構造化状態空間シーケンスモデルを用いた位置情報の長距離依存関係を利用したバイノーラル音声合成

Kitamura, Kentaro / 北村, 健太郎

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
19
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030611

構造化状態空間シーケンスモデルを用いた位置情報の長距離依存関係を利用したバイノーラル音声合成

Binaural Audio Synthesis with the Structured State Space sequence model

北村健太郎 (Kentaro Kitamura)*

法政大学大学院 情報科学研究科
kentaro.kitamura.2p@stu.hosei.ac.jp

Abstract

The structured state-space sequence model (S4) is a recent innovation in sequence modeling that has shown excellent performance in handling long-range dependencies across a variety of tasks and modalities. In the field of speech processing, it has been found to be an alternative to the self-attention model in automatic speech recognition and in speech synthesis. In this study, a new model for synthesizing binaural speech is developed that represents the long relationship between mono speech using S4 and the latent state space between speaker and source location information. Each layer is conditioned with information common to both left and right sides of the speech, which is processed by location, binaural time difference, and pre-trained binaural speech. Compared to conventional methods, our model shows that speech synthesis is possible with similar quality. These results indicate that our model has the potential to extend the applicability of S4 in sequence modeling and into the domain of conditional speech synthesis.

1 Introduction

シーケンスのモデリングは、機械学習における主要な課題の 1 つであり、自然言語処理から音声認識まで、幅広い領域で中心的な役割を果たしている。これらの領域では、シーケンスのアイテム間の複雑な依存関係を捉えるモデルが必要とされる。近年、構造化状態空間シーケンスモデル (Structured State-space Sequence Model, S4)[3] という新しいアプローチが提案され、シーケンスモデリングにおける長距離依存性の取り扱いに優れた性能を示している。 S4 モデルは状態空間モデル (SSM) の理論的な利点と効率的な計算を兼ね備えている。これにより、SSM のパラメータ化を通じて、効率と有効性の適切なバランスをとれる。S4 は、音声の分野で特に大きな進歩を示しており、自動音声認識や音声合成（TTS）[10] などのタスクにおいて、自己アテンションモデルに取って代わるものとなっている [10] 。 一方、バイノーラル音声合成は、音の位置情報を再現するための重要な技術であり、バーチャルリアリティ、ゲーム、聴覚障害者のアクセシビリティ向上など、様々な応用が期待されている。しかし、これらのタスクでは、音の空間的な配置と時間的な進行を同時に扱う必要があり、そのためには強力なシーケンスモデリング機能が必要だ。そのため、これらの問題に対する効果的な解決策を見出すことは重要な課題であり、新たなイノベーションが待ち望まれている。 本研究では、この問題を解決するために、S4 モデルを用いた新しいバイノーラル音声合成モデルを提案する。特に、モノラル音声、話者、音源位置の情報を潜在状態空間間の関係として表現することで、これらの情報を統合する手法を開発する。これにより、音声合成の性能を大幅に向上させる。 この新手法の性能を評価するために、バイノーラル音声生成の最先端研究との比較実験を行う。具体的には、本モデルを既存の最先端技術である BinauralGrad [7] と比較し、いくつかのメトリクスに基づいて評価する。その結果、我々のモデルは大幅な改善を達成し、全ての評価指標において優れた性能を示した。これらの結果は、S4 モデルが複雑なシーケンスモデリングの課題に対処し、その応用範囲をさらに拡大する可能性を示している。

2 Related Work

バイノーラル音声の生成方法は主に 2 つある。一つ目はデジタル信号処理で二つ目はニューラルネットを使った手法である。デジタル信号処理は線形時不変システムとして頭部伝達関数、室内音響、周辺雑音をそれぞれモデル化している [9, 15, 14]。これらの線形システムは、数学的にモデル化されており比較的軽量であり、知覚的にもっともらしい結果が得られている。しかし、実際の音響伝達には線形時不変システムではモデル化できない非線形な部分があり、移動音源の生成などの動的な場合においてモデル化できていない。本論文では、より忠実な音源の生成をするために非線形部分を対応できるニューラルネットを使った手法を用いる。ニューラルネットを

