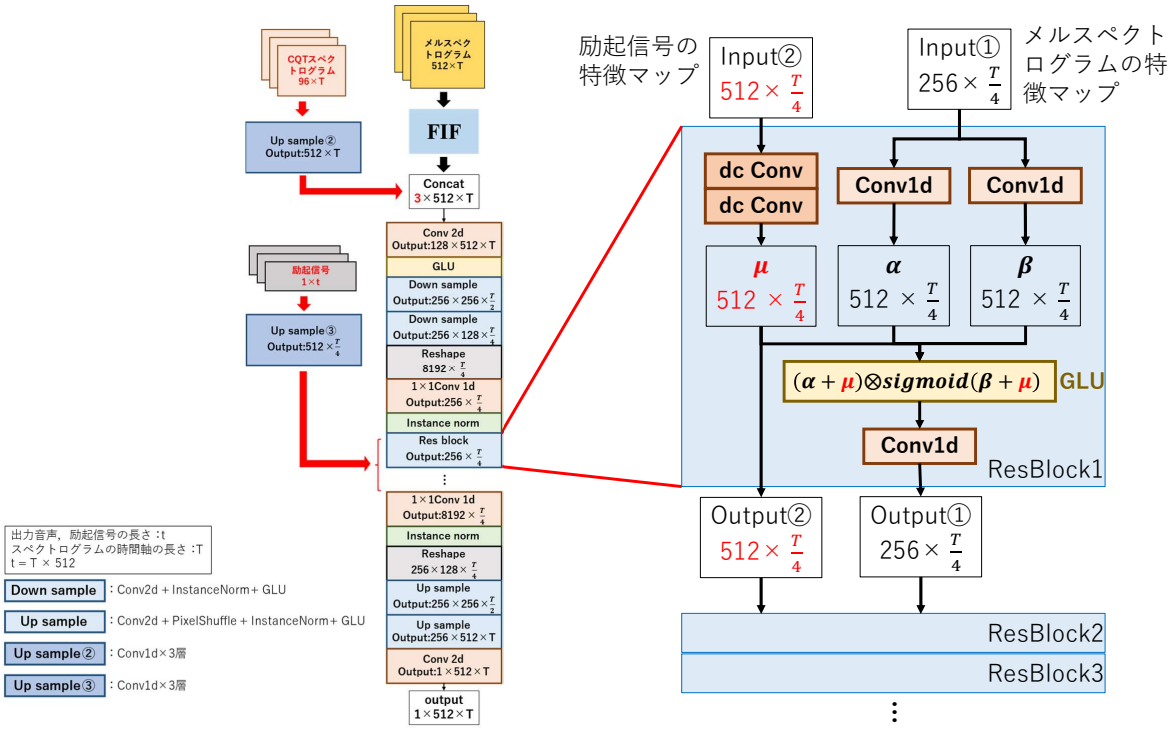
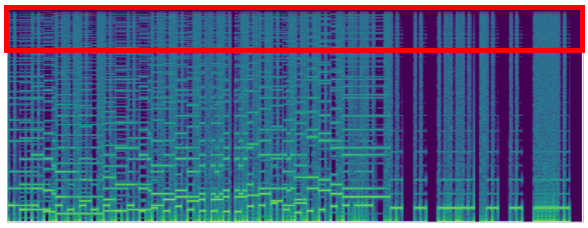


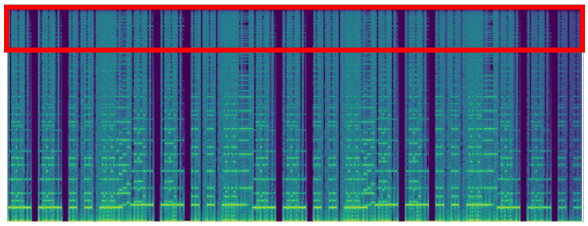
図 3.5: スタイル変換器の生成器全体構造

を行うように設計されたモデルであり，変換前と変換後で特徴は大きく変わらない．例えば，男性が「こんにちは」と話す音声と女性が「こんにちは」と話す音声は周波数構造の違いは少ない．しかし，本研究では扱うゲーム音楽とクラシック音楽は周波数構造が大きく異なる．特に，高周波成分の違いが大きく，スペクトログラムで見るとその違いは顕著である．ゲーム音楽のメルスペクトログラムを図 3.6 に，クラシック音楽のメルスペクトログラムを図 3.7 に示す．

図 3.6 と図 3.7 の赤い四角部分は同じ周波数帯域を示している．ゲーム音楽は高周波部分にも豊富に周波数成分を含んでいるが，クラシック音楽は高周波成分が少ない．これらの音声をそのまま訓練データとして使用した場合，MaskCycleGAN-VC がスタイル変換を正しく学習できない可能性があると考えた．例えば，識別機はデータセットのメルスペクトログラムと生成器の生成メルスペクトログラムの本物，偽物を識別するように設計しているが，高周波成分の有無を学習してしまう可能性がある．そこで，MaskCycleGAN-VC が意図しない方向に学習することを防ぐために，データセット音声にローパスフィルタを適用して高周波成分を除去する．図 3.6 と同じ音声にローパスフィルタで高周波成分を除去した音声のメルスペクトログラムを図 3.8 に，図 3.7 と同じ音声にローパスフィルタで

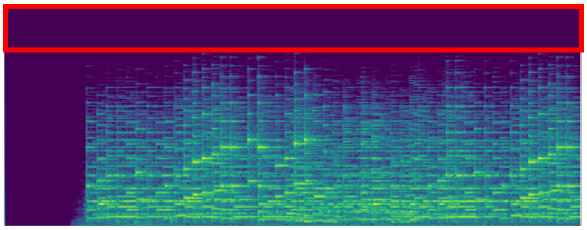


(a)

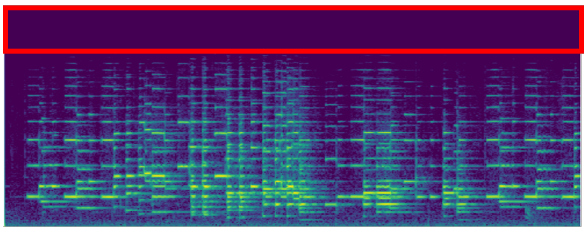


(b)

図 3.6: メルスペクトログラム (ゲーム音楽)



