社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

声道パラメータ表現および強化学習を利用したText-to-Action-to-Speech

小野　晶子†　加藤　徳啓††　高道慎之介†,†††

† 慶應義塾大学
†† 北海道大学
† 東京大学

あらまし　TTS モデルのほとんどは，テキストからの音声波形の出力を教師データを用いて学習するモデルである．しかし，人間は声道の制御によって教師データ無しで音声を生成する．調音パラメータ推定を強化学習に基づいて学習することで，人間の言語獲得のメカニズムを模倣することができる．調音パラメータを事前学習する際の教師データの品質および，報酬の設計が，モデル合成音の品質に大きく影響する可能性が示唆された．

キーワード　調音合成，強化学習，言語獲得

Akiko ONO†, Norihiro KATO††, and Shinnosuke TAKAMICHI†,†††

† Keio University
†† Hokkaido University
† The University of Tokyo

1. ま え が き

テキスト音声合成（text-to-speech; TTS）のほとんどは，大規模データセットを用いてテキストから音声特徴量や音声波形への変換を教師あり学習によって学習するアプローチである [1], [2]．近年主流である，end-to-end の TTS モデルでは，テキストから直接音声波形を生成することで，波形の各サンプル値までがモデルによる制御の対象となるため，高品質な音声合成が可能となる [1]．しかし，このような TTS の枠組みでは，ニューラルネットワークのパラメータが内部で暗黙的に記憶されているため，動作原理がブラックボックスであり，外部からの指示による制御能力や，説明可能性の点で限界がある [3]．また，高品質な合成のためには大量のペアデータを必要とし，実際，最新の TTS モデルは数万時間規模の音声コーパスで学習されている [4].

一方，人間の発話は，舌・唇・顎など複数の調音器官の協調運動によって物理的に生成される．各調音器官の役割は明確であり，調音器官の動きが異なると異なった音声が出力されるため、発せられた音声からその調音の仕組みを理解しやすい [5]．また，このため，話者は自分の発音をフィードバックを通じて学習できる．実際，乳児は口の動きを記憶するのではなく，音響フィードバックに基づく試行錯誤を通じて発話を習得する [6]．このような学習プロセスは，教師あり学習のように正解の音声データとの誤差を直接最小化していくものではなく，自ら発した音に対する聴覚フィードバックおよび周囲からの反応に基づく試行錯誤による過程である [6].

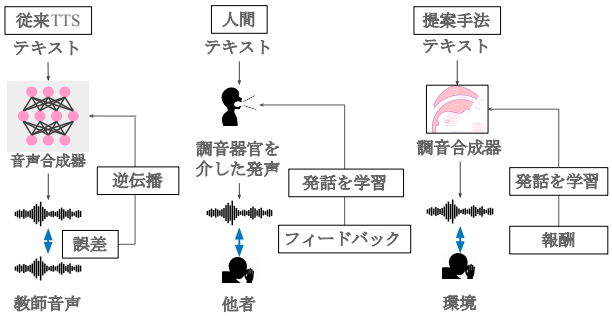

図 1: 従来手法の音声合成の学習（左），人間の音声生成の学習（中），提案手法の音声合成の学習（右）．従来は音声の教師データを用いて学習するが、人間は教師データなしに他者からのフィードバックにより学習する．本論文では、人間の学習の枠組みで音声合成を学習可能かを検討する．

図 1にコンセプトを示すように，本研究では，音声合成においてこのような人間と同様の学習手法が実現可能かを調査する．具体的には，テキストから音響特徴量へのマッピングを直接学習するのではなく，調音パラメータを介して声道を表現するモデルを構築し，強化学習によりその制御戦略を得る．これにより，人間の言語獲得メカニズムを模倣した学習が可能となる．この実現の狙いは (1) 音声合成の過程を声道運動により解釈可

能にすること，(2) 人間の学習と対比することでそれぞれの学習過程をそれぞれ明らかにすること，さらには (3) 音声生成の物理機構を持つ AI の実現に向けたシミュレーション (sim2real [7] の sim 部分) を実現することにある．

2. 関 連 研 究

2.1 音声合成モデルの学習

近年の TTS モデルのほとんどは，テキストから音声への中間表現として，特徴量を出力し，教師あり学習によりその予測モデルを訓練する [1],[2]．モデルの損失関数としては，特徴量間の L1/L2 損失や敵対的損失などが用いられる [8]．このような TTS は高品質な音声を合成するが，内部の動作原理がブラックボックスであり，柔軟な制御能力や，説明可能性の点で限界がある [3].

