深層学習を利用した歴史的演奏音源の回復
Restoration of historical audio considering using deep neural network

喜多修也 (Naoya Kita)*

法政大学 情報科学部 ディジタルメディア学科学科
naoya.kita.4u@stu.hosei.ac.jp

Abstract

This study proposes a method for historical recordings restoration using deep neural network. To simulate historical audio with narrow dynamics and bandwidth, and noise, trained data were pre-processed clean speech by limiting bit depth, adding noise, and applying a bandpass filter. After denoising and bandwidth extension using the audio super-resolution network AERO, attack-focused super-resolution was performed with reference to SREdgeNet. The results of the evaluation by estimating the signal-to-noise ratio using LSD and PSNR for the test data and WADA SNR for historical recordings showed that the proposed method was superior to the conventional method, with results of 1.18, 1.07 and 17.92 for the conventional method are 0.84, 0.38 and 319.60 for the proposed method, respectively.

1　はじめに

近年、レコード会社や博物館といったアーカイブを残す会社が、アナログレコードに記録された歴史的音源をデジタル化して保存する動きがある [1]。ここで、歴史的音声とは 20 世紀初頭前後の時期に記録された音源を指す。こうした歴史的音源の記録方法は、収録したい音をホーンで収音し、その音の振動をカッターヘッドに伝えレコードを掘る音響録音という手法と、カーボンマイクで収音し、カッターヘッドを振動させて掘る電気録音という手法で録音された。

しかし、これら音声は、現代の人間が聞いている CD 音質の音声と比較して音質が悪い [2]。当時のレコードは材質がレコード上でまばらに存在しており録音、再生の際に針が材質に引っかかり、秒間平均で 2000 回のクリックノイズが音声とともに記録、発音される。また、収音媒体の周波数特性により記録されている音声の帯域幅が狭い。音響録音では 160Hz から 2000Hz、電気録音では 20Hz から 4000Hz である。高周波数成分を記録できないことにより、音の発音する瞬間であるアタックという部分がぼやけることでこもって聞こえる。アタックは、楽器の識別をする上で重要な要素であり、微細な演奏のニュアンスを引き出す上で重要なため、こもってしまうと聞き手が受け取る演奏や音色の印象が異なってしまう [3]。低帯域が記録できないことで、楽曲中でリズムを支えるパートが小さくなり音の厚みが弱く聞こえる。また、記録できるダイナミックレンジが狭く、大きい音は記録可能な一方で小さい音を記録できず、強弱をつけた奏法も記録できない。リバーブをカットした音場で録音していることから、リバーブを記録していないといった問題がある。

本研究では、過去に録音された演奏音源にフォーカスして、上記のような歴史的音源のモデル化を提案し、教師あり深層学習を利用して歴史的音源を、デノイズされ、20kHz を超える帯域の幅拡張、現代で聞く HiFi 音声と同等になるように回復する。歴史的音源の回復は、録音当初にどのような演奏手法で演奏していたのかを理解できるようになるほか、文化的に失われた音声を再復興を見込め、歴史的音声のもつ歴史的価値観向上につながる。

2　関連研究

ここでは、スペクトルベースで行う帯域幅拡張のネットワークを利用した音源回復における問題と、歴史的音源のモデル化について言及する。

2.1　歴史的音源の回復

アタックの回復において、位相の考慮は絶対である。アタックのデータは一瞬であるため、位相を考慮しないでスペクトルベースの帯域幅拡張をすると正解データと位相がずれ、生成したアタック箇所の波形が損失計算の過程で一致せず、学習が進むにつれて正しい表現から遠くなる。

Mandel らは、低解像な音声を高周波数成分まで広げる、音声帯域幅拡張のネットワーク『AERO』[4] を提案した。入力された低サンプリングレートの音声を周波数領域で超解像する本ネットワークは、複素フーリエ変換によるスペクトル変換、正解音声と回復した音声との間で、実部と虚部を利用した STFT 損失（Short-Time Fourier Transform Loss）で位相のずれない帯域幅拡張を実現した。