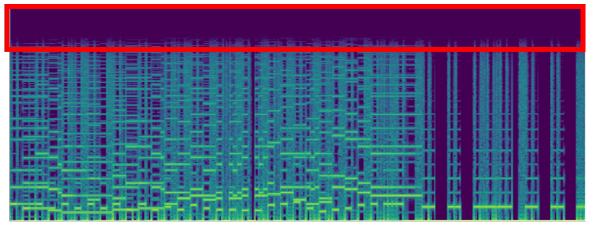
(c)



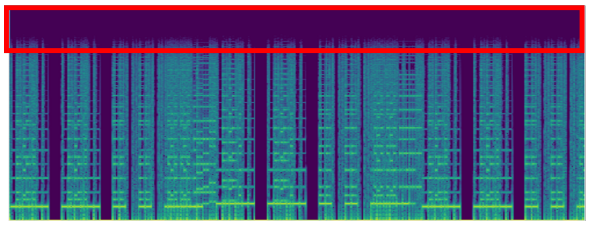
(d)

図 3.7: メルスペクトログラム (クラシック)

高周波成分を除去した音声のメルスペクトログラムを図 3.9 に示す．高周波成分を除去することでゲーム音楽とクラシック音楽のメルスペクトログラムは周波数帯域の差が少なくなり，モデルは周波数構造のスタイル変換を正しく学習することが期待できる．

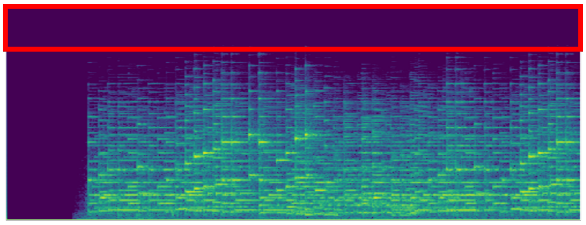


(a)

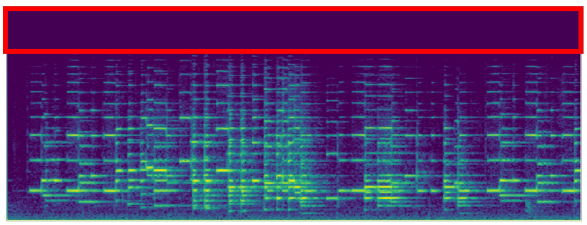


(b)

図 3.8: ローパスフィルタを適用したゲーム音楽のメルスペクトログラム



(c)



(d)

図 3.9: ローパスフィルタを適用したクラシック音楽のメルスペクトログラム

第4章　実験

4.1　データセット

本研究で使用したデータセットはクラシック音楽のデータセットとゲーム音楽のデータセットである．

4.1.1　ゲーム音楽のデータセット

NES MDB(Nintendo Entertainment System Music Database)[13] は，Nintendo Entertainment System という海外版のファミリーコンピュータに使用されていた音楽を集めたデータセットである．データセットには 397 種類のゲームで使用された 5278 曲のサンドトラックがあり，10 秒程度の短い効果音から 2 分程度の長い BGM などが収録されている．

4.1.2　クラシック音楽のデータセット

MusicNet[14] は，シューベルトやモーツァルトのクラシック音楽を収録したデータセットである．ピアノや管楽器，弦楽器などの楽器単体のソロ演奏や複数楽器の重奏などさまざまなパターンのクラシック音楽を含む．データセットには 300 秒程度の楽曲が 330 曲収録されている．

4.2　データセットの前処理

すべての音声は，サンプリングレートを 22050 にそろえる．また，音響特徴量スタイル変換器の学習のために，データセット音声にローパスフィルタを適用して高周波数成分を除去した音声を用意する．本研究で使用したローパスフィルタは，周波数 5000Hz 以上の周波数成分を除去する．また，Neural Vocoder，音響特徴量スタイル変換器の訓練用データを作成するために，音声からメルスペクトログラムを生成する．スペクトログラムに変換するための STFT パラメーターは，FFT サイズ 2048，FFT の移動幅 512，窓関数はハニング窓を使用する．メル周波数は 512 としてスペクトログラムからメルスペクトログラムに変換する．メルスペクトログラムの周波数軸のサイズは 512，時間軸のサイズは元の

