# Mamba の “処理時間” はヒトの読み時間と符合する

山本悠士 [1,2]　磯野真之介 [2]　河原吉伸 [3,5]　横井祥 [2,4,5]
[1] 総合研究大学院大学　[2] 国立国語研究所　[3] 大阪大学　[4] 東北大学　[5] 理化学研究所
{yuji.yamamoto, s-isono, yokoi}@ninjal.ac.jp　kawahara@ist.osaka-u.ac.jp

## 概要

本研究は，有力な状態空間モデルである Mamba における単語あたりの概念上の処理時間と，ヒトの単語あたりの読み時間との間に対応があることを示す．Mamba では，各層における再帰的な状態遷移には概念上の時間（離散化ステップ $\Delta_t$）がかかり，その長さは入力に応じて変化する．この離散化ステップを用いて，文章読解におけるヒトの単語あたりの読み時間を予測したところ，その予測性能は GPT-2 サプライザルなどの既知の主要な変数に比肩しかつ独立に有効であった．また，Mamba のアーキテクチャの数学的分析から，異なる時間スケールで動くモジュールや，入力情報やノイズが記憶表象の変化にどう影響するかを検討できる関係式などが見つかり，Mamba がヒトの言語処理の分析に新たな視座をもたらすモデルであることが示唆された．

## 1　はじめに

人間言語は，本来的に時間軸に沿って産出・理解される．では，流暢に人間言語を操る（単方向の）人工言語モデルの表象には，時間に対応するものが存在するのだろうか．そして，存在するとすれば，ヒトの言語使用における時間の流れとどのような関係にあるのだろうか．

この問いに対して，状態空間モデル（SSM）ベースの言語モデルである Mamba [1] は興味深い観点を提供する．SSM は時間を連続的に扱うが，シンボル列を扱う SSM 言語モデルはその状態遷移を離散的な時間で取り扱う．特に，通常の SSM は，一定間隔の離散化ステップに従って状態遷移を行う．一方で Mamba は，離散化ステップ $\Delta_t$ が入力に応じて動的に決まるという特徴を持つ（図 1）．この設計により，Mamba は他の SSM 言語モデルよりも文脈の変化に柔軟に対応することができ，複雑で非定常な構造をもつテキストデータに対して経験的に高性能な言語モデリングをおこなうことができる．

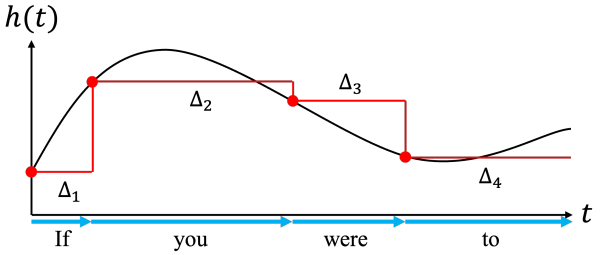


**図 1**　入力依存の離散化ステップ $\Delta_t$ が連続時間の状態 $h(t)$ を離散化するイメージ図．

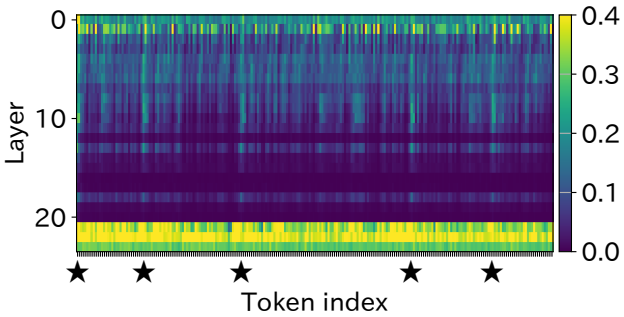


**図 2**　Natural Stories の先頭 5 文に対する 24 層の離散化ステップ $\Delta_t$ を可視化した．“★” は文頭の位置を示す．

本研究は，Mamba の単語単位の離散化ステップとヒトの読み時間の対応関係を示す．2 節で Mamba と読み時間モデリングについて概説し，Mamba の離散化ステップが単語の処理時間に概念上対応することをみたあと，3 節でコーパスに基づく読み時間モデリング実験の結果を報告する．4 節では Mamba における記憶のメカニズムの数学的検討から，ヒトの言語処理への示唆を論じる．