このような問題を解決しようとするアプローチのひとつとして，調音合成がある [2]．調音合成では声道の形状や声帯振動などの調音パラメータによって音声を生成するため，各パラメータの物理的意味が明確であり，また，多様な音声を柔軟に生成できる．本研究では，調音合成のアプローチを活用し，調音パラメータの生成による音声合成を学習することで，人間の音声合成の模倣を試みる．

2.2 人間の音声生成の学習

人間は成長過程で自ら声を出し，その結果を聞き取って調整するというサイクルを繰り返すことで音声生成を習得する．母親が乳児の発声にタイミングよく応答すると，乳児の喃語はより母親の発話に含まれる音韻パターンに近づくという研究報告もあり，社会的・聴覚的フィードバックが音声の獲得を促進することが示唆されている [6]．人間の音声生成の学習は，正解の音声を教わるのではなく，環境との相互作用の中で試行錯誤によって進む過程であり，教師あり学習とは異なる．乳児は自分の発声器官の動きを一つ一つ記憶するのではなく，自然に声を出しながらその結果を評価・調整し，正解の音声へと近似させていく学習を行っていると考えられる．本研究では，教師あり学習ではなく，エージェントが報酬を得て方策を学ぶ強化学習を用いることで，人間が試行錯誤しながら音声生成を学習する過程を模倣することを試みる．

2.3 声道のパラメータ表現

発話時のある瞬間における声道形状は，適切なパラメータの組によって表現することができる．古典的には，声道をいくつかの管セグメントに分割し各セグメントの断面積や半径で形状を表す面積関数モデルが用いられてきた [9]〜[11]．VocalTractLab (注1) は，調音パラメータによって高精度に声道をモデル化できるシステムである [12]．VocalTractLab は精緻な解剖学的モデルに基づき連続的な調音運動や共発音を再現できるが，パラメータ制御の計算負荷が大きいことが指摘されている [13].

一方，pink trombone (注2) は，リアルタイム動作を重視したオープンソースの調音合成システムである．pink trombone は声道を 44 個の管セグメントに分割した簡易物理モデルに基づき，各セグメントを伝播する音の流れをシミュレーションする [13]．pink trombone は簡素化されたモデルであり，計算効率と操作性の高さから調音合成に関する複数の先行研究に用いられている [13] [14]．本研究では，pink trombone を利用し，計算効率と解釈性を両立した調音パラメータ生成モデルを構築する．

2.4 音声からの調音パラメータ推定

ある音声波形が与えられたとき，それに対応する声道の調音パラメータを推定する問題は，音声-調音逆変換（acoustic-to-articulatory inversion; AAI）と呼ばれる [15]．AAI は前節で述べた調音合成の逆問題にあたり，音響信号から元の調音情報を復元する試みである．異なる調音動作で類似の音響が生じ得るため，音声から調音パラメータへのマッピングは一意に定まらないことも多く，音声解析の分野における挑戦的な課題とされている [15].

AAI 研究においては，合成された音声から声道形状を推定する，合成による解析（analysis-by-synthesis; AbS） [16] の手法が主に提案されてきた [17],[18]．これらの手法には，EMA や MRI を用いた直接的な声道形状の計測を避けられるという利点がある [13].

さらに，調音合成システム自体の微分可能な実装を利用し，勾配法に基づく変換を行う手法も提案されている．声道勾配降下法（vocal tract gradient descent; VGD）は，pink trombone を PyTorch 上に再実装することでホワイトボックス型の最適化を実現し，勾配降下により任意の母音音声から声道形状および声源パラメータを推定する手法である [14]．VGD では，音声のスペクトル特性と調音パラメータから計算される声道の周波数応答を比較する損失関数を定義し，その勾配を調音パラメータ空間に伝播させることで最適化を行う．また VocalTrax は，pink trombone のエンドツーエンドの最適化により音声から調音パラメータを自動推定する手法である [2]．VocalTrax では，既存手法に比べ異なるドメインの音声に対する再現性が向上したことが示されている．