* 指導教員：伊藤克亘 教授

使用する手法では、深層学習のモデルを用いて音響の特性を学習し、それに基づいてバイノーラル音声を生成 [11] する。このアプローチの大きな利点は、高度な非線形モデリング能力を有しているため、より複雑で現実に近い音響環境をシミュレートすることが可能であることである。WaveNet[13] のようなモデルは、元のオーディオサンプルから直接波形を生成し、その結果、非常に高品質で自然な音声を生成することができる。空間モデルを組み込んだニューラルネットでの音声合成手法 [2] では、オブジェクトの位置関係が空間の音響に由来しているとし、ビデオフレームを条件づけて音声合成している。しかしこの研究は音源と受聴者の位置の違いによる残響の効果や伝達の遅延をモデル化することはできていない。本研究では、ニューラルネットを用いたこれらの問題を解決したバイノーラル音声生成のための新しいアプローチを提案する。具体的には、音源と受聴者の位置関係を条件つけて音源の移動と音声の長期依存関係を学習し音声合成している。さらに、非線形な音響特性をより正確にモデル化し、よりリアルなバイノーラル音声生成を実現することを目指している。

2.1 Binaural Speech Synthesis Technology without Neural Networks

両耳音声合成の目的は、モノラル音声を両耳音声に変換することである。両耳音声合成の目的は、モノラル音声を両耳音声に変換することである。バイノーラル音声とは、音圧や時間差といった両耳への入力の違いを模倣することで、3 次元空間の体験や現実の音の方向や距離を再現する音響技術である。通常、バイノーラル音声の収録には、人間の耳の形状を模したダミーヘッドマイクロホンが用いられ、耳の形状によって生じる複雑な反響を再現する。音源が媒質を通って耳に届くとき、拡散、残響、反射などの空間的効果を受ける。 室内インパルス応答 (RIR) を使った研究では、部屋の材質、温度、媒質の違いによる音の伝わり方を再現するためにフィルタを使う [8]。また、頭部伝達関数（HRTF）を使った研究では、頭や耳の形状による音の反射や回折を表現することで、音の指向性を再現する [1]。 そのため、デジタル信号処理 (DSP) の手法では、一般的に様々な関数を使用するが、それぞれのデータセットが録音環境に特化され、音声の時不変性より最適な生成結果を得ることが難しいという課題がある。本論文では、モノラル音声と音源と聴取者の位置情報のみを用いたバイノーラル音声合成法を提案する。

2.2 Binaural Speech Synthesis Technology with Neural Networks

ニューラルネットワークを用いたバイノーラル音声合成法では、単純な線形処理では耳の形状による回折や反射の再現が困難な HRTF の再現が可能である。バイノーラル音声合成のための 2 つのモデル（Temporal ConvNet[11]）を用いて、HRTF による室内残響や音声の変化を再現する。バイノーラルオーディオを合成するために 2 つのモデルが使用される。1 つ目はソースの物理的特性とリスナーの両耳へのワープを学習し、2 つ目は部屋の残響と HRTF を学習するネットワークで構成される。BinauralGrad [7] は、拡散モデルと線形処理を組み合わせた 2 段階のバイノーラルオーディオ合成手法である。第 1 段階は、拡散モデルを使用して、両側のモノラルオーディオからバイノーラルオーディオの共通部分を生成する。第 2 段階は、第 1 段階の生成を基に、特徴的で忠実度の高いバイノーラルオーディオを生成する。本論文では、より忠実なバイノーラル音声を生成することを目的として、潜在状態空間におけるモノラル音声、話者、音源位置の関係を表現するために S4 モデルを用いる。

3 Background

バイノーラル音声のような時系列データは音源と受聴者の位置関係や残響が動的に変化することから長期依存関係がある。音声において長期依存関係をモデリングした WaveNet[13] は RNN[12] のように再帰的な構造を模倣した Dilated Causal Convolutions を取り入れており、畳み込みを使い時系列データを学習している。Dilated Causal Convolutions を使うことで、RNN より学習時間を短縮し、音声の波形を直接生成することに成功し、長期依存関係を効率的に学習することができた。本研究では、音声と位置情報の長距離依存関係に着目し、バイノーラル音声を生成する。モノラル音声からのバイノーラル音声のニューラル合成は、HRTF を暗黙的に模倣するためにコンボリューションを使用する ConvNet[11] がある。この手法では HRTF を暗黙的に学習しているため、音声の長期依存関係を適切にモデリングしていない。BinauralGrad は、Denoising Diffusion Probabilistic Models[5] をベースに作られた DiffWave[6] 使用して、高品質のバイノーラル音声を生成する。HMM を基礎概念にしていることから、音声波形の確率的な関係を学習している。この手法では音声波形は直接生成され、生成の過程でまず低周波情報が徐々に生成され、次の推論ステップで細部が反復的に生成される。これは、正確で忠実度の高い音声波形生成に適している。しかしこの手法では、音声波形の確率的な前後関係を学習しているところから、バイノーラル音声の確率的モデルであり、長距離依存関係の直接的な学習を行っていない。S4 を使用する現在の音声合成は、長距離依存関係をモデル化するための原理的なアプローチを取り入れている。S4 は、実際に耳で聞く音を原理的に模倣するために使用される。物体から放出された音波は媒質（自由空気）を伝搬するが、媒質中では音は拡散、残響し、伝搬中に壁などの物理的物体の影響を受ける。このような音の物理現象の特徴を、入力されたモノラル情報と位置情報から捉えることで、高音質な音を生成することを期待する。 本章では長距離依存関係のカギとなる SSM、S4 モデルについて説明し、次章では我々のアーキテクチャについて紹介する。