音声によって異なる．本研究で使用したデータセット音声をメルスペクトログラムに変換したものの一部を図 4.1，図 4.2，図 4.3，図 4.4 に示す．

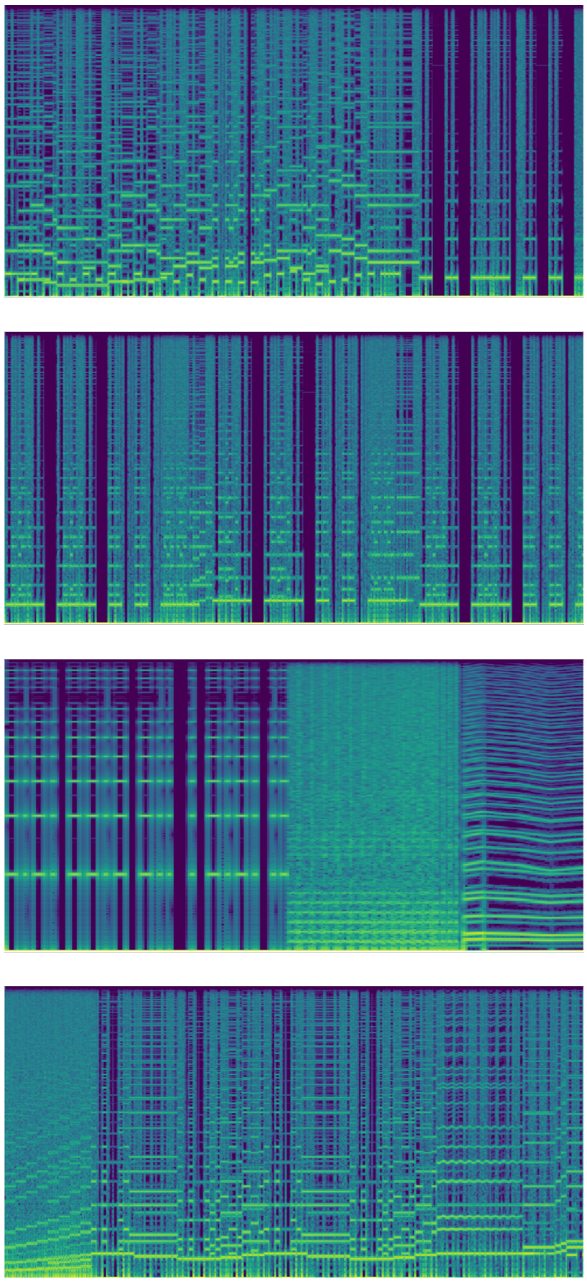


図 4.1: NES MDB データセットの音声をメルスペクトログラムに変換したものの一部

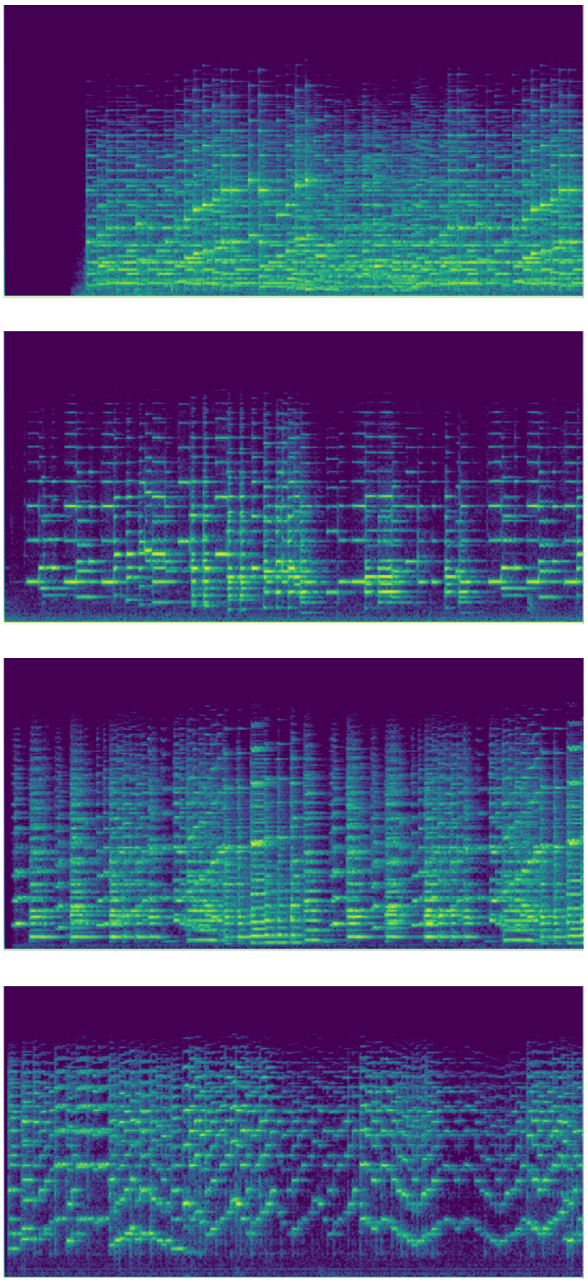


図 4.2: MusicNet データセットの音声をメルスペクトログラムに変換したものの一部

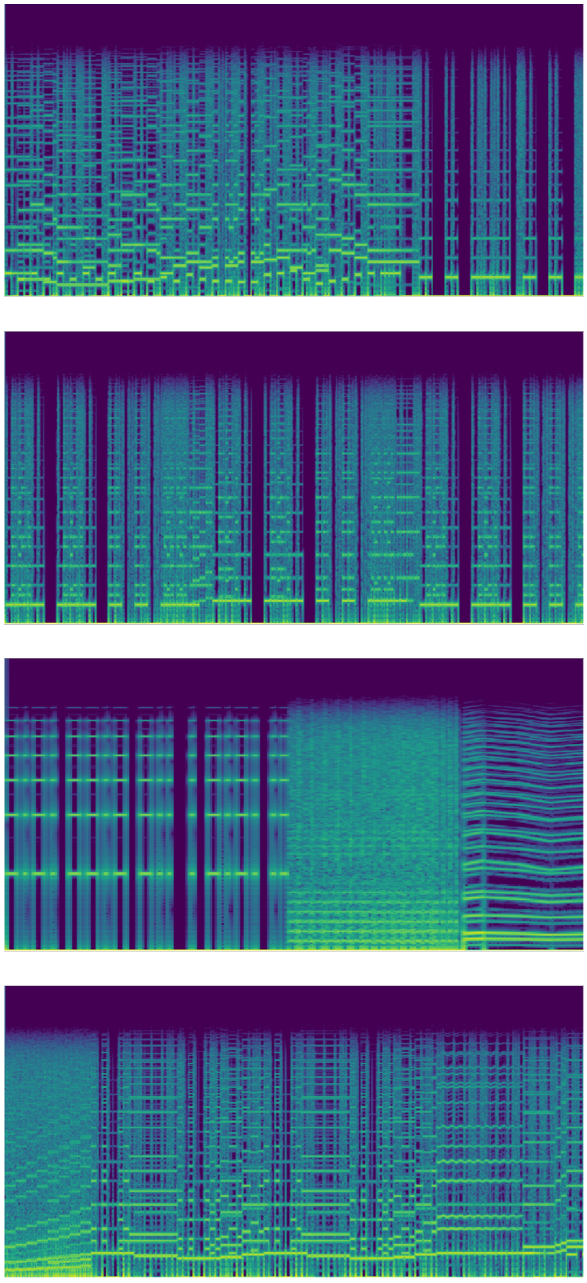


図 4.3: NES MDB データセットの音声にローパスフィルタを適用してメルスペクトログラムに変換したものの一部

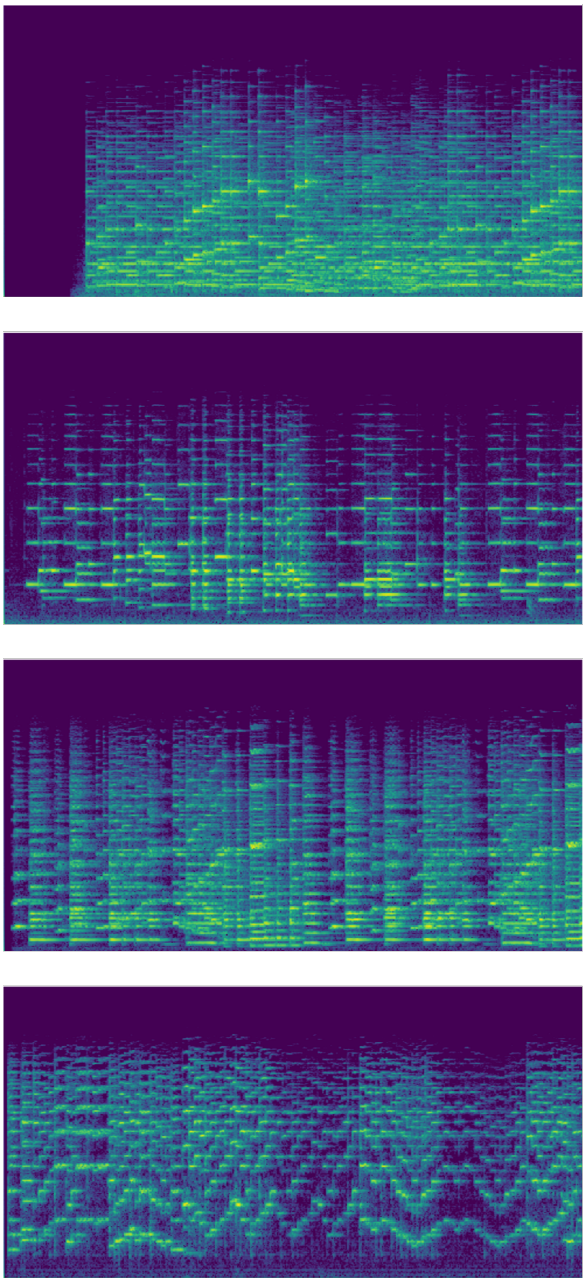


図 4.4: MusicNet データセットの音声にローパスフィルタを適用してメルスペクトログラムに変換したものの一部

4.3 評価方法

本研究ではアンケートを使用して生成音声を評価する．本研究で行う音楽のスタイル変換は正解データがない教師なし学習のため，正解データなどを使ってモデルを評価するこ

とが難しい．そこで，ベースモデルと提案手法の性能を第三者によって評価するためにアンケート調査を行う．アンケートによる評価は，Neural Vocoder と音楽スタイル変換の評価で合わせて 2 回行う．アンケートの方法は，被験者が既存手法，提案手法でモデルを学習して生成した音声を聞き，最も高品質な音声を選択する方法で行う．

4.4 実験

本研究では，3.1 節の Neural Vocoder と 3.2 節の音響特徴量スタイル変換器を組み合わせて音楽スタイル変換を行う．Neural Vocoder と音響特徴量スタイル変換器は別々に学習を行う．Neural Vocoder の訓練が完了したら，STFT Loss とアンケート評価を使って最適な Neural Vocoder を選択する．最後に，選択した Neural Vocoder と音響特徴量スタイル変換器を組み合わせて音楽のスタイル変換を行う．音楽のスタイル変換で生成した音声はアンケートを使って評価する．

4.4.1 Neural Vocoder の実験