本研究では，AAI を利用し，音声波形から調音パラメータを推定する．得られたパラメータ系列をテキストと対応付けて，強化学習の事前学習のための教師データとして利用する．

2.5 音声からの調音パラメータ推定における強化学習

AAI について，強化学習（reinforcement learning; RL）を用いて調音パラメータ推定を行う研究も存在する．この手法は，教師データを直接用意せずに環境との対話から方策を学習できるため，厳密な教師ありデータが得にくい音声模倣の問題に適している．

調音合成システムによる母音模倣タスクに深層強化学習を適用した研究 [19] では，まず pink trombone のような調音合成システム上での母音から母音への遷移をタスクに設定し，エージェントが一つの母音から別の目標母音を合成するよう学習した．

さらに，調音パラメータ制御を強化学習を用いて end-to-end で学習する手法も提案されている [3]．この研究では，限られた音節集合とはいえ，強化学習のみによって調音パラメータ推定の方策の獲得に成功したことが示されている．またこの研究

(注1)：https://www.vocaltractlab.de/
(注2)：https://dood.al/pinktrombone/

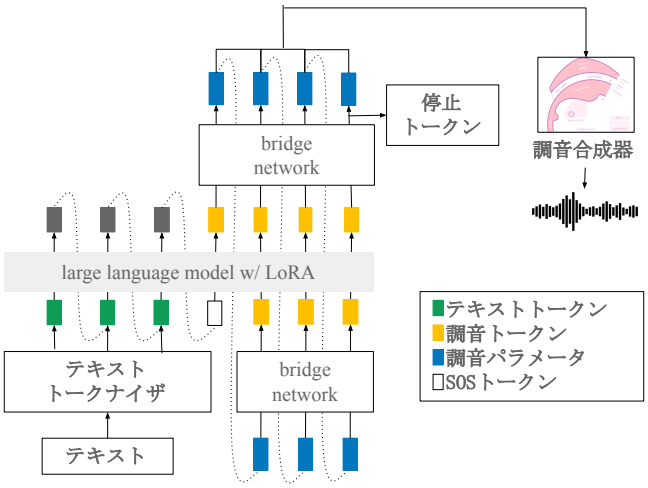

図 2: 提案するアーキテクチャ

では，大規模モデルに頼らず調音制御という解釈可能な中間表現を用いることで，音声生成プロセス全体の説明可能性が飛躍的に高まること，強化学習による調音パラメータ学習は人間の言語獲得メカニズムの模倣であることが強調されており，本研究の方向性と軌を一にする．

音声から調音パラメータを得る手法として強化学習を用いるアプローチは徐々に成果を上げつつある．本研究では，先行研究における音素・音節レベルの調音パラメータ推定からさらに発展させ，文章レベルのテキストから直接調音パラメータを推定する課題に取り組む．

3. 提 案 手 法

LLM を基盤としたニューラルネットワークを学習することにより，テキストから直接調音パラメータ系列を生成するアーキテクチャを提案する．提案するアーキテクチャを図 2に示す．

3.1 モデルの構造

3.1.1 LLM の利用

基盤モデルとして学習済み LLM を用いることで，その系列処理能力を活用する．モデルに入力としてテキストを与えると，LLM のデコーダが自己回帰的に調音パラメータに対応する連続ベクトル（埋め込み表現）を逐次出力する．そのベクトルを線形射影して調音パラメータに変換することで，各タイムステップで調音パラメータを出力する．調音パラメータは声道の物理モデル上での，各調音器官の形状・位置を定義し，声道の形状の表現を可能にすることで，声道から発せられる音声を合成できる．調音パラメータ出力学習のための LLM ファインチューニングには LoRA [20] を用いる．

3.1.2 停止トークンの予測

調音パラメータの出力と並行して停止トークンを予測する．停止トークンは tacotron 2 [21] にならい，現在の出力が系列の終了フレームである確率を表すものである．

