

# Auto Oshaberi: audio 2 speech.

Dreamerを応用した発声器官モデルによる模倣音声生成手法

下村 晏弘

仙台高等専門学校

石平 雄作

東京理科大学

堀 彰悟

石川工業高等専門学校

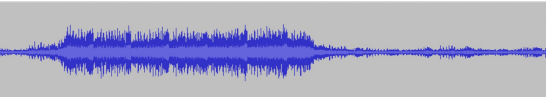
竹下 隼司

東京理科大学

## 1. 研究背景・目的

### 背景

対話AIシステムのために、任意の音響データを「声」に。

Ex. せみの鳴き声は  だよ ← Oh, Shocking...

→ せみの鳴き声は「みーんみんな」だよ ← OK!

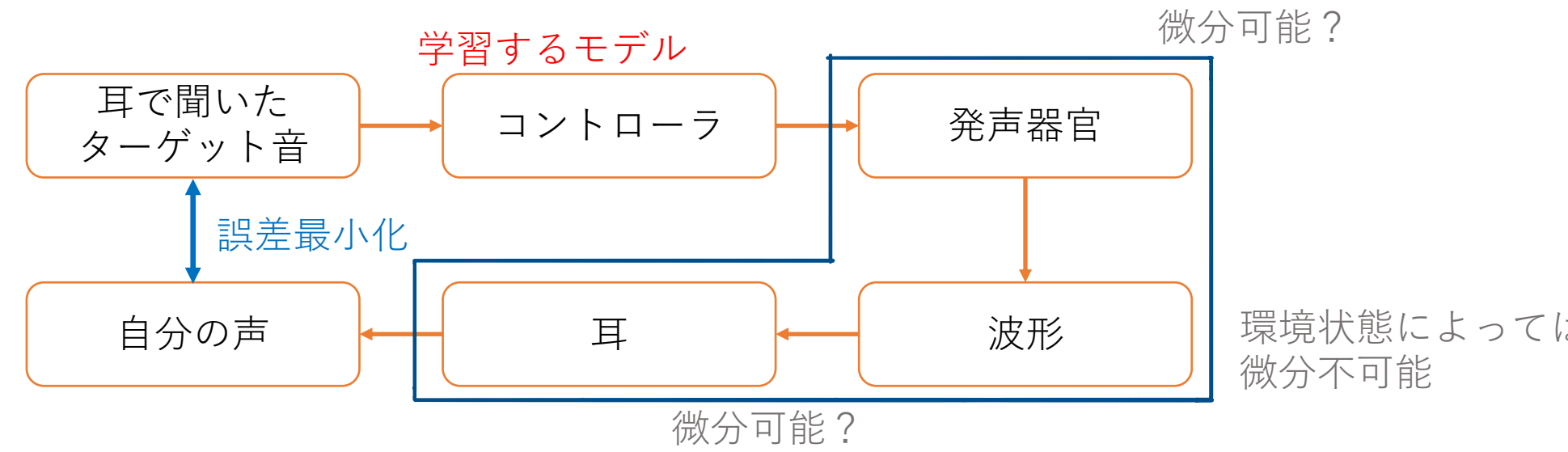
発声器官という制約された波形生成装置を使えば全て解決ではないか。

### 目的

環境音まで含めた任意音を「口」を使って模倣する

## 2. コンセプト

### 模倣のコンセプト



問題: 青枠部分の計算グラフが繋がらない

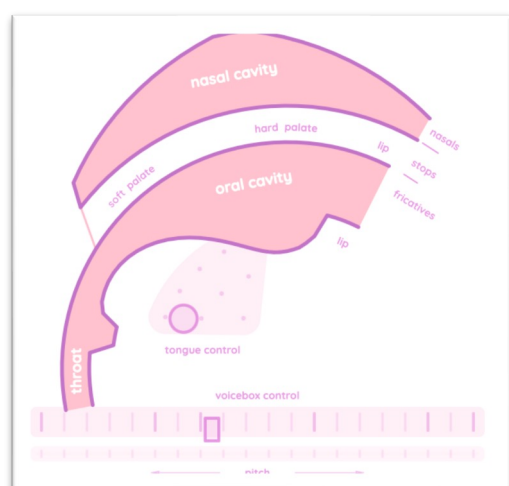