3.1 State Space Models

状態空間モデルは以下の簡単な式で定義される。この式は、1 次元の入力信号 $u(t)$ を N 次元の潜在状態 $x(t)$ に写像し、そ

れを 1 次元の出力信号 $y(t)$ に投影する。

$$\begin{aligned} x'(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{1}$$

ここで、SSM はディープシーケンスモデル内のブラックボックス表現として使用される。連続時間状態空間モデルを離散化するために、バイリニア法（2 次変換法）を用いる。これは状態行列 A をステップサイズ Δ で近似行列 $\overline{A}$ に変換する。これは入力の分解能を表す。離散状態空間モデルは次のようになる：

$$\overline{A} = (I - \frac{\Delta}{2} * A)^{-1} * (I + \frac{\Delta}{2} * A) \tag{2}$$

$$\overline{B} = (I - \frac{\Delta}{2} * A)^{-1} * \Delta B \tag{3}$$

$$\overline{C} = C \tag{4}$$

この方程式は、関数から関数へのマッピングではなく、配列から配列へのマッピングとなる：

$$x_k = \overline{A}x_{k-1} + \overline{B}u_k \tag{5}$$

$$y_k = \overline{C}x_k \tag{6}$$

これは RNN のように機能し、隠れ状態 x_k は時間を通して更新される。リカレント状態空間モデル（Recurrent SSM）は逐次的な性質を持つため、最新のハードウェアでは学習効率が悪い。その代わり、線形時不変状態空間モデル（LTI SSM）と逐次畳み込みの間には既知の関連がある。これに対応して、リカレント状態空間モデルは実際には離散畳み込みとして書ける。簡単のために、初期状態を $x_{-1} = 0$ とする。そして、展開すると次のようになる。

$$x_0 = \overline{B}u_0 \tag{7}$$

$$x_1 = \overline{AB}u_0 + \overline{B}u_1 \tag{8}$$

$$x_2 = \overline{A^2B}u_0 + \overline{AB}u_1 + \overline{B}u_2 \tag{9}$$

出力 y は次のようになる。

$$y_0 = \overline{CB}u_0 \tag{10}$$

$$y_1 = \overline{CAB}u_0 + \overline{CB}u_1 \tag{11}$$

$$y_2 = \overline{CA^2B}u_0 + \overline{CAB}u_1 + \overline{CB}u_2 \tag{12}$$

これをベクトル化し、コンボリューションとして表現できる。畳み込みカーネルの明示的な式は以下の通りである。

$$\begin{aligned} y_k &= \overline{CA^kB}u_0 + \overline{CA^{k-1}B}u_1 + \cdots + \overline{CAB}u_{k-1} + \overline{CB}u_k \\ y &= K * u \end{aligned}$$

ここで K は畳み込みカーネルまたはフィルタと呼ばれる：

$$K \in \mathbb{R}^L = (\overline{CB}, \overline{CAB}, ..., \overline{CA^{L-1}B}) \tag{13}$$

ここで L はカーネルサイズを示す。注意として、これは非常に大きなフィルタである。シーケンス全体と同じ大きさであり、多くのリソースを使い、学習時には非効率的である。長期依存性の鍵となる情報を含むカーネルは、そのままでは安定した学習ができず、また多くの計算機資源を使う。S4 はこの問題を解決する。