## 2　準備

### 2.1　Mamba

Mamba は，状態空間モデル（SSM）ベース言語モデルとして最有力のモデルである．SSM ベース言語モデルは Transformer とは異なり，系列全体を同時に処理するのではなく，ヒトと同様に，単語単位で入出力を逐次的に処理するため，メモリ消費量が

小さい．そのため，Mambaや，同様に再帰的構造を持つ Linear Attention [2] は，高効率な大規模言語モデルの開発に活用されている [3, 4, 5, 6].

Mambaの離散化ステップ $\Delta_t$ は，入力 $\boldsymbol{x}_t$ に応じて動的に決まり，各層において記憶の保持と忘却を制御するゲートとして機能する．MambaにおけるSSMは二つの式で構成される．一つは，記憶に対応する成分である隠れ状態 $\boldsymbol{h}_t$ を入力 $x_{ti}$ を用いて更新する式であり，もう一つは，隠れ状態 $\boldsymbol{h}_t$ から出力 $y_{ti}$ を生成する式である：

$$\boldsymbol{h}_t = \overline{A}_t \boldsymbol{h}_{t-1} + \overline{B}_t x_{ti}, \quad y_{ti} = C_t \boldsymbol{h}_t + D x_{ti}. \tag{1}$$

ただし，各係数 $\overline{A}_t, \overline{B}_t$ は，次のように与えられる：

$$\overline{A}_t = \exp(A\Delta_t), \quad \overline{B}_t = \Delta_t W_B \boldsymbol{x}_t \tag{2}$$

$$\Delta_t = \log(1 + \exp(W_\Delta \boldsymbol{x}_t + \boldsymbol{b}_\Delta)). \tag{3}$$

ここで，離散化ステップ $\Delta_t \in (0, \infty)$ は，記憶におけるゲートとしての役割を果たしている．例えば，$\Delta_t$ が大きい場合，係数 $\overline{A}_t$ の要素は0に近づき，その結果，前時刻の記憶 $\boldsymbol{h}_{t-1}$ の影響は弱まる．同時に，係数 $\overline{B}_t$ の絶対値は増大し，入力 $x_t$ の影響が強くなる．

## 2.2 読み時間モデリング

読み時間モデリングは，読み手が文中の各語を読むのにかける時間の変動を統計的に説明することで，文読解に伴う認知的負荷の要因を定量的に明らかにすることを目的とする研究手法である [7].

大規模言語モデルを用いた読み時間モデリングでは，ある単語 $w$ の文脈 $C$ での処理負荷が $-\log p(w \mid C)$ に比例するというサプライザル理論 [8, 9] がよく用いられてきた．Transformerに基づくサプライザルが読み時間を予測することが多くの言語で確かめられている一方 [10]，モデルサイズやコンテクスト幅が大きいとむしろヒトの傾向から離れていくことも指摘され，ヒトの記憶の制約との関係が示唆されている [11, 12]．本研究は，Transformerとは大きく異なる記憶の仕組みと時間の概念を持つMambaで，ヒトの読み時間が捉えられるかを見る．

## 2.3 処理時間としての離散化ステップ

本研究では，Mambaの離散化ステップ $\Delta_t$ が入力単語に対する処理時間に対応すると解釈する．Mambaは，入力単語 $w_t$ に応じて離散化ステップ $\Delta_t$ を変化させるように最適化されている．ヒトが文理解において高い認知負荷を要する単語に長く目を留めるように，Mambaは状態の適切な更新に要する処理時間を各単語に割り当てているとみることができる．