この節では，高品質なゲーム音楽，クラシック音楽を生成するための最適な Neural Vocoder の作成手法を探るための実験について述べる．本研究では Multi-Band Discriminator と CQT Loss の効果を調べるために，以下の 4 つの手法で実験を行った．

既存手法 ベースモデル (ParallelWaveGAN) をそのまま使用

提案手法 1-1 ベースモデルに Multi-Band Discriminator を追加

提案手法 1-2 ベースモデルに CQT Loss を追加

提案手法 1-3 ベースモデルに CQT Loss と Multi-Band Discriminator を追加

既存手法と提案手法で学習したモデルを使って生成したクラシック音楽とゲーム音楽の STFT Loss を表 4.1 に示す．また，アンケート調査の結果を表 4.2 と図 4.5，4.6 に示す．

表 4.1: 実験 1 で生成した音声の STFT Loss

実験手法	クラシック音楽	ゲーム音楽
既存手法	0.61966	0.49135
提案手法 1-1	0.57327	0.40650
提案手法 1-2	0.58777	0.53156
提案手法 1-3	0.55939	0.39106

表 4.2: 実験 1 で生成した音声のアンケート結果

実験手法	クラシック音楽	ゲーム音楽	クラシック音楽+ゲーム音楽
既存手法	9	7	16
提案手法 1-1	17	14	31
提案手法 1-2	9	5	14
提案手法 1-3	7	16	23

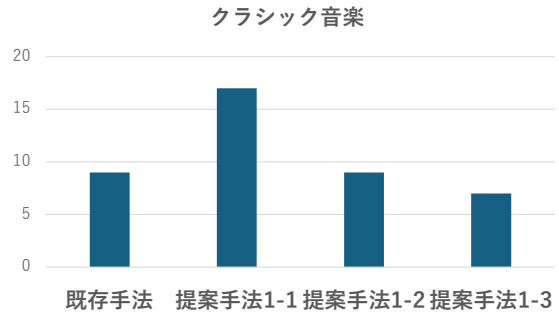


図 4.5: 実験 1 のアンケート結果 (クラシック音楽)

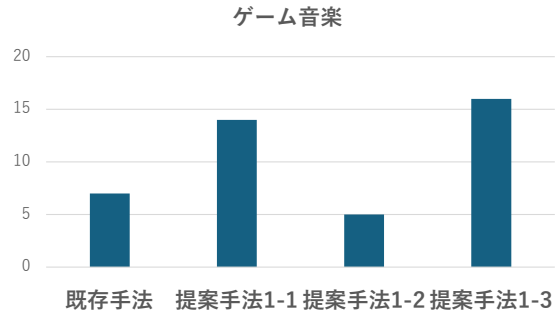


図 4.6: 実験 1 のアンケート結果 (ゲーム音楽)

既存手法で訓練終了後のモデルが生成したクラシック音楽は，高い音で音階が少し乱れたり，ロングトーンが不安定になっている．ベースモデルは楽器音の音階情報や楽器音特有の長く連続する音を学習できないため，楽器音で構成されるクラシック音楽をモデルがうまく学習できていないと考えられる．訓練終了後のモデルが生成したゲーム音楽は，ノイズが多く含まれている．ゲーム音楽は，図 4.1 のゲーム音楽のメルスペクトログラムからもわかるように，高周波成分を豊富に含んでいる．ベースモデルは人間の発話音声の周波数帯域を学習するように設計されているが，ゲーム音楽はその周波数帯域より広いため，高周波成分を学習できずにノイズを生成していると考えられる．