3.2 学 習 手 法

3.2.1 教師あり事前学習

強化学習のみで学習した場合と，事前学習を行った場合での

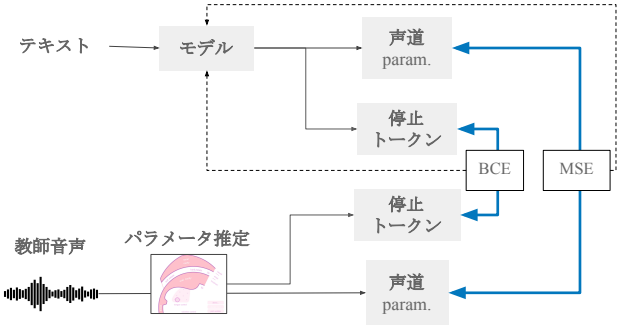

図 3: 事前学習のフロー

性能を比較するため，教師あり学習による事前学習を行う．事前学習のフローを図3に示す．テキストに対応する音声からパラメータ推定を行うことによって，テキストとそれに対応する調音パラメータ系列（および系列終了を示す停止トークンラベル）のペアからなる教師データセットを用意し，モデルに対して系列変換タスクの教師データを与える．モデルにはテキストを入力とし，正解のパラメータ系列を出力させるように学習させる．デコーダが予測したパラメータと正解との間で平均二乗誤差として損失を計算する．同様に，停止トークン予測については出力と教師データとの間で二値クロスエントロピー損失を与え，系列終了タイミングを学習させる．

学習初期は系列生成の学習を教師強制（teacher-forcing）によって行うが，モデルが十分に学習されてきた段階で，スケジュールドサンプリング（scheduled sampling）を行う [22]．学習の初期段階では常に教師強制により，直前の調音パラメータについての正解をモデルに与えていたものを，徐々にモデル自身が予測した出力を使用する割合を高めていく．このように学習過程を徐々に教師ありから自己回帰へ移行させることで，推論時の暴露バイアス（exposure bias） [23] による誤差蓄積問題を緩和し，モデルの頑健性向上が期待できる．十分な事前学習によってモデルはある程度正確にパラメータ系列を生成できるようになるため，その後の強化学習フェーズでは初期性能の低さによる学習の不安定化を避けつつ，報酬に沿った微調整を効率良く行うことが可能となる．

3.2.2 強 化 学 習

モデルによって出力された調音パラメータ系列に対して，強化学習アルゴリズムの一種である PPO（proximal policy optimization） [24] を用いて最適化を行う．強化学習時のフローを図 4に示す．先行研究に従い [3]，強化学習における報酬関数として，生成音声とターゲット音声（正解データ）の音節レベルの類似度を使用する．報酬の計算には，音声波形から音節を表現する埋め込みの系列を抽出できる訓練済みモデル sylber を用いる [25]．本手法ではまずモデル出力の調音パラメータ系列を VocalTrax で音声に変換し，sylber によってその合成音声とターゲット音声の双方から音節埋め込み系列を抽出する．次に，それら埋め込み系列間のコサイン類似度を総当たりで計算し，recall を報酬とする．ここで，sylber が合成音声から埋め込

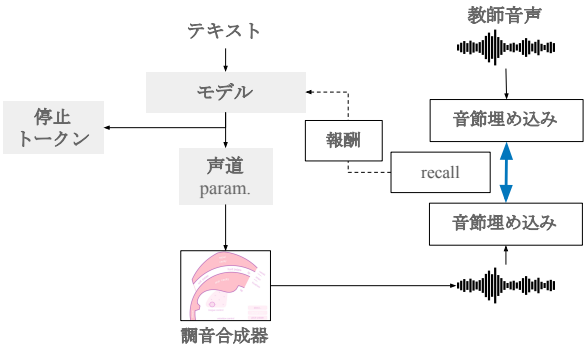

図 4: 強化学習のフロー

み系列を抽出できなかった場合は，先行研究 [3] に従い，−1 を報酬とする．このスコアは，合成音声がどれだけターゲット音声の音節系列に近いかを表す．報酬が増大するように方策を学習することで，モデルが出力するパラメータ系列は，教師音声の音響的特徴に近づくように学習される．