# 3 Mambaの離散化ステップはヒトの読み時間を予測する

Mambaの“処理時間”である $\Delta_t$ とヒトの読み時間との間に統計的に有意な関係があるかを見る．[1)]

## 3.1 実験設定

実験にはNatural Storiesコーパス [13] を用いた．同コーパスは10個の物語（英語，485文，10,245単語）からなり，それに母語話者181人が自己ペース読文法 [14] で行った単語ごとの読み時間が付されている．事前に参加者ごとの切片のみの線形混合効果モデル $\log(RT) \sim 1 + (1 \mid participant)$ をフィットし，単語ごとに残差の平均値をとって従属変数とした．

$\Delta_t$ の値はMamba-130m [2)] から取り，各時点での $\boldsymbol{\Delta}_t \in \mathbb{R}^d$ は次元方向に平均を取り，subwordに分かれた $\Delta_t$ は最大値を取ることで単語レベルに集約した．

分析は，読み時間を予測する線形回帰モデルを構築し，[3)] 10-fold交差検証を50回反復し単語ごとの平均二乗誤差（MSE）を評価することで行った．変数の追加によるMSEの減少（ΔMSE）が0より有意に大きいかを，置換検定により検定した．[4)]

## 3.2 結果

主要な結果を表1にまとめる．

まず，層ごとの $\Delta_t$ だけを入れた回帰モデルの予測力を切片のみのモデルとの比較で評価した．24ある層のうち19層が読み時間を有意に予測した[5)]（詳細は付録表2）．最も予測力が高かったのは第17層であった．次に，各層の $\Delta_t$ を組み合わせた場合の最善のモデルを探索した．層別の評価でΔMSEが高いものから順に，回帰モデルに加え，ΔMSEが有意に高くなるもののみを残した．その結果，14の層（第1,5,7,8,9,11,12,14,16,17,18,20,21,22層）が残った．

最後に，$\Delta_t$ の予測力がこれらの既知の変数から独立であるかを見るために，既知の変数を入れたベー

1) コードは https://osf.io/vnw5e/overview?view_only=93ad704fc6ea44438f3d3538b4b682eb で公開している．
2) https://huggingface.co/state-spaces/mamba-130m-hf
3) スピルオーバー（ある単語に起因する処理負荷が次単語以降の読み時間にも反映されること）を考慮し，$RT_t$ の予測には $w_{t-2}, w_{t-1}, w_t$ に対応する従属変数を用いた [10].
4) 層別の評価ではHolm法で $p$ 値を補正した．
5) ただし，このうち予測力の小さい一部の層は係数が負，すなわち $\Delta_t$ が大きいほど読み時間が短くなっている．この点は今後の検討が必要である．

**表 1**　既知の主要な変数と $\Delta_t$ に基づく回帰モデルの性能比較．位置は単語の文章中および文中の位置．ΔMSE は単語あたりの値を $10^3$ 倍．$\Delta_t$ 最善モデルは他モデルよりも変数が多いことに留意．

| | ΔMSE | $R^2$ |
|---|---|---|
| ベースライン（下記組み合わせ） | 3.70 | 0.52 |
| 文字数 | 1.31 | 0.19 |
| 頻度 | 1.54 | 0.22 |
| 位置 | 1.66 | 0.23 |
| GPT-2 サプライザル[6] | 1.45 | 0.21 |
| **第 17 層 $\Delta_t$** | **1.29** | **0.18** |
| **$\Delta_t$ 最善モデル** | **2.19** | **0.32** |
| ベースラインを加えた場合の比較： | | |
| **第 17 層 $\Delta_t$** | **0.06** | |
| **$\Delta_t$ 最善モデル** | **0.23** | |

スライン回帰モデルに $\Delta_t$ を加えた場合に予測力が向上するかを調べた．24 層中 14 層で有意な向上が見られた（詳細は付録表 3）．また，向上幅が最も大きかったのはやはり第 17 層であった．モデル選択では 8 層（第 0,7, 9,13,17,21,22,23 層）が残った．