3.2 S4

S4 は特定の状態空間モデル (SSM) の具体例であり、状態行列 A は対角プラス低ランク (DPLR) 行列としてパラメータ化される。すなわち、A を $Lambda + pq^*$ とする。このパラメータ化には 2 つの主要な性質がある。 一つ目は、このパラメータ化によって構造化表現が可能になり、計算が高速化されることである。S4 は特別なアルゴリズムを使って畳み込みカーネル K を非常に高速に計算する。 2 つ目の特性は、このパラメータ化には HiPPO 行列と呼ばれる特別な行列が含まれることである。これにより、SSM は理論的にも経験的にも長距離依存性をよりよく捉えられる。特に、HiPPO は特別な方程式 $x_0(t) = Ax(t) + Bu(t)$ を提供する。この方程式は A と B の値に対して閉形式の解を持つ。この特殊な A 行列は DPLR 形式で表現でき、S4 はこの A と B 行列を初期値として設定する。

4 Method

本節では、提案するフレームワークを紹介する。バイノーラル音声を合成するための前処理を説明し、S4 を組み込んだモデルの詳細を説明する。

4.1 Overview of the Framework

ネットワーク入力は、事前学習によって生成されたバイノーラル音声の共通情報と位置情報から、音声、話者と聞き手の位置、モノラル音声を線形変換する。この考え方は、BinauralGrad [7] に基づいている。

4.2 Common and positional information of binaural speech generated by pre-training.

このセクションでは、事前学習中に両耳音声の共通情報と位置情報を扱うために、BinauralGrad [7] モデルで使われているアプローチからヒントを得る。その目的は、順方向処理と逆方向処理の間で一貫した条件情報を維持しながら、左右の音声チャンネル間で共有される共通情報の生成を促すことである。事前学習では、ゴールデンバイノーラル音声の平均 $\overline{y}$ を順方向のクリーンデータとみなす。具体的には、我々のモデルは $x_{textavg} = \text{mean}(x_l^{\text{warp}}, x_r^{\text{warp}})$ と示される、ワープしたバイノーラル音声の平均を条件情報とする。この事前学習段階を含めることで、我々のモデルは、左右の音声チャンネルで共有される共通の特徴と、ニュアンスに富んだ位置情報の両方を捉えて生成することができ、両耳音声合成の品質に貢献する。

4.3 Definition of task

本研究では、両耳音声合成のためのタスクを定義する。入力データはモノラル音声信号とリスナーと音源の時系列位置データである。音声データは 48kHz でサンプリングされ、位置データは 120Hz でサンプリングされる。位置データには、リスナーと音源の頭の位置を表す（x,y,z）座標と、頭の向きを表す四元数（qx,qy,qz,qw）が含まれる。音声と位置データは同期しており、400 サンプルごとに新しい位置データが提供される。このタスクの出力は 2 チャンネルのバイノーラルオーディオで、サンプリングレートは 48kHz である。

4.4 Evaluation Metrics

提案手法を評価するために、いくつかのメトリクスを用いる。まず，リスニング実験により方向検出の精度を評価する．また，元データと生成音の位相差と波形差を評価値として用いる．さらに，提案手法を従来手法と比較し，その性能の優位性を検証する．

4.5 Dataset

実験にはバイノーラル音声合成（BinauralSpeechSynthesis）に含まれるデータセットを使用する。このデータセットには、8 人の被験者のトレーニングデータと、8 人の被験者のテストデータと追加検証シーケンスが含まれている。各被験者のディレクトリには、モノラル音声信号（mono.wav）、バイノーラル録音（binaural.wav）、2 つの位置ファイルが含まれている。音声ファイルは 48kHz のサンプリングレートで記録され、位置ファイルは 120Hz のサンプリングレートで頭の位置と向きを追跡する。位置ファイルは 1 行につき 1 つの追跡サンプルを持ち、120 行で 1 秒間の追跡位置を表す。位置は (x,y,z) 座標で表現され、頭の向きは四元数 (qx,qy,qz,qw) で表現され、各行は 7 つの浮動小数点数 (x,y,z,qx,qy,qz,qw) を含む。