図 1 に、歴史的音源の回復結果を示す。

上記図のように、位相のずれを考慮してもアタックの回復は上手く行えていない。理由は損失計算の方法とそれによる学習

* 指導教員：伊藤克亘 教授

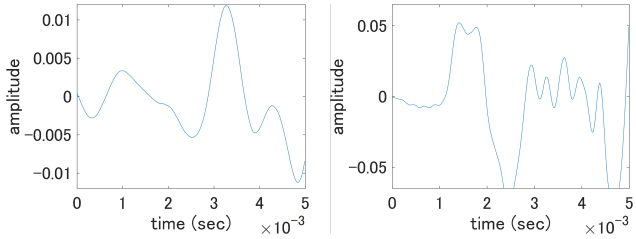

図 1　回復した歴史的音声と HiFi 音声のアタック箇所の波形

過程が起因する。AERO で利用している STFT 損失は、音声全体の正解音声と回復した音声の平均二乗誤差で損失計算を行っている。しかしこれを損失として最小化する方向へ学習するにつれ、音声データ全体が平滑化される [5]。それにより、アタックが考慮できず、これも回復過程において一瞬であるアタックが潰れ、こもって聞こえる要因となる。

アタック箇所以外においては、帯域幅拡張をする上で、中音域が過度にデノイズされてしまうという問題もある。

図 2 に左上から右に歴史的音声、CD 音質のピアノの対数振幅スペクトル、左下から右に歴史的音声を従来手法でデノイズ、帯域幅拡張した対数振幅スペクトルを示す。

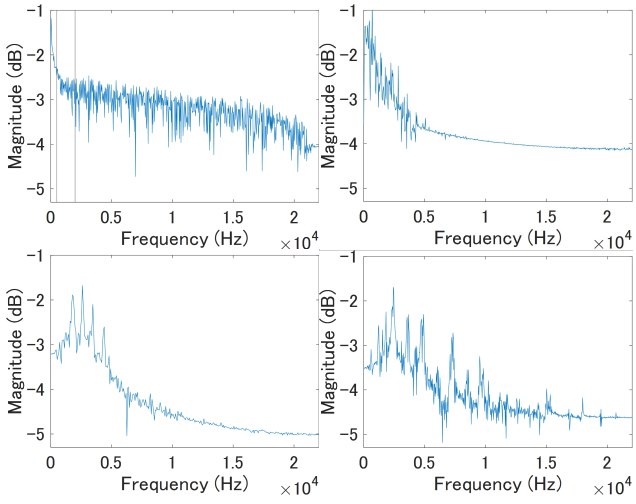

図 2　歴史的音声と高音質音声の対数振幅スペクトル

スペクトルを観察すると、各帯域の回復は、現在の技術で記録できる高音質音源のように帯域幅拡張を行えない。電気録音、音響録音はそれぞれ 2k から 6kHz、2k から 4kHz 程度までしか復元できない。また対数振幅スペクトルを観察すると、高音質音源には存在しない余分な倍音成分が現れ、効果的な帯域幅拡張が行えていない事が分かった。これは、過度なデノイズが原因であるほか、回復処理の工程順番も起因する。従来研究では、デノイズを行ってから帯域幅拡張を行っているが、これにより、帯域幅拡張をする際に必要な周波数成分がノイズとともに除去されてしまうことで起きている。

2.2　歴史的音源のシミュレーション方法

従来手法での歴史的音声のシミュレーションでは、実際の歴史的音声の再現が上手く行えない。現行の研究では、高音質な正解データに対して 3kHz のローパスフィルター、ホワイトノイズを付加することで歴史的音源のシミュレーションを行っている [6]。しかし、上記の手法では、はじめにで述べた、ノイズの種類や帯域幅の狭さ、ダイナミクスの狭さといった歴史的音声の音響特徴を再現しきれない。特に帯域の狭さのシミュレートについては、3000Hz のローパスフィルターだけでは足りず、低い音がはっきり聞こえる。またダイナミクスの制限は考慮されず、音の強弱の情報を残してしまっている。

3　提案手法

本研究では、歴史的音声のモデル化と、深層学習を利用した歴史的音声のアタック箇所の回復方法を提案する。歴史的音声に対する帯域幅拡張、デノイズ、アタック強調のすべてを従来研究の AERO のネットワークの一部を変更して行った。