手法 1-1 では，ベースモデルに Multi-Band Discriminator(以降では MBD とする) を追加して実験を行った．MBD を追加することで，訓練終了後のモデルが生成した音声が改善された．表 4.1 の STFTLoss の値は，クラシック音楽とゲーム音楽の両方でベースモデルから結果が改善したことがわかる．これは，MBD によって周波数帯域ごとに詳しく分析することができ，周波数構造の再現能力が向上し，STFT Loss の結果が改善されたと考えられる．特に，ゲーム音楽は高周波数成分を豊富に含んでいるため，MBD による STFT Loss の減少率が大きい．手法 1-1 から，MBD によってクラシック音楽，ゲーム音楽の生成能力が向上し，ゲーム音楽を分析する際には高周波成分を考慮することが結果に大きく影響を及ぼすことが分かった．

ベースモデルに CQT Loss を追加した場合，高い音の生成能力が少し改善された．ベー

スモデルが生成したクラシック音楽，音が高くなるほど上手く再現できずにノイズが生成されているが，CQT Loss を追加することで高い音でもメロディが乱れず，生成さ音声の音階は安定している．これは，学習が困難な高周波成分においても，CQT スペクトログラムを使って学習することで高周波部分を音階として学習することができ，生成能力が向上したのではないかと考えられる．ゲーム音楽は，既存手法と比較すると高い音階の生成能力は少し向上したが，ノイズが多く含まれる点は改善されていない．また，ゲーム音楽の STFT Loss の値を比較すると，手法 1-2 は手法 1-1 のベースモデルより悪くなっている．これは，STFT Loss と CQT Loss の学習タスクが異なり，お互いの学習が競合した結果 STFT Loss が悪くなったのではないかと考えられる．

手法 1-3 では，ベースモデルに CQT Loss と Multi-Band Discriminator の両方を追加して実験を行った．STFT Loss の値を見ると，手法 1-3 はクラシック音楽とゲーム音楽の両方で最も良い値となっている．特にゲーム音楽では，CQT Loss のみで学習した場合はベースモデルより STFT Loss の値が悪くなっているが，MBD と CQT Loss を組み合わせることで手法 1-1 の MBD のみで学習したときよりも良い値になった．これは，MBD が周波数成分の学習を補助することで，STFT Loss と CQT Loss の学習の競合を抑えられているのではないかと考えられる．訓練済みのモデルが生成音声したクラシック音楽は大きな改善は見られなかったが，高い音のメロディの乱れやロングトーンの不安定さが改善された．生成したゲーム音楽は高周波成分の再現力が向上し，ノイズが低減された．さらに，音階の乱れも改善された．アンケート調査の結果を見ると，クラシック音楽は手法 1-1 が最も良い結果になった．手法 1-2 と手法 1-3 で生成した音声は音階の滑らかな変化を表現できていないため，知覚品質が低下したのではないかと考えられる．STFT Loss の結果とアンケート調査の結果から，クラシック音楽の生成には手法 1-1 で作成したモデル，ゲーム音楽の生成には手法 1-3 で作成したモデルを使用した．

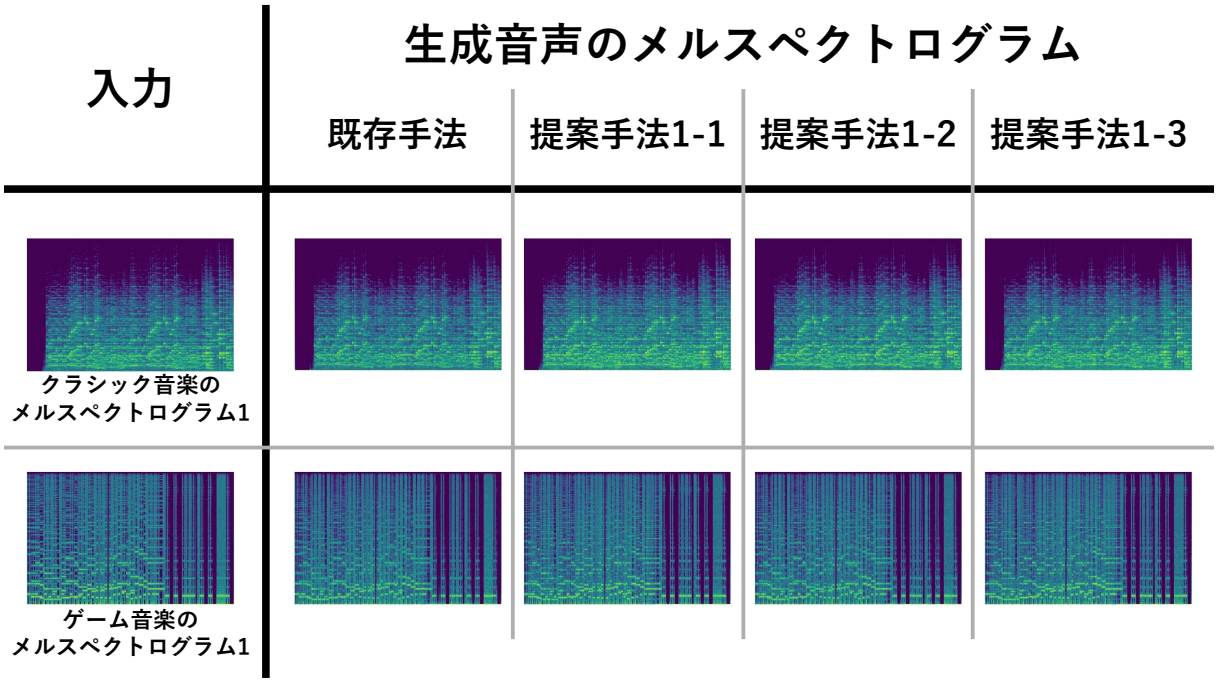


図 4.7: Neural Vocoder 生成音声のメルスペクトログラム

4.4.2 音響特徴量スタイル変換器の実験

この節では，ゲーム音楽とクラシック音楽で，メルスペクトログラムのスタイル変換を行うスタイル変換器の実験について述べる．本研究では入力の音響特徴量，音声の前処理を変更して以下の 4 つの手法で実験を行った．