4.6 Proposed model

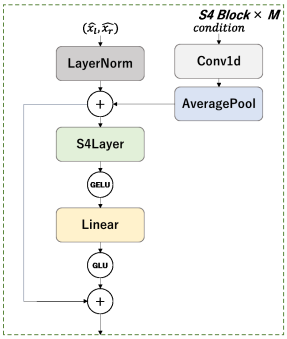

図 1　S4 Block

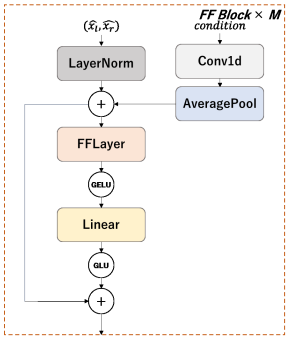

図 2　FF Block

BinauralS4 モデルは S4 層に基づくネットワークアーキテクチャである。モデルの次元は $d_{\text{model}} = 64$ で、各ダウンプールレベルに $n_{\text{layers}} = 8$ の残差（S4）～残差（FF）ブロックを持つ。これらのブロックは、ネットワークが、全く新しい、参照されない関数を学習しようとするのではなく、レイヤの入力に関して残差関数 [4] を学習することを可能にする。ダウンプーリングのプーリング係数は [4, 4]、特徴拡張は 2、FF 逆ボトルネックの拡張係数は 2、双方向層を使うかどうかは bidirectional=False である。また、unet-like アーキテクチャを使用する場合は unet=True とする。ドロップアウト率は 0.0 とする。 モデルはダウンブロック、センターブロック、アップブロックからなる。ダウンブロックではダウンサンプリングと特徴拡張を行い、センターブロックでは S4 レイヤーを適用する。アップブロックはアップサンプリングを行い、ダウンブロックからスキップされたコネクションを取り込むことで情報の流れを改善する。ネットワークへの入力は (x_r, x_l) である。条件として、$(mono, view, x_{avg})$。x_r, x_l は、モノラル音声と位置ビューを用いた線形変換により、以下のように変換される。$x = (y_l, y_r)$ は、n サンプルの長さを持つモノラル音声源 $x \in \mathbb{R}^N$ と、音源とリスナー間の相対空間位置 p が与えられたとき、我々の目的は、左右の耳用の 2 つの音声チャンネルを含むバイノーラル音声 $y = (y_l, y_r)$ を生成することである。この変換は次のように表される：

$$(y_l(n), y_r(n)) = f(x(n), p) \tag{14}$$

ここで、y_l と y_r は $\mathbb{R}^N$ の中にあり、f は我々の提案するフレームワークでパラメータ化された変換関数を表す。また、両耳音声の平均を $\overline{y} = \text{mean}(y_l, y_r)$ と定義する。

ここで、$p_s = (p_x, p_y, p_z)$ は座標で示される空間位置を表し、$p_\alpha = (q_x, q_y, q_z, q_w)$ はリスナーから音源への頭の向きを示す四元数を表す。リスナーは静止しており、座標系の原点に位置していると仮定する。その結果、軸 (p_x, p_y, p_z) はそれぞれ正面、右方向、上方向を示す。さらに、リスナーの左耳と右耳の空間位置をそれぞれ $p_{\text{l\_lstn}}$ と $p_{\text{r\_lstn}}$ とする。

左右の耳の時間的な差を揃えるために、音源とリスナーの距離を考慮したノンパラメトリックな方法であるジオメトリック・ワーピングを採用している：

$$\rho(n) = n - C \cdot ||p_{\text{src}}(n) - p_{\text{lstn}}(n)|| \tag{15}$$

ここで、n は現在のタイムスタンプを表し、C はオーディオのサンプリング レートと音速の比に基づいて計算される定数である。予測されたワープフィールド $\rho(n)$ は通常浮動小数点値を含むので、線形補間を使っ てワープ信号を計算する：

$$x_{\text{warp}}(n) = (\lceil \rho(n) \rceil - \rho(n)) \cdot x_{\lfloor \rho(n) \rfloor} + (\rho(n) - \lfloor \rho(n) \rfloor) \cdot x_{\lceil \rho(n) \rceil} \tag{16}$$

ここで、$\lceil \cdot \rceil$ と $\lfloor \cdot \rfloor$ はそれぞれ天井と床の関数を表す。左右両耳のワーピングは、位置 $p_{\text{lstn}}(n)$ を調整することで実現できる。結果として得られるワーピングされたバイノーラルオーディオは $(x_{l\text{warp}}, x_{r\text{warp}})$ と表記される。しかし、この方法はオーディオの回折を考慮しないので、オーディオ品質が低くなる可能性があることに注意することが重要です。Condition は、図 3 中のネットワークを用いて特徴量を抽出する。条件は畳み込み層と活性化関数で構成され、それぞれ音声（モノラ