強化学習のみによる学習および，事前教師あり学習を行ったモデルに対しての強化学習を行う．

4. 実験的評価

4.1 実 験 設 定

データセット：事前学習済み LLM として，Swallow-7b-hf [26] を用いた．JSUT コーパス [27] の BASIC5000 より，1 発話 (BASIC5000_0001）を使用した．以降，この発話を ground truth と呼ぶ．事前教師あり学習のため，VocalTrax [2] を用いて ground truth から調音パラメータ系列を推定し，これを教師データとした．調音パラメータ系列推定のための VocalTrax のハイパーパラメータは，サンプリング周波数を 16000Hz とし，そのほかは公式実装(注3)において config に記載されている初期値にしたがった．VocalTrax が推定する調音パラメータは，唇の狭窄の度合いを表す lipconstriction，声道のうち声門にもっとも近い地点の狭窄の度合いを表す throatconstriction，舌先の位置を表す tongue index，舌先が位置する地点の狭窄の度合いを表す tongue diameter，緊張度を表す tense の 5 次元である．これらの調音パラメータには，基本周波数（$F0$）は含まれていないため，CREPE [28] を用いて $F0$ を推定し，独自に調音パラメータに加えた．また，この推定結果より，$F0$ が検出されない場合は 0，検出される場合は 1 をとる有声・無声フラグの系列を求め，調音パラメータに加える．学習の偏りを防ぐため，$F0$ の系列を線形補間した．以上 7 次元の系列を学習する調音パラメータとした．

また，音節単位での報酬を与えるために，sylber によって ground truth における音節の境界（開始時刻・終了時刻）および音節埋め込みを推定した．得られた境界に基づき，発話を時刻で切り出し，音節ごとの ground truth 波形，調音パラメータ系列，埋め込みを対応付けた．さらに，各時刻に対応する音素列を音素アラインメント（HTS ラベル形式）から取得することで，音節区間と時間的に重なるテキストを当該音節に対応するテキストとして用いた．

作成したデータセットを単音節データセットとし，それらのうち時系列上で連続する 2 つの音節について，対応する ground truth 波形，調音パラメータ系列，埋め込みを結合したものを 2 音節データセットとした．これら 2 つのデータセットを利用し，それぞれについて学習を行った．

モデル：bridge network では，4096 次元の LLM 隠れ層として出力された表現を 7 次元に線形射影することで調音パラメータとする．LoRA においては，r と α をともに 16 とした．ファインチューニング対象の層は q，k，v，o の各層とし，dropout は 0.05 とした．

強化学習：単音節もしくは 2 音節の生成を 1 エピソードとし，PPO アルゴリズムで学習を行った．報酬割引率 γ を 0.99，GAE パラメータ λ を 0.95，バッチサイズを 1，学習率を $3e-5$，1 エピソードあたりの PPO 更新反復回数を 4 とした．停止トークンを学習対象とすると，学習の初期に一定長の系列が生成されず，音節の推定が行われないために，報酬が変動せず，学習が行われない．報酬を変動させ学習を進めるために，停止トークンは学習せず系列長は教師データから与えた．学習の進度による性能比較を行うため，事前学習を行う場合と行わない場合のそれぞれについて，60 エポックの学習と 120 エポックの学習を行った．

教師あり学習：入力テキストから 7 次元の調音パラメータ系列を出力させる学習を行った．最適化には Adam を用いた．学習率を $3e-5$，重み減衰を $1e-2$，バッチサイズを 8 とした．最初のエポックから第 5 エポックまでは完全な教師強制で学習を行った後，自己回帰的な生成の割合を第 6 エポックから第 30 エポックまで線形に増加させ，第 31 エポック以降は自己回帰的な生成の割合を 0.7 で固定して学習を進める．

損失関数は式3に表されるように，系列の MSE および停止トークンの BCE の重み付き和とした．ここで，$\boldsymbol{s}$ および $\hat{\boldsymbol{s}}$ は各フレームにおいて停止トークンが出力される確率の教師データ（停止トークンラベル）およびモデル出力，$\boldsymbol{v}$ および $\hat{\boldsymbol{v}}$ は各フレームにおける調音パラメータの教師データおよびモデル出力，$\mathrm{num}_{\mathrm{pos}}$ および $\mathrm{num}_{\mathrm{neg}}$ は教師データにおいて停止トークンラベルが 1 であるフレームの数および 0 であるフレームの数である．

$$\mathrm{loss}_{\mathrm{pos}} = \mathrm{BCE}(\mathbf{s}, \hat{\mathbf{s}}) \cdot \mathbf{s} \tag{1}$$

$$\mathrm{loss}_{\mathrm{neg}} = \mathrm{BCE}(\mathbf{s}, \hat{\mathbf{s}}) \cdot (\mathbf{1} - \mathbf{s}) \tag{2}$$

$$\mathrm{loss} = \mathrm{MSE}(\mathbf{v}, \hat{\mathbf{v}}) + \left(\frac{\mathrm{loss}_{\mathrm{pos}}}{\mathrm{num}_{\mathrm{pos}}} + \frac{\mathrm{loss}_{\mathrm{neg}}}{\mathrm{num}_{\mathrm{neg}}} \right) \tag{3}$$

