7ZB-06

# コンサートホール音像再現のための Diffusion モデルを用いたバイノーラル音声変換 *

北村健太郎 †,　伊藤 克亘 †,

## 1　まえがき

コンサートホールでの録音は会場上部に吊り下げられているマイクでされている。その録音機は HRTF(頭部伝達関数) を意識された録音がされておらず会場の臨場感を録音することができない。バイノーラル録音できる機材は市販されているものの、値段が高く手に届かないことやバイノーラルマイクの見た目がコンサートホールの景観を損なうなどの理由で使われていない。その問題を解決するために、研究では空間音響と HRTF を音源と音源の位置を用いて学習したモデルを使い、通常のステレオ音源をバイノーラル音源へと変換する。この研究により音楽配信サービスなどのストリーミング音声の臨場感の向上や、バイノーラルオーディオ作成のコストが格段に減ることを期待する。

## 2　関連研究

### 2.1　バイノーラル音声の合成

線形の信号処理でのバイノーラル音声の生成は RIR(室内インパルス応答) と HRTF を組み合わせて両耳の音声を合成する。部屋のインパルス計測や耳の形状の測定にコストがかかる。そのため、線形の信号処理を用いた手法は汎用的な関数を利用することが多く、収録環境がデータセットによって大きく異なるため、最適とはいえない生成結果になる。

マルチタスク学習を用いたモノラル音声からバイノーラル音声を生成するための学習 [2] は、ステレオ音源の生成タスクと左右のチャンネルが反転された音源の分類タスクという 2 つのタスクから特徴量を抽出する。この学習モデルは、視覚と聴覚の入力から空間特徴を抽出し、左右の音声チャンネルを予測し、左右のチャンネルが反転しているかどうかを判定する。まず、ビデオフレームから ResNet を用いて視覚的特徴を抽出する。次に、視覚的特徴に基づく別々のサブネットワークを用いて、ステレオ音源の生成と左右のチャンネルが反転された音源の分類を行う。本学習法では、2 つのタスクの損失の重み付け和に基づいて全体の損失を最適化する。FAIR-Play データセットと YouTube-ASMR データセットでモデルを学習・評価した。この手法にはいくつか制限がある。まず、損失の重みを手動で調整しなければならない。重みが異なると、結果も異なる可能性がある。また、複数の音源が存在するシーンではバイノーラル音声の精度と方向の知覚品質が低下する。これは、コンサートホールにおける多数音源の生成には向いていない。

拡散確率モデルを使ったバイノーラル音声の合成モデルである BinauralGrad[1] は 2 段階のフレームワークを用いて音声を合成している。第 1 段階では、モノラル音声を入力し、1 チャンネルの拡散モデルによりバイノーラル音声とモノラル音声の共通している特徴量を生成し、第 2 段階では、2 チャンネルの拡散モデルによりバイノーラル音声を生成する。BinauralGrad は正確かつ高忠実度のバイノーラル音声を生成することができる。ダミーヘッドにバイノーラルマイクを取り付け録音し、音源と聞き手の位置も記録したデータセット [3] で学習・評価した。この手法は知覚品質や、生成音声の質は極めて高いが、残響成分がバイノーラル音声の生成過程で消えており、コンサートホールでの反響が消えるという問題がある。またこのモデルも、単一音源の生成しかできず、コンサートのような複数音源に対応はしていない。

## 3　提案手法

従来研究では、残響の生成や音像定位を実現する際の高音域の忠実な再現ができていない。残響の生成に関して従来の生成モデル [3] でノイズ除去の効果がある。生成音声モデルの評価方法を変えることで解決する。高音域の忠実な再現に関して、両耳時間差を計算をする際に加算を行うことにより、位相さによる干渉が起こる影響で楽器音の忠実な再現ができない。　これは、残響の生成と同様にモデルの評価方法を変えることで解決する。

本研究では、残響の生成と音像の配置を線形の信号処理、信号処理された音声を Diffusion モデルで明瞭化を行う。音声の明瞭化は BinauralGrad[3] をベースにモデルを作る。システムの全体像図 1 に示す。システムでは最初にモノラル音声に対して、両耳の到達時間差を考慮した音像配置を行う。ここで生成された音声は、ローパスフィルタがかかっている。そのため処理した音声と、音源を観測する位置情報をニューラルネットワークに入力し、音声の明瞭化とバイノーラル化を行う。音源を観測する位置情報は、音源に対しての相対的な三次元座標と姿勢を表すクオータニオンとする。入力された音声に対し、作成した変換モデルは高音域強調と残響の生成を行う。

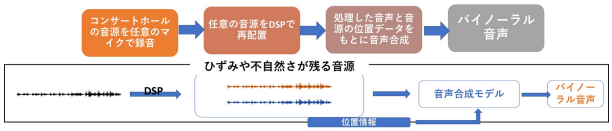


**図 1.** コンサートホールの音声を再現するシステムの全体像

### 3.1　損失関数

BinauralGrad[1] では差を用いて損失を計算していた。スペクトログラムや波形の差を損失関数とすると、音の周期的な構造を効率的に学習することができない。今回損失関数として Ben 氏らの研究 [4] を用いる。Ben 氏らは、従来の実空間上の音声館の差分は局所解には

* :Binaural Audio Transformation Using Diffusion Model for Concert Hall Sound Image Reproduction *Kentaro Kitamura (Hosei Univ.) et al.*