既存手法 ベースモデル (MaskCycleGAN-VC) をそのまま使用

提案手法 2-1 ベースモデルに励起信号を追加で入力

提案手法 2-2 高周波成分を除去した音声を使用

提案手法 2-3 ベースモデルに励起信号，CQT スペクトログラムを追加で入力

クラシック音楽からゲーム音楽への変換では，手法 2-3 が最もゲーム音楽の特徴に近いメルスペクトログラムを生成できていることがわかる．これは，CQT スペクトログラムの特徴とゲーム音楽のメルスペクトログラムの特徴が似ているため，変換の際に CQT スペクトログラムの特徴をうまく反映したメルスペクトログラムを生成するように学習できたのではないかと考えられる．一方で，ゲーム音楽からクラシック音楽への変換では，手法 2-3 が最も悪い結果となっている．手法 2-3 のゲーム音楽からクラシック音楽への変換は，どんな入力に対してもモデルの出力が同じになるモード崩壊を起こしている．これ

は，ゲーム音楽のメルスペクトログラムを生成する場合とは違ってクラシック音楽のメルスペクトログラムと CQT スペクトログラムの特徴が大きく異なるため，CQT スペクトログラムの特徴が，メルスペクトログラムを生成するための学習を妨げているのではないかと考えられる．クラシック音楽からゲーム音楽，ゲーム音楽からクラシック音楽の両方の変換結果を比較すると，全体的にクラシック音楽からゲーム音楽への変換がうまくいっているものが多い．これは，それぞれの音楽に含まれる情報量の多さが違うことが要因ではないかと考えられる．データセットのメルスペクトログラムを比較すると，ゲーム音楽のメルスペクトログラムは高周波成分を豊富に含むが周波数構造は単純であり，逆にクラシック音楽のメルスペクトログラムは高周波成分は少ないが，周波数構造が複雑である．クラシック音楽からゲーム音楽への変換は複雑な周波数構造から単純な周波数構造への変換のため，学習が簡単なのに対し，ゲーム音楽からクラシック音楽への変換は単純な周波数構造構造から複雑な周波数構造への変換のため，足りない情報を補間するように生成器が学習する必要があり，学習が困難なのではないかと考えた．一方で，手法 2-2 の結果を見ると，ゲーム音楽からクラシック音楽への変換が少し良くなっている．既存手法や手法 2-1 では低周波成分がぼやけており，高周波成分は不自然な連続構造となっていが，手法 2-2 では高周波成分が少なく，低周波の構造にぼやけが少ない．このことから，ゲーム音楽からクラシック音楽への変換においては，変換前後の音声で周波数帯域をそろえることで周波数構造を学習し，ノイズを低減できることがわかった．手法 2-2 のクラシック音楽からゲーム音楽への変換結果を見ると，全体的に結果が悪い．ゲーム音楽のメルスペクトログラムは低周波から高周波まで幅広い周波数帯域で規則的な連続構造となっている．しかし，高周波成分を除去することでゲーム音楽の周波数の連続構造の学習が困難になり，学習を妨げているのではないかと考えられる．

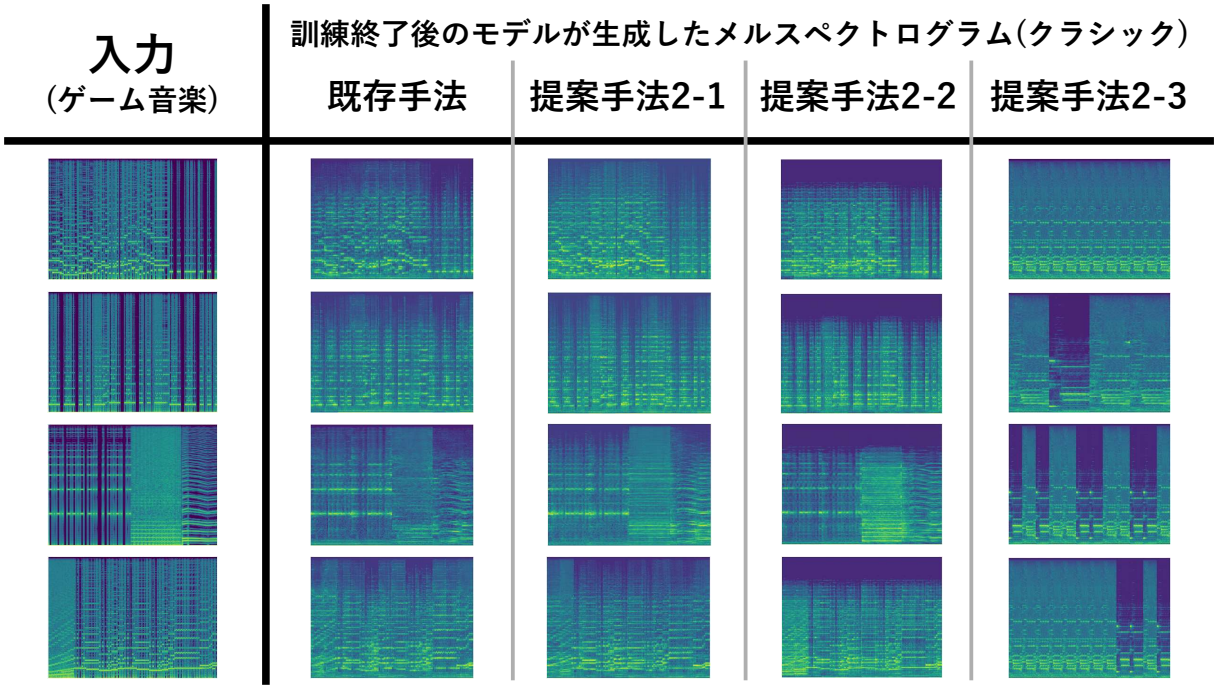


図 4.8: メルスペクトログラムのスタイル変換結果 (ゲーム音楽からクラシック音楽に変換)