解法: 世界モデルで近似する

## 3. アプローチ

声道モデル「PinkTrombone」：円筒列によって声道を近似

操作項目 (行動 $a_t$ )	説明
Pitch shift	音程調節
Tenseness	声のかすれ具合
Trachea	気管となる円筒の直径の大きさ
Epiglottis	喉頭蓋となる円筒の直径の大きさ
Velum	軟口蓋となる円筒の直径の大きさ
Tongue index	舌となる円筒の位置
Tongue diameter	舌となる円筒の直径の大きさ
Lips	口先の円筒の直径の大きさ

発声器官の状態 $v_t$	説明
Frequency	声門の周波数
Pitch shift	音程調節
Tenseness	声のかすれ具合
Tract diameters	声道円筒の直径列
Nose diameters	鼻腔円筒の直径列



その他の観測:

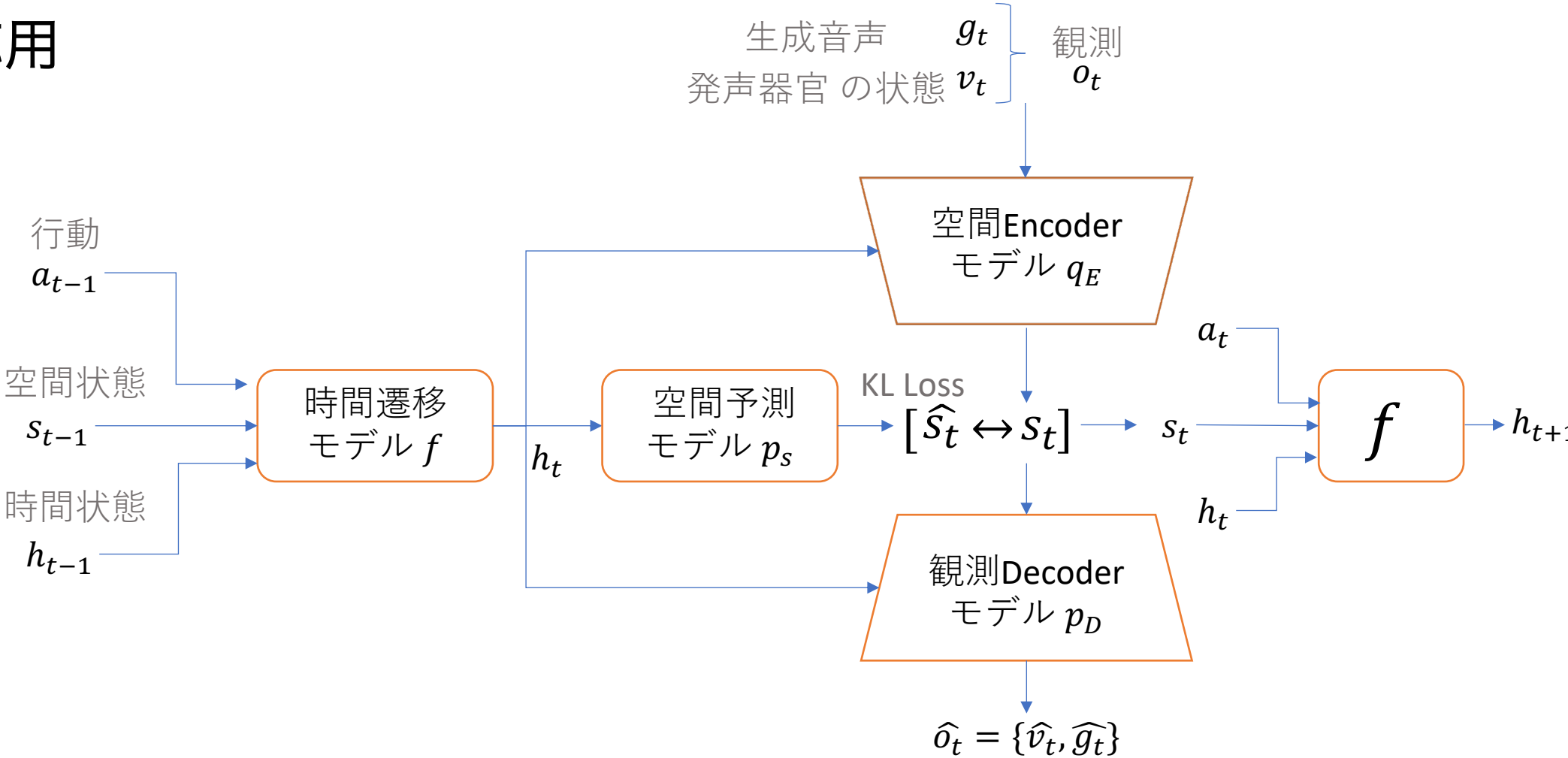
生成音声  $g_t$ , 次のステップのターゲット音声  $\tau_{t+1}$