総合すると，多くの層の $\Delta_t$ が読み時間を予測し，GPT-2 サプライザルを含む既知の主要な変数に比肩しかつ独立に有意な予測力を持つことがわかった．

## 3.3　追加分析

$\Delta_t$ の読み時間に対する予測力は何に由来するのかを理解するため，$\Delta_t$ について追加分析を行った．

### 3.3.1　言語学的特徴にみる層間の分担

文章中の各単語 $w_t$ の言語学的な諸特徴と $\Delta_t$ との相関を図 3 に示す．諸特徴としては，読み時間の回帰に用いたもののほか，文頭・文末（該当すれば 1，しなければ 0），および前単語からの統語木（Penn TreeBank 形式）上のパス長を含めた．

全層でほぼ一貫した傾向としては，文中の位置（何単語目か）と $\Delta_t$ が負に相関し，文頭と $\Delta_t$ は正に相関した．これは，多くの層が文頭でその保持する情報を大きく変化させる一方，文の内部では情報を保持する傾向にあることを示す．また，統語木のパス長とは正の相関がみられた．これは，構造上の切れ目では層の持つ情報が大きく変化することを示す．いずれも，Mamba の $\Delta_t$ が言語の基本的な構造を捉えていることを示唆する．

層間での分業も見られる．第 16・17 層は周辺の層とはっきりと違った特徴を見せる．物語中の位置

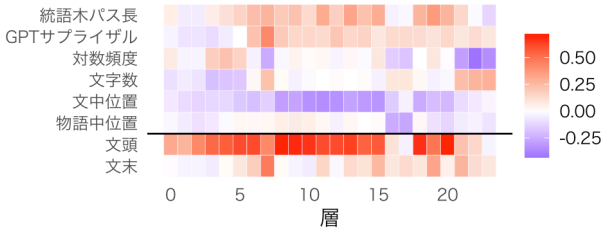


**図 3**　単語 $w_t$ の言語学的な特徴と $\Delta_t$ の相関．「文頭」「文末」については全単語で算出，それ以外については文頭・文末のトークンを除いて算出（読み時間分析に従う）．

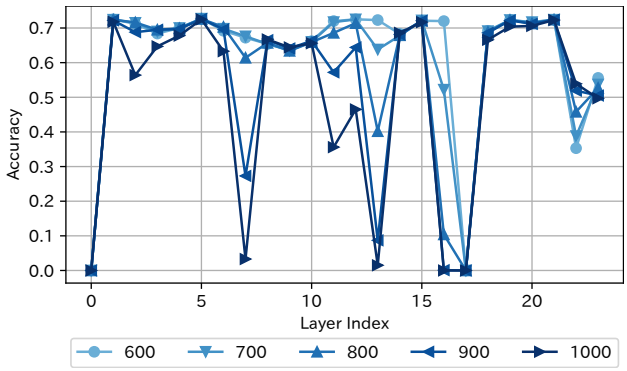


**図 4**　各層を個別にノックアウトしたときの Passkey Retrieval タスクの正解率．正解率は，1,000 通りの入力例を用いて評価した．凡例は入力長を示す．

と負の相関がある一方，文中位置・文頭・統語木パス長との相関が弱い．これは，これらの層が文内部の局所的な構造にあまり反応せず，むしろ物語全体の情報を保持していることを示唆する．第 16・17 層は読み時間の予測力が際立って高く（表 1，付録表 2），このことはこれらの層がヒトの物語読解における長い文脈の処理を捉えていることを示唆する．

### 3.3.2　介入実験による長距離依存性の分析

前節の観察を踏まえ，本節では，各層（特に第 16・17 層）が長距離依存関係を捉えているかを Passkey Retrieval タスク [15] における介入実験により調査する．本タスクは，モデルが長距離依存関係を処理できるかどうかを評価するために用いられ，ノイズトークン列の中に埋め込まれた特定の文字列を正しく復元できるかを測定するタスクである．