損失関数が十分に収束するまで学習を行ったところ，単音節では 60 エポック，2 音節では 70 エポックを要した．

4.2 評　　価

音響的類似度，話者類似度．定性的評価の 3 観点を用いて，分析合成音（VocalTrax 推定パラメータで再合成）を ground truth と比較して評価したほか，モデル合成音（本モデルの学習データセットに含まれるテキストに対し，本モデルが生成したパラ

(注3)：https://github.com/PapayaResearch/vocaltrax

メータで合成）と ground truth，分析合成音とモデル合成音をそれぞれ比較した．なお，強化学習のみおよび事前学習あり強化学習によるモデルを用いた音声合成については，停止トークンの推論を行わず，系列長を教師パラメータより与えた．

音響的類似度：LSD で ground truth と合成音とのスペクトル間の距離，および分析合成音とモデル合成音とのスペクトル間の距離を評価した．librosa を用いて音声波形の対数パワースペクトル間距離をフレームごとに計算し，フレーム間の平均をとった．波形の時間長が一致しない場合は，ゼロパディングを行ったうえで算出した．フレーム t のパワースペクトルを $P_t(\omega)$，対応する比較対象を $\hat{P}_t(\omega)$ とすると，LSD は式 (4) のようにあらわされる．

$$\mathrm{LSD} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\log P_t(\omega) - \log \hat{P}_t(\omega)\right]^2 d\omega\right)^{\frac{1}{2}} \quad (4)$$

さらに，MCD で ground truth と合成音とのメルケプストラム系列間の距離，および分析合成音とモデル合成音とのメルケプストラム系列間の距離を評価した．オープンソースの計算ツールを使用した(注4)．メルケプストラム係数（0 次を除く）$\mathbf{c}_t \in \mathbb{R}^K$ と $\hat{\mathbf{c}}_t$ に対し，MCD は式 (5) のようにあらわされる．

$$\mathrm{MCD} = \frac{10}{\log 10}\cdot\frac{1}{T}\sum_{t=1}^{T}\sqrt{2\sum_{k=1}^{K}(c_{t,k}-\hat{c}_{t,k})^2} \quad (5)$$

話者類似度：話者埋め込み抽出モデル resemblyzer(注5) で抽出した埋め込みのコサイン類似度により，話者性の保存度を測った．話者埋め込みを $\mathbf{v}(\cdot)$ とすると，話者類似度は式 (6) のようにあらわされる．

$$\mathrm{SpkSim}(x,\hat{x}) = \frac{\mathbf{v}(x)^\top \mathbf{v}(\hat{x})}{\|\mathbf{v}(x)\|\|\mathbf{v}(\hat{x})\|} \quad (6)$$

定性的評価：音節ごとに ground truth，分析合成音，モデル合成音を聴取し，言語的内容の再現（母音脱落・子音の欠落・音節の重複），音高の知覚的なずれを中心に観察した．

その他：学習時の各エピソードにおける報酬および，各エポックにおける報酬の平均を観察した．

4.3 結果と考察

本節では，データセットに用いた音節長（単音節/2 音節）ごとに，教師あり学習，強化学習，事前学習あり強化学習を比較する．

4.3.1 考察の観点

考察の観点は以下である：

事前教師あり学習のための教師データは適切か：教師あり事前学習に用いた教師パラメータは，学習の土台として妥当か．

強化学習のみで，どれだけ学習できるか：教師あり事前学習を行わず，強化学習だけで調音パラメータ系列をどの程度まで獲得できるか．

強化学習のエポック数はどの程度が適切か：強化学習の収束エポック数はどの程度か．

(注4)：https://github.com/Takaaki-Saeki/DiscreteSpeechMetrics
(注5)：https://github.com/resemble-ai/Resemblyzer