世界モデル: Dreamerを応用

行動に対して次にどのような  
音声が生産されるかをモデル化  
行動 → 音声フィードバック

最小化する損失

$$\mathcal{L} = \|\hat{o}_t - o_t\| + D_{KL}[q_E \| p_S]$$



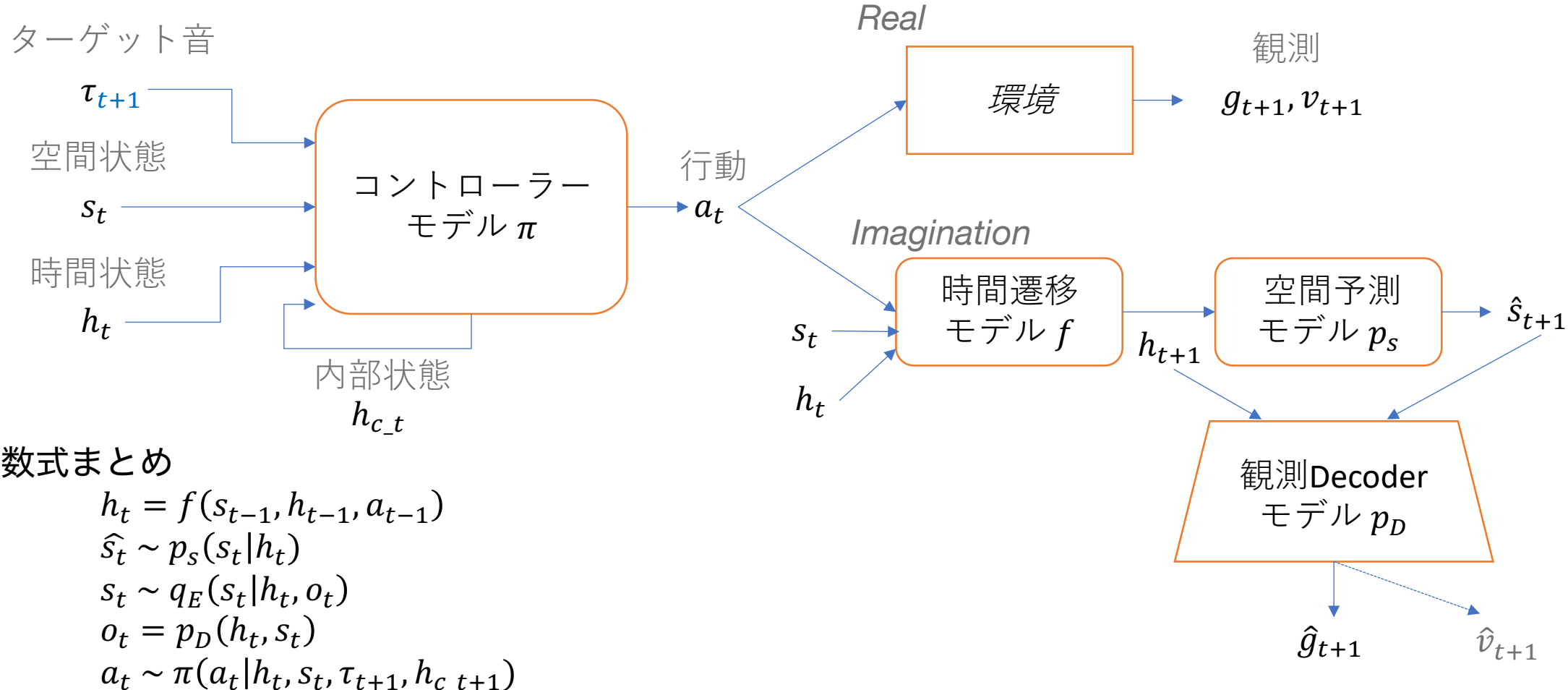
発声器官のコントローラモデル

世界モデルを通して  
再構成誤差の勾配を計算。

その勾配を用いて  
コントローラモデルを学習

最小化する損失

$$\mathcal{L} = \|\tau_{t+1} - \hat{g}_{t+1}\|$$



数式まとめ  
 $h_t = f(s_{t-1}, h_{t-1}, a_{t-1})$   
 $\hat{s}_t \sim p_S(s_t | h_t)$   