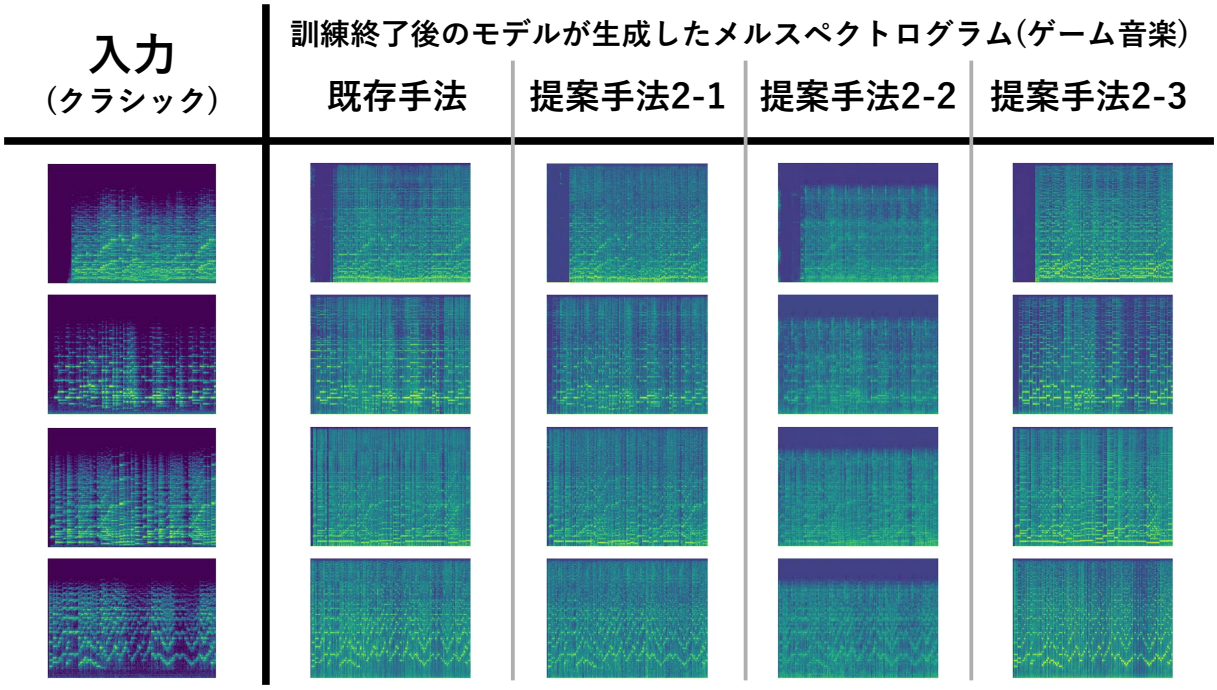


図 4.9: メルスペクトログラムのスタイル変換結果 (クラシック音楽からゲーム音楽に変換)

4.4.3　音楽のスタイル変換

この説では，4.4.1 節で行った Neural Vocoder の実験と，4.4.2 節で行ったスタイル変換器の実験から得られたモデルを組み合わせて End-to-end の音楽スタイル変換の実験ついて述べる．スタイル変換器は，実験 2 の 4 つのモデルを使用する．Neural Vocoder は，実験 1 のアンケート結果から，ゲーム音楽生成用の Neural Vocoder に手法 1-3 で作成したモデル，クラシック音楽生成用の Neural Vocoder に手法 1-1 で作成したモデルを使用した．なお，既存手法の実験では Neural Vocoder にベースモデルを採用した．実験に使用したスタイル変換器と Neural Vocoder の組み合わせを以下にまとめる．

既存手法 ベースモデル

クラシック音楽からゲーム音楽への変換
既存手法のスタイル変換器 + 既存手法の Neural Vocoder

ゲーム音楽からクラシック音楽への変換
既存手法のスタイル変換器 + 既存手法の Neural Vocoder

提案手法 3-1 スタイル変換器に励起信号を追加で入力

クラシック音楽からゲーム音楽への変換
手法 2-1 のスタイル変換器 + 手法 1-3 の Neural Vocoder

ゲーム音楽からクラシック音楽への変換
手法 2-1 のスタイル変換器 + 手法 1-1 の Neural Vocoder

提案手法 3-2 データセット音声の高周波成分を除去

クラシック音楽からゲーム音楽への変換
手法 2-2 のスタイル変換器 + 手法 1-3 の Neural Vocoder

ゲーム音楽からクラシック音楽への変換
手法 2-2 のスタイル変換器 + 手法 1-1 の Neural Vocoder

提案手法 3-3 スタイル変換器に励起信号と CQT スペクトログラムを追加で入力

クラシック音楽からゲーム音楽への変換
手法 2-3 のスタイル変換器 + 手法 1-3 の Neural Vocoder

ゲーム音楽からクラシック音楽への変換
手法 2-3 のスタイル変換器 + 手法 1-1 の Neural Vocoder

4 つの手法で生成した音声のアンケート調査結果を表 4.3 と図 4.10，4.11 に示す．クラシックからゲーム音楽への変換でアンケート結果が最も良かったのは，手法 3-3 であった．手法 3-3 で生成したゲーム音楽は，ノイズが少なく，音階の違いがはっきりと判る音声が

多い．このことから，音階情報を学習するうえで，CQT スペクトログラムを使って学習することが効果的であることがわかった．また，異なる楽器が複数使用されているクラシック音楽からの変換においても，手法 3-3 で生成した音声が最も良く，各楽器のメロディを反映した音声を生成することができた．一方で，手法 3-3 のモデルは，同じ楽器の和音が含まれているクラシック音楽からの変換に弱く，生成された音声は主旋律よりも副旋律を強調されていた．これは，CQT スペクトログラムが音声の細かい特徴を消したものであり，CQT スペクトログラムの特徴を強調して変換を行う手法 3-3 のスタイル変換器で和音の特徴の学習に失敗しているのではないかと考えられる．手法 3-3 に次いで生成音声のアンケート結果が良かったものは手法 3-1 である．手法 3-1 が生成したゲーム音楽には少し目立つノイズが含まれているが，同じ楽器の和音を含むクラシック音楽からの変換に強い．クラシック音楽からゲーム音楽への変換で最も悪かったのは，手法 3-2 である．手法 3-2 で生成したゲーム音楽はノイズが多く含まれて特徴の変換もできていない．これらの実験結果から，クラシック音楽からゲーム音楽への変換では，励起信号と CQT スペクトログラムを使った学習が効果的であり，データセット音声の高周波成分を除去することは学習に逆効果であることがわかった．

表 4.3: 実験 3 の生成音声のアンケート結果

実験手法	クラシック音楽からゲーム音楽	ゲーム音楽からクラシック音楽
既存手法	3	6
提案手法 3-1	12	12
提案手法 3-2	2	12
提案手法 3-3	13	0