AERO は、4 層のエンコーダとデコーダから構成される U-Net アーキテクチャを採用し、音声の帯域幅拡張を行うモデルである。本ネットワークは、入力音声を時間・周波数領域の特徴空間へ変換し、ノイズ除去と帯域幅拡張を段階的に行い、高品質な広帯域音声の再構築を目指す。

エンコーダでは、まず周波数変換ブロック (FTB) を用いた畳み込み処理を行う。これは時間方向への畳み込みによ特徴量抽出し、この情報からデノイズを実現する。その後、周波数領域への畳み込み、逆畳み込みにより、スペクトルの帯域幅拡張をする。

学習時の損失関数は、STFT 損失と特徴損失 (Feature Loss) の重み付き和を利用する。特徴損失は、高次の音響特徴を考慮し、音声の知覚的な品質向上を目的とする。

図 3 に、本研究用に変更した AERO のネットワークと各工程での処理を示す。MSD は STFT 損失、特徴損失の計算を示し、加えてアタックに関する計算の損失も追加した。

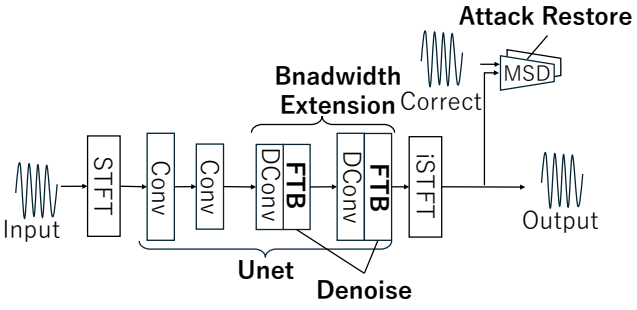

図 3　AERO のネットワークと処理

従来研究より、デノイズしてから帯域幅拡張を行うとノイズと一緒に、帯域幅拡張の分析に必要な低音域も除去されてしまうため、FTB をデコーダの最終部分に配置し、帯域幅拡張してからデノイズを行った。

3.1　歴史的音声のシミュレーション

本研究では、歴史的音源の音響特性を忠実に再現するために、学習データに対して劣化のシミュレーションを施す。このシミュレーションは、当時の録音技術、レコードの製造技術の制約による音質の変化を再現し、それらの劣化データを利用してネットワークによる教師あり学習を通して補正することを目的に行う。

まず、カーボンマイク録音の再現を目的として、マイクの非線形特性と周波数特性をモデル化する。カーボンマイクは入力音の振幅に対して線形な応答を示さず、大きな音量では歪みが生じる。これを再現するため、クリーンな音声の振幅に対して非線形関数を適用し歪みを再現する。特に強い音の部分で発生する音質の劣化を再現したと同時に、負の振幅成分が 0 に近づくようなモデル化を実現し、その補正方法を学習できるようにする。

次に、低域および高域の周波数成分が強く制限され、中域のみが残るような音質となるマイクの周波数特性を再現するため、400Hz から 3000Hz の通過帯域を持つチェビシェフ 2 型フィルタを適用する [8]。この処理によって、録音された音声が持つ本来の帯域幅を狭め、こもった音質と低帯域の成分の少ない音質を再現し、歴史的音源に近い特性を持つデータを生成できる。

次にクリックノイズを付加する。この時、クリックノイズの振幅は録音したい信号の振幅が大きいほど、対応するクリックノイズも大きくなる傾向があることを確認した。この特性を反映させるため、音声の振幅に基づいてクリックノイズの大きさを調整し、実際の録音環境により近いノイズ特性を再現する。

これらの処理を施した劣化音声を学習データとして用い、HiFi なモノラル音楽データを正解音声として教師あり学習を行う。これにより、ネットワークは歴史的音源の特徴的な劣化を補正し、より広帯域でクリアな音声への変換が期待される。また、劣化のシミュレーションが現実の歴史的音源に近いほど客観的に評価でき、今回のシミュレーションは単なるデータ拡張ではなく、モデルの適用可能性を高めるための重要なステップとなる。これらの劣化シミュレーションを精密に設計することで、歴史的音源の復元精度を向上させることを目指す。