† 法政大学大学院 情報科学研究科

まりやすい問題 [4] を音声を複素数に変換し複素空間で勾配降下することで解決した。複数の正弦波のパラメータを同時に最適化をすることで生成波形の周波数構造を忠実に再現することを期待する。

### 3.2 Denoising Diffusion Probabillistic Model

拡散確率モデルは画像や超解像画像生成、テキスト画像生成、テキスト音声合成、音声強調などの様々なタスクで SOTA を実現している。特に、音声合成においては拡散確率モデルはスペクトログラムと波形の両方をモデル化することに強い能力を示している。本研究では、高解像度のバイノーラル音声波形を生成するために拡散確率モデルを使用する。

### 3.3 残響の生成

周波数信号に対し、周波数フィルターを掛け算して、逆 FFT をすれば、フィルタリング処理された信号を生成することができる。この手法の欠点として窓のつなぎ目で振幅、位相のずれが生じ、処理音声が劣化してしまう。今回はこの問題を解決するために、時系列上で処理をする。周波数応答関数を逆 FFT するとインパルス応答になる。このインパルス応答を時系列データと畳み込み積分することで、フィルターされる。今回は拍手をインパルス応答に見立て、その収録音を用いて畳み込みを行った。インパルス応答は様々な方法で測定することができるが、暗騒音のある音場では十分な SN 比を確保するには多数回の測定を行い、その応答を平均する必要がある。条件によっては十分な SN 比を取るには長時間の録音時間を要し、系の時不変性が問題となる。短時間に安定したインパルス応答の測定を行う方法として SweptSine 信号=SS 信号=Time Stretched Pulse TSP 信号を提案している。これはインパルスの位相を周波数の 2 乗に比例して変化させることにより、時間軸を引き延ばした信号である。一般的な SS 信号は

$$H(k) = \begin{cases} \exp(2\pi i J k^2/N^2) & 0 \le k \le \frac{N}{2} \\ \exp\{-2\pi i J (N-k)^2/N^2\} & \frac{N}{2}+1 \le k \le N-1 \\ J = N/2 \end{cases} \tag{1}$$

を逆フーリエ変換して得られる信号 $p(n)$ である。SS 信号に対する系の応答 $q(n)$ と SS 信号の時間軸反転をさせた $p(-n)$ を畳み込み演算するとその系のインパルス応答を求めることができる。

ただし、音場のインパルス応答が $N$ より長い場合には SS 信号を複数回提示して安定した後、その応答を収録し、畳み込みを行う必要がある。測定したインパルス応答に対し「誠にありがとうございます」と発話したドライ音声を畳み込む (図 2)。インパルス応答を

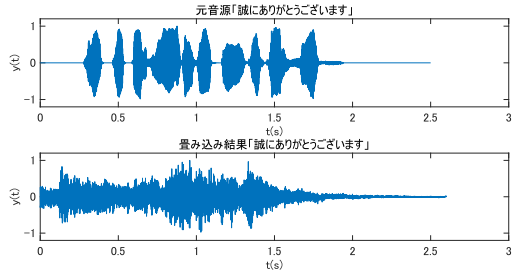


**図 2.** 上：「誠にありがとうございます」と発話しているドライ音声、下：上の音声に階段で収録したインパルス応答を畳み込んだもの

畳み込むことで、元音源にある音素間のポーズが無くなっていることが確認できる。音源に対してリバーブが生成されているが、実際に聞いた音にくらべ生成音はより残響感強く残っていることが確認できた。

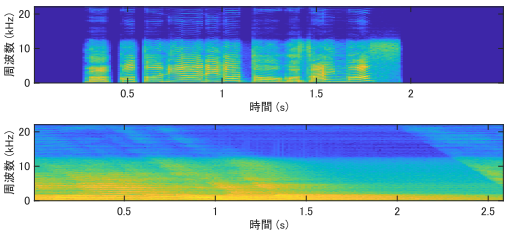


**図 3.** 「誠にありがとうございます」のスペクトログラム、上：元信号、下：フィルタリング処理された信号

フィルタリング処理された周波数成分を観察すると、全体的にパワーが増しており、雑音成分が増えている。また、解像度も下がっていることから、音色が曇ったように感じ取れる。

## 4 まとめ

従来のバイノーラル音声の生成方法は、音声の忠実な再現をすることができないことがわかった。これは、干渉の影響によるものであり、具体的には、両耳時間差をもちいた立体音響を生成する際に移送がずれることが原因だ。またインパルス応答を用いた残響音の生成は、SN 比が高くならず、良質な音声の生成をすることができないことがわかった。これは、残響や環境音を正確にモデル化することができないためである。

## 5 今後の予定

まず、残響が残るような音源を生成するために、BinauralGrad の評価に残響成分の比較を追加し、残響が残るか実験する。残響評価は室内インパルス応答から計算することが多い。BinauralGrad は生波波形を直接生成するモデルなので、波形やスペクトログラムから残響評価する方法を探す。

次に、楽曲向けに多チャンネルでの生成をするためのアーキテクチャを考案する。

## 参考文献

[1] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiang-Yang Li, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis, 2022.

[2] Sijia Li, Shiguang Liu, and Dinesh Manocha. Binaural audio generation via multi-task learning, 2021.

[3] Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Butler, Fernando de la Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021.

[4] Joseph Turian and Max Henry. I'm sorry for your loss: Spectrally-based audio distances are bad at pitch, 2020.