表 1: ground truth との比較

音節数	教師あり学習	強化学習	強化学習エポック数	LSD	MCD	話者類似度
	分析合成音			32.323	7.693	0.968
	✓			33.244	8.116	0.952
1		✓	60	36.038	15.380	0.959
		✓	120	34.936	14.607	0.963
	✓	✓	60	33.119	10.085	0.953
	✓	✓	120	33.268	11.289	0.963
	分析合成音			31.951	7.792	0.806
	✓			32.182	9.102	0.796
2		✓	60	32.717	12.389	0.816
		✓	120	32.495	11.858	0.779
	✓	✓	60	32.253	9.632	0.816
	✓	✓	120	34.342	17.699	0.767

表 2: 分析合成音との比較

音節数	教師あり学習	強化学習	強化学習エポック数	LSD	MCD	話者類似度
	✓			12.399	4.750	0.980
		✓	60	24.779	16.612	0.988
1		✓	120	23.095	16.649	0.990
	✓	✓	60	17.055	11.377	0.982
	✓	✓	120	21.854	13.762	0.974
	✓			14.560	8.607	0.898
		✓	60	18.623	13.923	0.865
2		✓	120	20.701	12.714	0.829
	✓	✓	60	16.104	8.689	0.897
	✓	✓	120	30.368	18.980	0.814

データセットの長さとして単音節および 2 音節は適切か：単音節および 2 音節というごく短い時間長を持つデータセットで，安定に学習が可能か．

4.3.2 結果と考察

評価指標の数値を表 1，表 2に示す．

事前教師あり学習のための教師データは適切か：分析合成音と ground truth の客観指標の差が大きく，教師パラメータ自体が ground truth から乖離している可能性が示唆された．この場合，教師あり学習で得られる初期モデルは教師信号の再現には向くが，実音声に整合する調音制御の獲得には不利になり得る．

強化学習のみで，どれだけ学習できるか：単音節における事前学習なし強化学習と事前学習あり強化学習を比較すると，事前学習は各指標を若干改善させたが，大幅な改善ではなかった．このことから，ある程度強化学習のみで調音パラメータを学習できることが示唆された．しかし，強化学習における指標が教師あり学習を上回る傾向は確認できなかった．具体的には，強化学習による合成音のなかで，有声と推論すべきところを無声と推論していたり，$F0$ が ground truth より非常に高くなっていたりするサンプルが存在した．本研究では，sylber 埋め込みに基づく類似度を主な報酬としているため，音響類似度や話者類

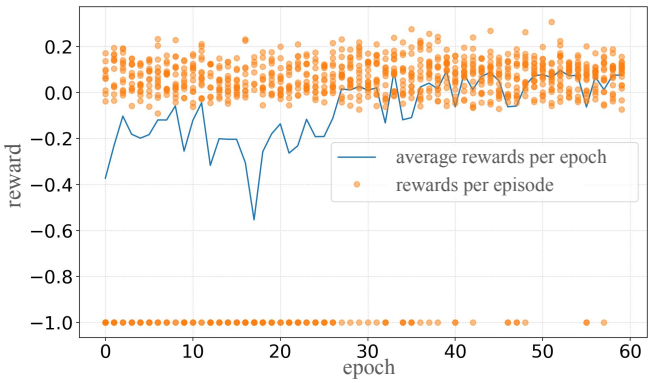

図 5: 単音節データセットを用いて，事前学習なし強化学習を 60 エポック行った際の，各エピソードにおける報酬および，各エポックにおける報酬の平均の推移

似度などの客観指標と，最適化すべき目標が一致しない可能性がある．

強化学習のエポック数はどの程度が適切か：エポック数以外の条件が同じとき，60 エポックと 120 エポックにおいて各指標を比較すると，大きな改善が見られなかった．図5に示すように，強化学習の初期において，報酬が −1 となるエピソードが存在する．これは，そのエピソードにおいて，音節埋め込みを検出できなかったことを示す．しかし，遅くとも 60 エポックまでに，全音節について音節埋め込みが安定に抽出できるようになる．その後は報酬が収束し，学習が進みにくくなる．さらなる性能向上のためには，報酬設計を変更するなどの解決策が考えられる．

データセットの長さとして単音節および 2 音節は適切か：単音節で学習した場合では sylber 埋め込みが安定に抽出できないケースがあり，また LSD や MCD が 2 音節で学習した場合より悪化した．本研究では ground truth の時間長を 2 音節程度まで長くすることで，埋め込み抽出と学習が安定し，評価指標も解釈しやすくなる可能性が高い．