Mamba の各層の SSM を 1 つずつノックアウト[7]した状態で Passkey Retrieval タスクを実行した．モデルには，6 桁の数字からなるパスキーが与えられ，続いて複数のノイズ文が提示され，最後にパスキーを再現するよう指示するプロンプトが与えら

7）　具体的には，式 (1) における入力に対する係数 $\overline{\boldsymbol{B}}_t$ を全時刻で 0 に設定した．これにより $\boldsymbol{h}_t = 0$ が常に維持される．

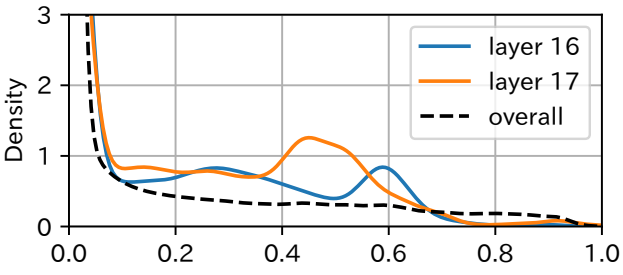


**図 5**　遷移行列 $\exp(A)$ の固有値分布．全層の固有値の分布と，第 16,17 層の固有値の分布をプロットした．

れる．[8]結果を図 4 に示す．図より，第 0,7,13,16,17 層をノックアウトした場合，正解率が顕著に低下した．特に，第 0・17 層では入力長を 600 トークンまで短縮しても正解率が 0% のままであり，これは，これらの層が長距離にわたる情報伝達を必要とする推論において不可欠であることを示している．

# 4　Mamba から見るヒトの言語処理

ここまでは，Mamba の “処理時間” がヒトの読み時間を予測できることを実証的に見た．以下では Mamba の設計や内部動態の理論的分析を通じて，Mamba が時間の中で展開するヒトの言語処理の理解に有用な示唆をもたらしうることを述べる．

## 4.1　遷移行列の固有値から見る記憶保持

状態空間モデルの各層における情報の長期保持能力は，遷移行列 $\exp(A)$ の固有値により評価できる．固有値の絶対値が 1 に近いほど情報は保持されやすく，0 に近いほど忘却が進む．図 5 に，全層および $\Delta_t$ が文境界に反応しない第 16・17 層における固有値分布を示す．同図より，全層の固有値は 0 付近にピークを持つ一方で，第 16・17 層の固有値分布は全体の分布と比較して 0.5 付近の固有値の割合が高いことが分かる．したがって，これらの層は $\Delta_t$ の挙動に加え，SSM のダイナミクス上も情報を長期保持しやすい性質を持つことが示唆される．

つまり，Mamba の各層は，遷移行列の固有値によって表現できるような異なる記憶保持能力を持っていることがわかる．ヒト脳も異なる時間スケールで展開する言語情報を追跡できることがデコーディング研究からわかっており [16]，Mamba との比較は新たな視座をもたらす可能性がある．

8）　具体的には，次のような入力を与える：「The passkey is 317451. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again. ... The passkey is」

## 4.2　離散化ステップから見る状態遷移の不確かさ

離散化ステップ $\Delta_t$ が，状態遷移における不確かさに対応することを形式的に示す．まず，Mamba の隠れ状態遷移を確率モデルとして解釈するために，形式的にノイズ項を導入する．具体的には，元々の遷移関数は式 (1) で与えられているが，これにガウスノイズ項 $w(t) \sim \mathcal{N}(0, Q_c)$ を加える：

$$h'(t) = A(t)h(t) + B(t)x(t) + w(t). \tag{4}$$

これは，状態遷移を次のような確率モデルとみなしていることを意味する：

$$p(h_t \mid h_{t-1}) := \mathcal{N}\left(h_t\,;\, \overline{A}_t h_{t-1} + \overline{B}_t x_t, Q_d\right). \tag{5}$$