 $s_t \sim q_E(s_t | h_t, o_t)$   
 $o_t = p_D(h_t, s_t)$   
 $a_t \sim \pi(a_t | h_t, s_t, \tau_{t+1}, h_{c,t+1})$

学習: Dreamerの手法を応用

声道モデルを含む環境とインタラクションし、世界モデルの学習用データを収集。

学習された世界モデルにより生成された軌道を通し、コントローラモデルを学習。

上記のステップを繰り返す。

## 4. 実験

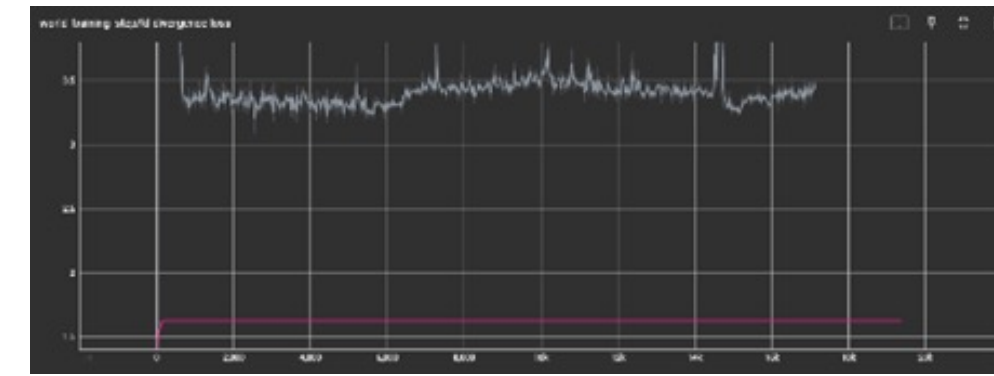
### 実験設定

学習・評価には声道モデルを用いて一つあたり2~3秒ほどのランダムに生成した音声データを約10時間分生成したものを用いた。  
本研究ではパラメータ数の異なる2つのモデルで学習を行った。各コンポーネントのパラメータ数を下表に示す。