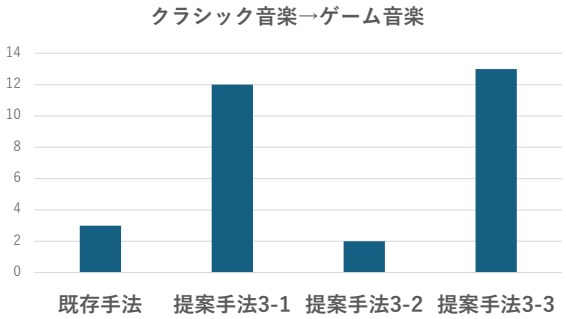


図 4.10: 実験 3 のアンケート結果 (クラシック音楽からゲーム音楽への変換)

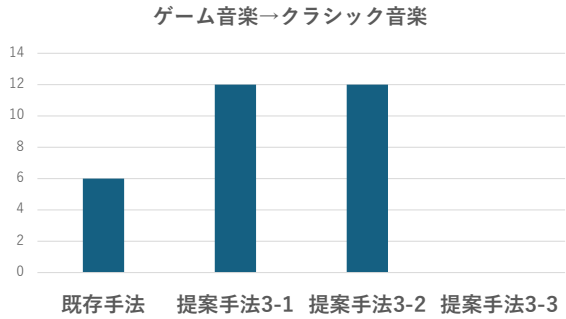


図 4.11: 実験 3 のアンケート結果 (ゲーム音楽からクラシック音楽への変換)

ゲーム音楽からクラシック音楽への変換では，手法 3-1 と手法 3-2 がベースモデルに比べて良い結果となった．既存手法で生成したクラシック音楽は，楽器に一貫性がなくメロディの音の高さによって異なる楽器が使用されている．例えば、入力ゲーム音楽の高い音

の部分はフルートのような高い管楽器に変換され，低い音はピアノのような楽器に変換される．これらの変換が不自然に行われるため，既存手法の生成音声の知覚品質が悪くなっているのではないかと考えられる．手法 3-1 では楽器の不自然な変換は少なくなり，生成音声はフルートのような管楽器に統一されている．手法 3-2 も同様に不自然な楽器変換が少なくなり，生成音声はピアノやフルートなどの複数の楽器を再現している．手法 3-3 に使用したスタイル変換器は，ゲーム音楽からクラシック音楽への変換の学習に失敗したため，生成音声はメロディ情報が完全に失われた音声となっている．ゲーム音楽からクラシック音楽への変換では，提案する 2 つの手法がベースモデルを上回る結果となったが，ゲーム音楽への変換ほど顕著な差とはならなかった．生成された音声を比較しても，提案手法が生成したクラシック音楽はクラシック音楽の楽器などの特徴を再現することができなかった．手法 3-2，手法 3-3 の結果を比較すると，クラシック音楽からゲーム音楽への変換とは対照的に，手法 3-2 が良い結果となり，手法 3-3 は変換に失敗した．このことから，変換タスクによって音声の前処理や入力に使用する音響特徴量の最適な組み合わせが異なるのではないかと考えられる．

第5章　まとめ

本研究では，機械学習を用いてゲーム音楽とクラシック音楽を相互に変換する音楽のスタイル変換手法を提案した．本研究では MaskCycleGAN-VC と ParallelWaveGAN を組み合わせた声質変換手法をベースモデルとして採用した．2 つのベースモデルは人間のスピーチ音声を分析，生成するように設計されているため，本研究で対象とするゲーム音楽，クラシック音楽を分析するために MBD と CQTLoss を組み合わせた Neural Vocoder と，複数の音響特徴量を入力してスタイル変換を行うスタイル変換器を提案した．

Neural Vocoder の実験から，ベースモデルの識別機を Multi-Band Discriminator に変更することで，クラシック音楽の生成音声の品質が向上した．また，ベースモデルの識別機を Multi-Band Discriminator に変更し，CQT Loss 組み合わせて学習を行うことで，ゲーム音楽の生成音声の品質が向上した．

スタイル変換器と Neural Vocoder を組み合わせた音楽スタイル変換の実験では，アンケート調査の結果から，提案手法を組み合わせることでクラシック音楽からゲーム音楽，ゲーム音楽からクラシック音楽の両方の変換においてベースモデルを上回ることができた．

一方で，提案モデルには課題があり，スタイル変換器はクラシック音楽からゲーム音楽，ゲーム音楽からクラシック音楽変換タスクによって最適なモデルが異なっていた．今後の課題として，スタイル変換器の 2 つの生成器について，変換タスクによって入力特徴量を変えたり学習方法を変更するなどの改良を行い，両方向の変換タスクを両立できる学習方法を模索し，生成音声の品質向上を目指す．

謝辞

本研究においては，ご多忙の中，熱心に指導をしてくださった中野浩嗣教授，伊藤靖朗教授に深く感謝致します．研究を進めるにあたり，とても貴重なご意見をいただきました．

また，研究しやすい環境を作ってくださった研究室の皆様にも，心より感謝しています．

参考文献

[1] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved Cyclegan-based Non-parallel Voice Conversion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6820–6824.

[2] ——, "MaskCycleGAN-VC: Learning Non-Parallel Voice Conversion with Filling in Frames," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5919–5923.

[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *9th ISCA Speech Synthesis Workshop*, 2016.

[4] R. Yamamoto, E. Song, , and J. M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *in Proc. ICASSP*, 2020.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *in Advances in neural information processing systems*, pp. 2672–2680, 2014.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A.Efros, "Unpaired Image-To-Image Translation using Cycle-Consistent Adversarial Networks," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[7] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/6804c9bca0a615bdb9374d00a9fcba59-Paper.pdf

[8] J. Kong, J. Kim, and J. Bae, "Hifi-Gan: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural*

Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

[9] F. Chen, R. Huang, C. Cui, Y. Ren, J. Liu, Z. Zhao, N. J. Yuan, and B. Huai, "SingGAN: Generative Adversarial Network For High-Fidelity Singing Voice Generation," *Proceedings of the 30th ACM International Conference on Multimedia*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238857196

[10] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 3918–3926. [Online]. Available: https://proceedings.mlr.press/v80/oord18a.html

[11] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis," *ArXiv*, vol. abs/2009.01776, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221470340

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-To-Image Translation With Conditional Adversarial Networks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[13] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley, "LakhNES: Improving multi-instrumental music generation with cross-domain pre-training," 2019.

[14] J. Thickstun, Z. Harchaoui, , and S. Kakade, "Learning features of music from scratch," *in International Conference on Learning Representations (ICLR)*, 2017.