この標準的な仮定の下で，状態遷移 $\boldsymbol{h}_t$ に関する条件付きエントロピー（不確かさ）は $\Delta_t$ によって決まり，$\Delta_t$ が大きくなるほど不確かさも増大することが分かる：

$$H[h_t \mid h_{t-1}] = \frac{1}{2}\sum_{i=1}^{n} \log\left(e^{2A_{ii}\Delta_t} - 1\right) + \text{const.}, \tag{6}$$

導出については，付録 A を参照されたい．

では，Mamba は入力に応じてどのように離散化ステップ $\Delta_t$ を調整しているのだろうか？図 2 や 図 3 より，いくつかの層では，離散化ステップ $\Delta_t$ は文頭トークンにピークをとる．これより，実際に Mamba は，直前の文脈から予測することが難しい箇所で $\Delta_t$ を増大させていることが示唆される．

この不確かさは，ヒトがノイズのある表象を使って言語処理を行う際に直面するものに概念的に近い．ノイズ下でのヒトの言語処理は近年注目を集めるトピックであるが，ノイズは単語の確率的除去として実装されている [17, 18]．Mamba の連続的な記憶表象へのノイズの付加は，ノイズ下のヒトの言語処理にも示唆をもたらす可能性がある．

# 5　おわりに

本研究では，Mamba の “処理時間” といえる離散化ステップ $\Delta_t$ が，ヒトの単語ごとの読み時間を予測し，その予測力は GPT-2 サプライザルなどの既知の変数をベースラインに入れても有意であることを示した．また，Mamba のアーキテクチャの分析からは，時間の展開とともに記憶を更新しながら行われるヒトの言語処理に，Mamba が新たな視座をもたらしうることが示唆された．

## 謝辞

本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2104，JST 創発的研究支援事業 JPMJFR2331，および JSPS 科研費 JP22H05106，JP25K22996 の支援を受けたものである．

本研究は，国立国語研究所において毎週開催されている「317 カフェ」[9] での議論を端緒として着想されたものである．ここに，317 カフェの参加者の皆様に感謝の意を表する．

## 参考文献

[1] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In **First Conference on Language Modeling**, 2024.

[2] Songlin Yang and Yu Zhang. Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism, January 2024.

[3] Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. Falcon mamba: The first competitive attention-free 7b language model. **arXiv preprint arXiv:2410.05355**, 2024.

[4] Preferred Networks, Kaizaburo Chubachi, Yasuhiro Fujita, Shinichi Hemmi, Yuta Hirokawa, Kentaro Imajo, Toshiki Kataoka, Goro Kobayashi, Kenichi Maehashi, Calvin Metzger, et al. Plamo 2 technical report. **arXiv preprint arXiv:2509.04897**, 2025.

[5] Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. Kimi linear: An expressive, efficient attention architecture. **arXiv preprint arXiv:2510.26692**, 2025.

[6] Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, et al. Nvidia nemotron 3: Efficient and open intelligence. **arXiv preprint arXiv:2512.20856**, 2025.

[7] 栗林樹生, 大関洋平, Ana Brassard, 乾健太郎. ニューラル言語モデルの過剰な作業記憶. 言語処理学会 第 28 回年次大会 発表論文集, March 2022.

[8] John Hale. A probabilistic Earley parser as a psycholinguistic model. In **Second Meeting of the North American Chapter of the Association for Computational Linguistics**, 2001.

[9] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, pp. 1126–1177, 2008.

[10] Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. Testing the predictions of surprisal theory in 11 languages. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 1451–1470, 2023.

[11] Byung-Doh Oh and William Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 336–350, 2023.

[12] Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. Context limitations make neural language models more human-like. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10421–10436, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. **Language Resources and Evaluation**, Vol. 55, No. 1, pp. 63–77, 2021.

[14] Marcel Adam Just, Patricia A. Carpenter, and Jacqueline D. Woolley. Paradigms and processes in reading comprehension. **Journal of Experimental Psychology: General**, Vol. 111, No. 2, pp. 228–238, 1982.