式 1 にカーボンマイクの非線形特性をモデル化した式 $Q(x)$ を示す。係数 α が 0.4 の時、歴史的音声の特性に近づけられる [7]。

$$Q(x) = (1-\alpha)x + \alpha x^2 + \alpha^2 x^3 + \alpha^3 x^4 + \alpha^4 x^5 \qquad (1)$$

図 4 の左に従来手法のローパスフィルタの概形（青）と、提案手法で利用した、カーボンマイクの周波数特性をモデル化したチェビシェフ 2 型フィルタの概形（赤）、及び非線形特性を示す。ここで、青線で描かれた概形は α が 0.4 の時のものである。

表 1 に、無音区間、有音区間のノイズパワーを示す。歴史的音声のような、SN 比を測るために必要なクリーンな音声が無い状態で SN 比を比較するために、ブラインドな音声に対して

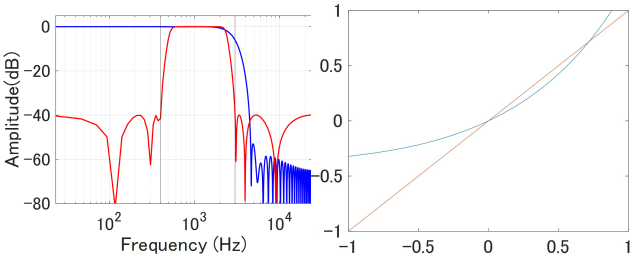

図 4　フィルタと非線形特性

SN 比を計算できる、WADA SNR というアルゴリズムを利用した [8]。

SNR は値が小さいほど、信号に対してノイズが大きいことを表すことから、無音区間のノイズが、有音区間のノイズよりも小さいことがわかる。

表 1　無音、有音区間でのクリックノイズのレベル (dB)

無音区間 (dB)	有音区間 (dB)
3.05	5.01

3.2　アタックの回復

アタックは、音が発音されてから 20 ミリ秒以内に存在する一方で、それ以降の音は周期成分のあるサステインやディケイにあたる。そのため、本研究ではこの 20 ミリ秒の区間をアタックとした。

本研究では、従来の AERO で使用していた音声全体の STFT 損失に加えて、アタック部分に着目した損失計算を行うことで、より精密なアタック回復を目指す。損失計算は従来手法に加えて 2 つを利用する。

1 つ目は、アタックのタイミングが正しく生成されているかを評価する損失である。アタックのタイミングのずれを測定するために、まず正解音声と生成音声の両方に対してアタック検知を行い、その検知された時間の差を計算する [9]。アタックの検知には、隣接するフレーム間における各周波数成分の振幅スペクトルの変化量を計算するスペクトラルフラックスオンセットを用いる。

以下式 2 にフラックスオンセットを記す。

$$SF(n) = \sum_{\omega=0}^{\frac{N}{2}+1} H\left(|X_\omega(n)| - |X_\omega(n-1)|\right)^2 \qquad (2)$$

アタック部分が正しい音質で生成されているかを評価する損失である。アタックの音質評価には、正解音声の発音開始から 20 ミリ秒の区間に対して、異なる FFT サイズ（512, 1024, 4096）を用いたスペクトル類似度を計算し、その類似度を損失として評価する。

t]

