4Y-07

構造化状態空間シーケンスモデルによるバイノーラル音声合成と音源と位置情報の長距離依存関係への応用 *

北村健太郎 †, 伊藤 克亘 ‡,

1 まえがき

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)[2] 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。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。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両耳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。

2 関連研究

2.1 バイノーラル音声合成

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ーを使う [5]。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。

2.2 ニューラルネットを使ったバイノーラル音声合成

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モデル（Temporal ConvNet[6]）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 [4] 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 提案手法

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。

3.1 システムの全体像

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 [4] に基づいている。

3.2 提案モデル

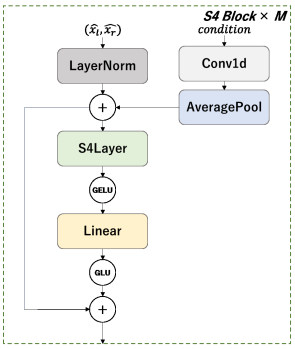

図 1. S4 Block

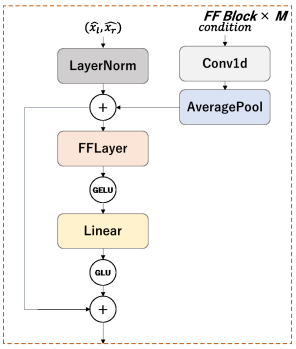

図 2. FF Block

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。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、レイヤ入力に関して残差関数 [3] を学習することを可能にする。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。) ブロックはダウンブロックの前に導入される。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

* :Binaural Audio Synthesis with Structured State Space sequence model *Kentaro Kitamura (Grad. School of CIS, Hosei Univ.) et al.*

† 法政大学大学院 情報科学研究科
‡ 法政大学情報科学部

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：

$$(y_l(n), y_r(n)) = f(x(n), p) \tag{1}$$

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：

$$\rho(n) = n - C \cdot ||p_{\text{src}}(n) - p_{\text{lstn}}(n)|| \tag{2}$$

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：

$$x_{\text{warp}}(n) = (\lceil \rho(n) \rceil - \rho(n)) \cdot x_{\lfloor \rho(n) \rfloor} + (\rho(n) - \lfloor \rho(n) \rfloor) \cdot x_{\lceil \rho(n) \rceil} \tag{3}$$

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ある。コンディションレイヤー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、$\overline{y}$ 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。

4 実験

4.1 評価手法

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：テストデータにあるバイノーラル音声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： テストデータにあるバイノーラル音声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： テストデータにあるバイノーラル音声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。

4.2 データセット

実験に BinauralSpeechSynthesis[6] 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。

4.3 結果

表 1. バイノーラル音声合成の性能比較

Models	L2($\times 10^{-3}$)	Amp	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。

5 あとがき

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。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、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。

参考文献

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

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

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

[4] 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.

[5] 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.

[6] 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.