[15] Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 54567–54585. Curran Associates, Inc., 2023.

[16] Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Remi King. Hierarchical dynamic coding coordinates speech comprehension in the human brain. **PNAS**, Vol. 122, No. 42, p. e2422097122, 2024.

[17] Richard Futrell, Edward Gibson, and Roger Levy. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. **Cognitive Science**, Vol. 44, No. 3, p. e12814, 2020.

[18] Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. A resource-rational model of human processing of recursive linguistic structure. **PNAS**, Vol. 119, No. 43, 2022.

9) https://sites.google.com/view/317cafe/

## A　状態遷移の不確かさの導出

本節では，離散化ステップと状態遷移の不確かさの対応関係を示唆する式 (6) の導出をする．

ガウスノイズ項 $w(t)$ を含む SSM の状態方程式：

$$h'(t) = A(t)h(t) + B(t)x(t) + w(t),\quad w(t) \sim \mathcal{N}(0, Q_c) \tag{7}$$

に対して Zero-Order Hold (ZOH) 離散化を適用すると次の差分方程式が得られる：

$$\boldsymbol{h}_t = \overline{A}_t \boldsymbol{h}_{t-1} + \overline{B}_t x_t + \boldsymbol{w}_t \tag{8}$$

$$\boldsymbol{w}_t \sim \mathcal{N}(\mathbf{0}, Q_d),\quad Q_d = \int_0^{\Delta_t} e^{A\tau} Q_c e^{A^\top \tau} d\tau. \tag{9}$$

ただし，$Q_d$ は連続時間ノイズ強度 $Q_c$ をサンプリング間隔 $\Delta_t$ で離散化したものであり，対角行列と仮定する．このとき，$\boldsymbol{h}_{t-1}$ が与えられた下での $\boldsymbol{h}_t$ は以下の分布：

$$p(h_t \mid h_{t-1}) := \mathcal{N}\left(h_t\,;\, \overline{A}_t h_{t-1} + \overline{B}_t x_t, Q_d\right). \tag{10}$$

に従い，式 (8) により $\boldsymbol{h}_t$ を $\boldsymbol{w}_t$ に変数変換することで以下も成り立つ：

$$p(h_t \mid h_{t-1}) = \mathcal{N}(\boldsymbol{w}_t\,;\, \mathbf{0}, Q_d). \tag{11}$$

Mamba の遷移行列 $e^A$ は対角行列であるため，式 (9) の被積分関数は対角行列より，積分は次のように解ける：

$$(Q_d)_{ii} = \int_0^{\Delta_t} e^{2A_{ii}\tau} (Q_c)_{ii} d\tau = (Q_c)_{ii} \frac{e^{2A_{ii}\Delta_t} - 1}{2A_{ii}} \tag{12}$$

このとき，状態遷移における不確かさ、すなわち条件付き微分エントロピー：

$$H[h_t \mid h_{t-1}] = -\int_{\mathbb{R}}^{n} p(h_t \mid h_{t-1}) \log p(h_t \mid h_{t-1}) dh_t \tag{13}$$

は，式 (11) より，正規分布 $\mathcal{N}(\boldsymbol{w}_t\,;\, \mathbf{0}, Q_d)$ についての微分エントロピーと等しい．したがって，状態遷移における不確かさは次のように与えられる：

$$H[h_t \mid h_{t-1}] = \frac{1}{2} \sum_{i=1}^{n} \log\left(e^{2A_{ii}\Delta_t} - 1\right) + \text{const.} \tag{14}$$

ただし，$\Delta_t$ に依存しない項を定数項 const. としてまとめた．最後に，§4.2 の議論は，決定論的モデルとしての元来の Mamba からは自然に導かれないことに留意する．