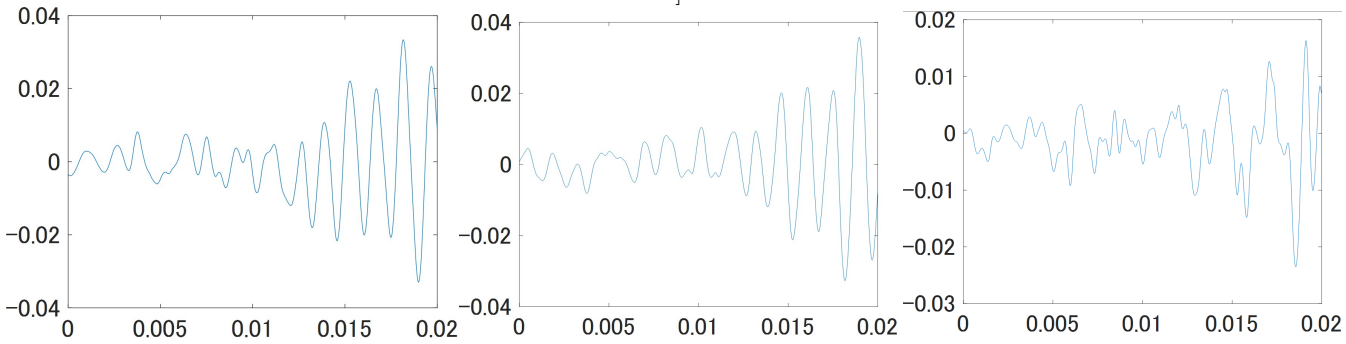

図 5　テストデータのアタック箇所の波形 (正解、提案手法、従来手法)

4　評価手法と評価結果

4.1　学習データ

本研究では歴史的音声の再現を行うため、リバーブの無いドライなデータセットを探す必要があった。しかし、オーケストラのようなアンサンブル形式の音声データセットでは、ドライな音声で記録されているものが少なく、そのような音声を探す事は非現実的である。そのため、本研究では歴史的音声のシミュレートを単一の楽器で行う。

adl-piano データセットは、プロのピアノ独奏のデータセットである [10]。ジャズやブルース、ゲーム音楽等、さまざまなジャンルの音楽を midi データで記録したものになり、総数約 100 時間存在する。このデータセットをドライなピアノ音色を利用して音声ファイルに変換し、学習データとして扱う。特にデータのうち無作為に抽出した 80% を学習データ、10% を検証データ、10% をテストデータに利用する。データ長を揃えるため、各データを 1 分ずつにトリミングした。

また、実際の歴史的音声のデータは、The Great 78 Project を利用する [11]。これは、1900 年から 1960 年までに発行された、約 20 万曲の SP レコードをデジタル化したアーカイブだ。ここには、クラシック音楽、童謡、当時の pops が収録されており、サンプリングレートは 96kHz で保存されている。この音声を利用して、本ネットワークの精度を主観的に評価する。

4.2　評価方法

4.2.1　客観評価

本研究では、2 つの客観評価によってテストデータの再現度合いを評価する。

回復した音声とクリアな音声の間の対数スペクトルの類似性を測るため、LSD(Log-Spectral-Distance) を利用する。この値が小さいほど、振幅スペクトルが近い事を意味しており、これは正解データと近いレベルで歴史的音源を回復できた事を意味する。これを音声全体、アタック箇所に対して行う。

式 3 は LSD を示す。ここで、Y は正解の音声、$\hat{Y}$ は歴史的音声を帯域幅拡張した音声を示す。

$$\mathrm{LSD} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{K}\sum_{k=1}^{T}\left(|Y_{t,k}|^2 - \left|\hat{Y}_{t,k}\right|^2\right)} \tag{3}$$

また、テストデータ、および歴史的音声の SN 比がどれほど改善されたかを観察するために、WADA SNR[8] を利用した。

4.2.2　主観評価

本研究では、回復した音声が CD 音質レベルで音の立ち上がりをはっきり聞き取る事ができるか、低帯域の音声が聞こえるか、デノイズができているかの観点で主観評価実験を行った。

聞いた音声について、1 を最低、5 を最高と評価する手法の、MOS(Mean Opinion Score) を利用する。テストデータ、及び歴史的音源の両方で評価し、被験者対象は、大学生 15 人で行った。

4.3　評価結果

図 5 に正解のテストデータのアタック箇所の波形、テストデータを提案手法で回復したアタック箇所の波形、従来手法で回復したアタック箇所の波形を示す。

表 1 に、回復した音源について、音声全体の LSD と、アタック箇所だけに着目した LSD による客観評価を行った結果を示す。ここで、AERO は従来手法、AERO(+A) は提案手法を指す。従来手法では 1.18、1.07、提案手法は 0.84、0.38 となり、提案手法の方が低い値を示し、比較して精度よく音声全体の帯域幅拡張、およびアタックの回復が行えていることが分かった。