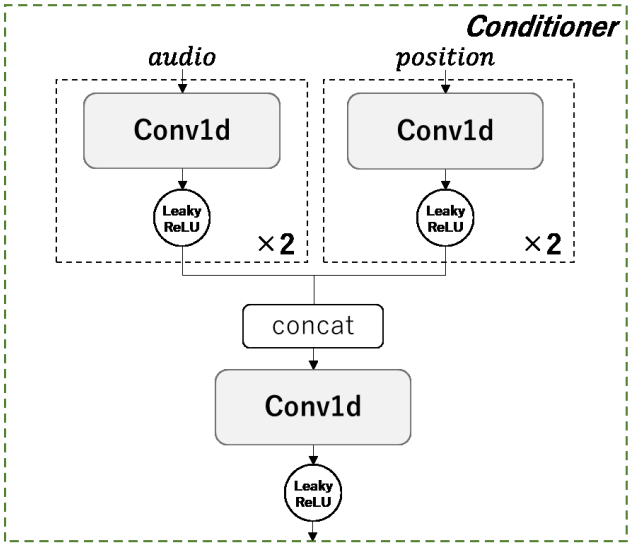

図 3　Conditioner

ル、$x_{l\text{warp}}$、$x_{r\text{warp}}$、x_{avg}）と位置（視点）を処理した後、合成する。x_{avg} は、*overliney* に基づいて事前に訓練されたモデルを用いて、生成されたゴールデンオーディオの左右共通部分を表す。 結合された特徴量は ResidualBlock に渡される。この条件は residualBlock で conv と avgpool によってリサイズされ、ダウン／アップサンプル層で $(x_{l\text{warp}}, x_{r\text{warp}})$ に結合される。

5　Experiments

5.1　Evaluation Metrics

本研究では、提案した BinauralS4 モデルの性能を評価するために 4 つの評価指標を用いた： 1. L2 誤差： グランドトゥルースのバイノーラルオーディオと生成されたバイノーラルオーディオ間の L2（ユークリッド）距離。これは 2 つの信号間の全体的な波形の類似性を測定する。 2. 振幅誤差： グランドトゥルースと生成されたバイノーラル音声の振幅間の平均絶対誤差（MAE)。この指標は、生成された信号の振幅の正確さを評価する。 3. MRSTFT 誤差： 多重解像度短時間フーリエ変換（MRSTFT）損失。スペクトル収束、対数マグニチュード損失、線形マグニチュード損失を考慮することで、多重解像度スペクトル損失をモデル化する。この指標は、両信号の周波数成分の時間的整合性を評価する。 4. 位相誤差： グランドトゥルースと生成されたバイノーラル音声の位相間の平均絶対誤差（MAE)。この指標は、生成された信号の位相情報の正確さを評価する。

5.2　Evaluation Procedure

BinauralS4 モデルを評価するために、様々なモデル構成で実験を行った。具体的には、異なるレイヤー数（6、8、16）と異なる次元（64 と 128）のモデルを探索した。各構成について、訓練データセットでモデルを訓練し、テストデータセットでその性能を評価した。 評価中、テストデータセットの各サンプルについて、グランドトゥルースのバイノーラル音声と生成されたバイノーラル音声の間の L2、振幅、MRSTFT、位相誤差を計算した。この誤差をすべてのサンプルで平均し、各指標の最終評価スコアを得た。さらに、レイヤーの数と次元を変えることがモデルの性能に与える影響を明らかにするために、異なるモデル構成の結果を比較した。

5.3　Results

表 2 は、異なるモデル構成での評価結果を示している。層数を増やすと一般的に性能が向上することが確認され、8 層モデルは L2 と MRSTFT で最も低いエラー値を達成した。16 層モデルは、バイノーラルオーディオに必要なフェーズで最も高い性能を示した。レイヤー数の増加に伴う性能向上には相関が見られなかった。また、レイヤー数を増やすとトレーニング時間も大幅に増加した。このモデルは、ダウンサンプル係数に起因する不自然なノイズがあったが、プールサイズを小さくすることで不自然なノイズが取り除かれた。

表 1　Evaluation Results for Different Model Configurations

Layers	L2($\times 10^3$)	Amplitude	Phase	MRSTFT
4	0.121	0.033	0.901	1.663
6	0.119	0.032	0.901	1.626
16	0.121	0.032	0.858	1.687