## B　読み時間モデリング結果の層ごとの詳細

**表 2**　各層の $\Delta_t$ のみを用いたモデリングの結果．Holm 法にて補正の上，$p < 0.05$ の層を抜粋．ΔMSE は単語あたりの値を $10^3$ 倍．係数および $R^2$ はデータ全体にフィットした場合の参考値．

| 層 | ΔMSE | $p$ 値 | 係数 $_t$ | $_{t-1}$ | $_{t-2}$ | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | 0.04 | 0.029 | 0.04 | 0.05 | 0.03 | 0.01 |
| 4 | 0.04 | 0.013 | -0.07 | -0.15 | -0.09 | 0.01 |
| 5 | 0.10 | 0.000 | -0.08 | -0.14 | -0.11 | 0.01 |
| 7 | 0.08 | 0.000 | 0.21 | 0.24 | 0.02 | 0.01 |
| 8 | 0.05 | 0.000 | -0.01 | -0.07 | -0.10 | 0.01 |
| 9 | 0.09 | 0.000 | -0.02 | -0.10 | -0.11 | 0.01 |
| 10 | 0.04 | 0.002 | -0.03 | -0.09 | -0.05 | 0.01 |
| 11 | 0.07 | 0.000 | 0.19 | -0.16 | -0.22 | 0.01 |
| 12 | 0.03 | 0.025 | 1.06 | 0.31 | 0.09 | 0.01 |
| 13 | 0.04 | 0.014 | 0.14 | -0.09 | -0.16 | 0.01 |
| 15 | 0.05 | 0.001 | 0.98 | -0.68 | -0.46 | 0.01 |
| 16 | 0.94 | 0.000 | 14.75 | 16.69 | 9.80 | 0.13 |
| 17 | 1.29 | 0.000 | 16.90 | 15.87 | 9.78 | 0.18 |
| 18 | 0.03 | 0.014 | 0.21 | -0.11 | -0.23 | 0.01 |
| 19 | 0.06 | 0.001 | 3.11 | -1.19 | -0.96 | 0.01 |
| 20 | 0.09 | 0.000 | 10.46 | 0.56 | 0.39 | 0.01 |
| 21 | 0.17 | 0.000 | 0.15 | 0.12 | 0.01 | 0.02 |
| 22 | 0.57 | 0.000 | 0.29 | 0.38 | 0.21 | 0.08 |
| 23 | 0.28 | 0.000 | 0.45 | 0.81 | 0.47 | 0.04 |

**表 3**　ベースラインモデルに各層の $\Delta_t$ を加えたモデリングの結果．表記は表 2 に同じ．

| 層 | ΔMSE | $p$ 値 | 係数 $_t$ | $_{t-1}$ | $_{t-2}$ | $R^2$ |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.000 | 0.17 | 0.15 | -0.05 | 0.53 |
| 1 | 0.03 | 0.009 | 0.04 | 0.04 | 0.00 | 0.53 |
| 2 | 0.02 | 0.030 | 0.08 | 0.07 | -0.02 | 0.53 |
| 5 | 0.02 | 0.009 | -0.06 | -0.05 | -0.06 | 0.53 |
| 8 | 0.04 | 0.000 | -0.05 | -0.07 | -0.06 | 0.53 |
| 9 | 0.04 | 0.000 | -0.05 | -0.06 | -0.07 | 0.53 |
| 10 | 0.03 | 0.000 | -0.05 | -0.07 | -0.04 | 0.53 |
| 11 | 0.03 | 0.000 | -0.02 | -0.12 | -0.11 | 0.53 |
| 13 | 0.02 | 0.003 | -0.03 | -0.07 | -0.12 | 0.53 |
| 16 | 0.04 | 0.000 | -2.38 | 2.14 | 5.67 | 0.53 |
| 17 | 0.06 | 0.000 | -0.80 | 3.58 | 5.82 | 0.53 |
| 18 | 0.02 | 0.017 | -0.03 | -0.15 | -0.09 | 0.53 |
| 19 | 0.02 | 0.030 | 0.20 | -1.08 | -0.86 | 0.53 |
| 22 | 0.04 | 0.000 | 0.13 | 0.09 | 0.03 | 0.53 |