5. ま と め

本研究では，人間の言語獲得を模倣した強化学習による調音器官を介した音声合成手法を提案した．1 発話の単音節単位および 2 音節単位という条件下で，提案アーキテクチャは調音パラメータ系列を学習可能であった．教師あり学習は教師データへの近似を達成するが，モデル合成音は教師パラメータの妥当性に強く依存することから，教師あり学習によるさらなる性能向上のためには教師パラメータをより ground truth を再現したものにする必要性が示唆された．sylber 埋め込みに基づく強化学習は，ground truth への接近を意図した設計と整合する挙動を示すが，学習時の報酬の停滞や，音響的距離や有声・無声の推論の正確性に課題が残った．

謝辞：本研究は，JST 創発的研究支援事業 JPMJFR226V の支援を受けて実施した．

文 献

[1] Chenshuang Zhang et al. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI, 2023.
[2] Luke Mo et al. Articulatory Synthesis of Speech and Diverse Vocal Sounds via Optimization. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.
[3] Akshay Anand et al. Teaching Machines to Speak Using Articulatory Control, 2025.
[4] Detai Xin et al. RALL-E: Robust Codec Language Modeling with Chain-of-Thought Prompting for Text-to-Speech Synthesis, 2024.
[5] Catherine Anderson. *2.2 Articulators*. eCampusOntario, 2022.
[6] Michael H. Goldstein and Jennifer A. Schwade. Social Feedback to Infants' Babbling Facilitates Rapid Phonological Learning. *Psychological Science*, 19(5):515–523, 2008. PMID: 18466414.
[7] Entong Su et al. Sim2Real Manipulation on Unknown Objects with Tactile-based Reinforcement Learning, 2024.
[8] Jaehyeon Kim et al. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *ICML 2021*, 2021.
[9] J. L. Kelly and C. C. Lochbaum. Speech synthesis. In *Proceedings of the Fourth International Congress on Acoustics, Copenhagen, pp. 1-4, September 1962*, 1962.
[10] P. Mermelstein. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, 04 1973.
[11] Shinji Maeda. *Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model*, pages 131–149. Springer Netherlands, Dordrecht, 1990.
[12] Peter Birkholz. Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. *PLOS ONE*, 8(4):1–17, 04 2013.
[13] Mateo et al. Cámara. Parameter optimisation for a physical model of the vocal system. *EURASIP J. Audio Speech Music Process.*, 2025(1), July 2025.
[14] David Südholt et al. Vocal Tract Area Estimation by Gradient Descent, 2023.
[15] Leena G Pillai and D. Muhammad Noorul Mubarak. Acoustic to Articulatory Inversion of Speech; Data Driven Approaches, Challenges, Applications, and Future Scope, 2025.
[16] O. Fujimura. Some Remarks on the Analysis-by-Synthesis as a Model of Speech Perception. *STUF - Language Typology and Universals*, 21(1-6):48–52, 1968.
[17] B. S. Atal et al. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. In *The Journal of the Acoustical Society of America, 63(5), 1535–1553*, 1978.
[18] Yingming Gao et al. Articulatory Copy Synthesis Based on a Genetic Algorithm. In *Interspeech*, pages 3770–3774, 09 2019.
[19] Denis Shitov et al. Deep Reinforcement Learning for Articulatory Synthesis in a Vowel-to-Vowel Imitation Task. *Sensors*, 23(7), 2023.
[20] Edward J. Hu et al. LoRA: Low-Rank Adaptation of Large Language Models, 2021.
[21] Jonathan Shen et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, 2018.
[22] Samy Bengio et al. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks, 2015.
[23] Kushal Arora et al. Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation, 2023.
[24] John Schulman et al. Proximal Policy Optimization Algorithms, 2017.
[25] Cheol Jun Cho et al. Sylber: Syllabic Embedding Representation of Speech from Raw Audio. In *ICLR*, 2025.
[26] 藤井一喜 et al. 継続事前学習による日本語に強い大規模言語モデルの構築. **言語処理学会第** *30* **回年次大会**, 3 2024.
[27] Ryosuke Sonobe et al. JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, 2017.
[28] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A Convolutional Representation for Pitch Estimation, 2018.