表 2 に歴史的音源に対して WADA SNR を利用して SNR を予測した結果を表す。従来手法では 17.92 であるのに対し提案手法は 19.60 となり、SNR の値が元の歴史的音声と比較して改善されている事がわかった。

表 2　各手法の LSD, アタック箇所の LSD の評価

	LSD ↓	LSD(A) ↓
AERO(test)	1.18	1.07
AERO(+A)(test)	0.84	0.38

表 3 に、CD 音質相当の音声、歴史的音源、歴史的音源を従

表 3　歴史的音声に対して回復させた音の WADA SNR ↑

historical	AERO	AERO(+A)
7.05	17.92	19.60

来手法で回復させた音声と提案手法で回復させた音声の MOS の結果を示す。CD は CD 音質の音源、Hist は歴史的音源、Test は歴史的音源をシミュレートしたテストデータを表す。Attack はアタック箇所が鮮明に聞こえたか、Low frequency は低い帯域まで音が聞こえたか、clean はノイズがなく聞こえているかの項目を表す。結果では、テストデータに対するアタック回復は、従来手法 3.42、提案手法 3.71、となり比較してよい結果を示した一方、低帯域の回復、デノイズについては、従来手法が 3.32、2.62 となった一方、提案手法では 2.62、2.03 となり、比較して悪い値となった。歴史的音声の回復については、アタック回復、低帯域の回復、低帯域の回復は従来手法では 1.92、2.42、1.85、提案手法では 1.60、2.21、1.33 となり、手法を下回っていた事が分かった。

表 4　各手法の MOS の結果

	Attack	Low frequency Band	Clear
CD	4.21	4.70	4.50
Hist	3.46	3.32	1.33
Test	3.63	2.82	1
Test(AERO)	3.42	3.32	2.62
Test(AERO(+A)	3.71	2.62	2.03
Hist(AERO)	1.92	2.42	1.85
Hist(AERO(+A))	1.60	2.21	1.33

提案手法は、従来手法と比較して音声全体、アタック箇所ともに LSD が低くなり優位な結果を示した。従来手法よりも提案手法のほうが音声全体、アタック箇所の両方でスペクトルの回復が行えた事を示している。実際に波形を確認すると、提案手法と比較して急激かつ非周期的な振幅を確認でき、アタックの再現に成功している事が確認できた。WADA SNR の値も提案手法が従来手法よりも約 2dB 高い値を示し、より回復できたことを示せた。

しかし、MOS を確認すると客観評価と対比して、提案手法が従来手法よりもテストデータと歴史的音声ともに、有意な結果を示していない。特にノイズが無くクリアに聞こえるかという調査ではアタック、低音域の回復の項目と比較して従来手法よりも低い値を示した。これは、デノイズの部分でまだ課題があることを表している。生成した音声を聞き、スペクトログラムを確認したところ無音区間におけるノイズの除去は成功している一方、有音区間でのデノイズは改善の必要がある。特に、低い帯域に関する調査では 2 未満の値をとっていないことから、相対的に中音域でのデノイズがうまくいっていないものとした。

以下図 6 に FTB の概略図を示す。

まず、有音区間でのデノイズが上手くいかなかった原因は、Unet 中のデコーダ層にある FTB ブロックの構造にあるとした。FTB でデノイズが行える理由は、時間方向の畳み込みと、時間-周波数領域の注意機構によるものにある。時間方向の畳み込みをすることで、定常的に表れているノイズの除去を行う事が出来る。しかし、歴史的音声のような、信号の振幅に比例して大きくなるノイズが混在している場合、このような振幅成分のノイズは信号と同様に非定常的であることから、非定常的なノイズの除去が上手く行えないとした。