表 2　Evaluation Results for Different Model Configurations

Models	L2($\times 10^{-3}$)	Amplitude	Phase	MRSTFT
DSP	0.725	0.060	1.584	2.140
Warpnet	0.144	0.036	**0.804**	1.755
BinauralGrad	0.129	**0.030**	0.837	**1.282**
ours	**0.121**	0.032	0.851	1.747

全体として、BinauralS4 モデルはすべての評価指標で有望な結果を示し、Pool 値を下げた 16 層モデルが最高の性能を達成し、より深く広いモデル構成がより良いバイノーラルオーディオ合成品質につながることを示している。この結果は、提案手法が従来のモデルと同レベルの品質を生成できることを示している。モデル構成の選択は、特にリアルタイムアプリケーションの場合、性能と計算コストのバランスを考慮すべきである。

6　conclusion

構造化状態空間シーケンスモデル（Structured State-Space Sequence Model: S4）は、シーケンスモデリングにおける最近の革新的技術であり、様々なタスクやモダリティにまたがる長距離依存性の処理において卓越した性能を示す。状態空間モデル (SSM) をパラメータ化することで、S4 は理論的な利点を維持しながら効率的な計算を可能にし、シーケンスモデリングの分野に大きな進歩をもたらした。音声処理の分野では、S4

は自動音声認識（ASR）や音声合成（TTS）などのタスクにおいて、自己アテンションモデルの代替となる可能性を示している。本研究では、両耳音声合成のために S4 アーキテクチャに基づく新しいモデルを開発し、潜在状態空間におけるモノラル音声とリスナーと音源の位置の関係を表現する。 その結果、Wave L2、Amplitude L2、Phase L2、Multi-Resolution Short-Time Fourier Transform (MRSTFT) の各メトリクスの観点から、我々のアプローチが同等の音声合成品質を達成できることが示された。これらの結果は、音声合成分野における S4 モデルの可能性を裏付けるものである。この結果は、我々のモデルがシーケンスモデリングにおける S4 の適用可能性を拡張し、条件付き音声合成の領域における可能性を示していることを示している。両耳音声合成における S4 モデルの応用の成功は、バーチャルリアリティオーディオ、音の空間化、パーソナライズされた聴覚体験など、様々なオーディオ関連アプリケーションにおける新たな可能性を開くものである。 将来的には、このモデルのアーキテクチャをさらに改良し、より広範で多様な音声データに対する性能を調査する予定である。さらに、音源分離や音声ノイズ除去など、音声に関連する他のタスクを処理するためにモデルを拡張する予定である。全体として、我々の研究の有望な結果は、シーケンスモデリング技術の進歩に貢献し、音声合成分野での S4 の可能性を示すものである。

参考文献

[1] D. Begault. 3-d sound for virtual reality and multimedia. 09 2001.

[2] I. D. Gebru, D. Marković, A. Richard, S. Krenn, G. A. Butler, F. De la Torre, and Y. Sheikh. Implicit hrtf modeling using temporal convolutional networks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3385–3389, 2021.

[3] A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces, 2022.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020.

[6] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021.

[7] Y. Leng, Z. Chen, J. Guo, H. Liu, J. Chen, X. Tan, D. Mandic, L. He, X.-Y. Li, T. Qin, S. Zhao, and T.-Y. Liu. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis, 2022.

[8] Y. Lin and D. Lee. Bayesian regularization and nonnegative deconvolution for room impulse response estimation. *IEEE Transactions on Signal Processing*, 54(3):839–847, 2006.

[9] T. Lokki, L. Savioja, R. Vaananen, J. Huopaniemi, and T. Takala. Creating interactive virtual auditory environments. *IEEE Computer Graphics and Applications*, 22(4):49–57, 2002.

[10] K. Miyazaki, M. Murata, and T. Koriyama. Structured state space decoder for speech recognition and synthesis, 2022.

[11] A. Richard, D. Markovic, I. D. Gebru, S. Krenn, G. A. Butler, F. Torre, and Y. Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021.

[12] A. Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, Mar. 2020.

[13] 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, 2016.

[14] D. Zotkin, R. Duraiswami, and L. Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6(4):553–564, 2004.

[15] 平原達也, 大谷真, and 戸嶋巌樹. 頭部伝達関数の計測とバイノーラル再生にかかわる諸問題. 電子情報通信学会, 基礎・境界ソサイエティ,fundamentals review. 2(4):468–485, 2009.