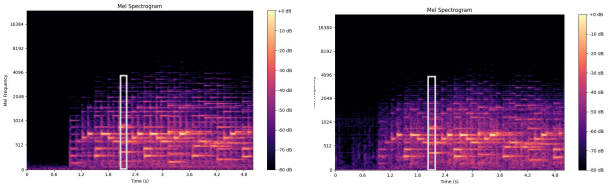

図 6　スペクトログラム

時間領域の特徴抽出は時間方向のみの畳み込みで実現しているが、周波数成分についての情報取得を全結合層によるものでなく、周波数方向への畳み込みを行うことで、音声とノイズのスペクトル情報を分析でき、デノイズ能力が向上するとした。

また、デノイズが上手くいっていない理由は他にも、学習段階でノイズ特徴をうまくとらえられていないからだとした。この原因は Unet の層の数にあるものとした。Unet はエンコーダで入力された情報の特徴抽出をし、その情報をもとにデコーダで正解の情報に近づけるという構造をしている。この層の数が多いほどより詳しく特徴抽出を行える。現在の実験では、層の数を 4 層にして学習を行っているが、本実験においてはノイズ特徴を抽出するには不十分であった可能性がある。そのため、Unet の層を深くすることで、より強く特徴抽出が行えるため、層を 5 つにして学習を行うなどの工夫が必要だとした。

5　おわりに

本研究では、深層学習を用いた歴史的音源の回復手法、および歴史的音源のモデル化を提案した。前者では従来の歴史的音源の回復手法に加えて、アタックのタイミングのずれを考慮する損失と、アタック部分の音質を評価する損失を組み合わせることで、アタック回復を目指し提案した。後者では、カーボンマイクの非線形特性や帯域制限の再現、クリックノイズの追加を行うことで、より現実的な学習データを生成する手法を提案した。

実験の結果、提案手法を用いた音声復元は客観評価においては従来手法より優位な結果を示した一方、主観評価では従来手法よりも劣る評価を受けた。特にデノイズ能力が劣っており、その原因として、本研究で用いた周波数変換ブロック（FTB）が、定常的なノイズの除去には有効である一方で、音声の振幅に比例して変化するノイズの除去には弱いことをあげた。一方、デノイズが弱くなった場合、元の信号の成分も残ることから、トレードオフな関係でもある。

今後は、これらの結果を受けて、周波数方向への畳み込みによって周波数情報についても特徴抽出することで非定常的なノイズの除去も実現する。

また、主観評価には MOS（Mean Opinion Score）を使用したが、被験者数が少なかったため、今後、40 人に拡張した再実験を実施する予定である。

本実験のデータセットは独奏のピアノ演奏であったため、データセットを拡張して、オーケストラといった、ピアノ以外の楽器による演奏音源についても回復することを目指す。

参考文献

[1] Nový Fonograf. et al. Digitalizuje, eviduje a chrání zvukové nahrávky z historických nosičů, konkrétně z fonografických válečků a gramofonových desek 2018-04-12 [Online]. Available: https://novyfonograf.cz/

[2] T. J. Park et al. Digital Audio Restoration–A Statistical

Model Based Approach The International Series in Engineering and Computer Science, vol 437. Springer, Boston, MA.

[3] K. Siedenburg et al. Specifying the perceptual relevance of onset transients for musical instrument identification The Journal of the Acoustical Society of America, February 2019

[4] M. Mandel et al AERO: Audio super resolution in the spectral domain Proc. ICASSP, Rhodes, 2023, pp. 1–5.

[5] Kim, K. et al SREdgenet: Edge enhanced single image super resolution using dense edge detection network and feature merge network Computer Vision and Pattern Recognition (cs.CV)

[6] E. Moliner et al A two-stage u-net for high-fidelity denoising of historical recordings ICASSP, 2022

[7] S. Oksanen et a. Modeling of the carbon microphone nonlinearity for a vintage telephone sound effect Conference on Digital Audio Effects (DAFx-11), pp. 27-30

[8] W. Hsu et al Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation in ASR, pp. 16–23, 2017b.

[9] Shier, J et al. Differentiable modelling of percussive audio with transient and spectral synthesis The Proceedings of Forum Acusticum, Sep 2023

[10] L. Ferreira et al Computer-generated music for tabletop role-playing games , AIIDE-20, pp59–65.

[11] The great 78 project. Feb. 11, 2023. [Online]. Available: https://great78.archive.org