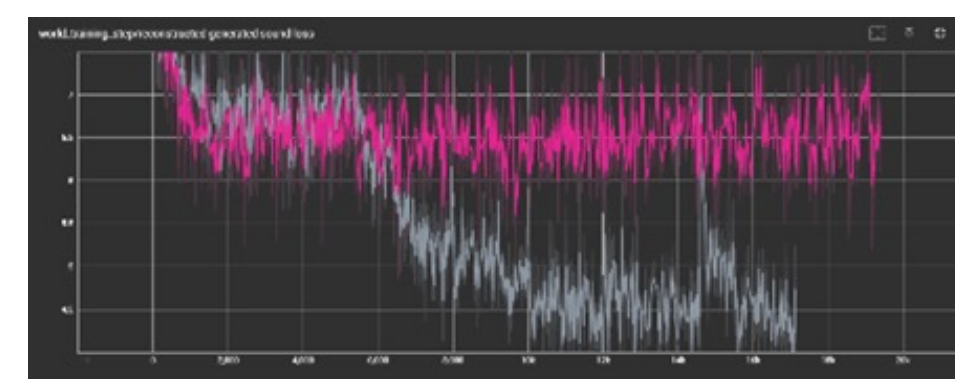
size	transition	prior	posterior	decoder	controller
small	857K	592K	4.95M	36.5M	2.80M
large	1.84M	1.58M	18.3M	37.5M	3.78M

### 結果

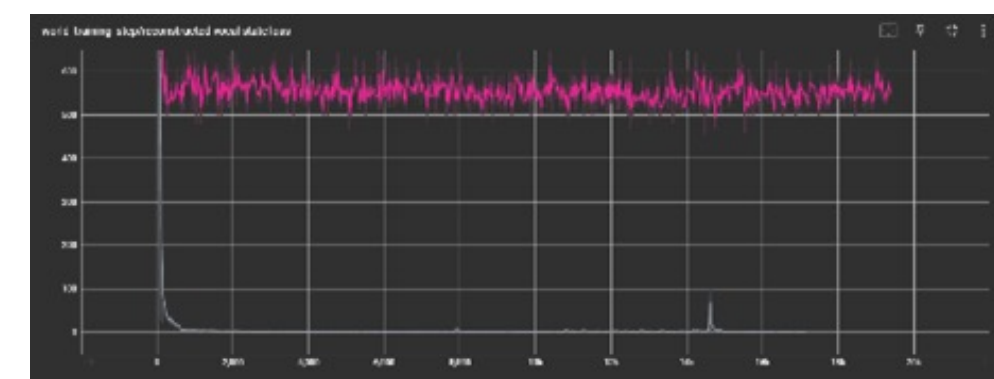
右の3つのグラフにおいてピンク線がsize large, グレー線がsize small にあたる。  
パラメータ数を増やすと再構成誤差が大きくなり、kl誤差は小さくなった。



kl誤差の推移

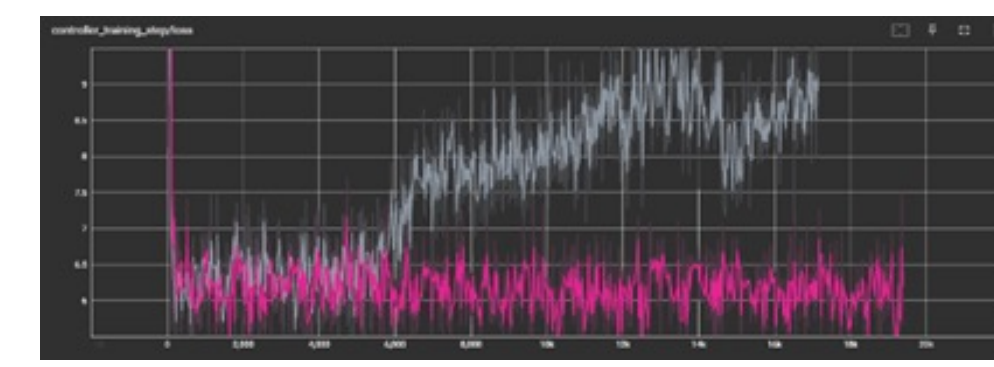


音声の再構成誤差



声道の状態の再構成誤差

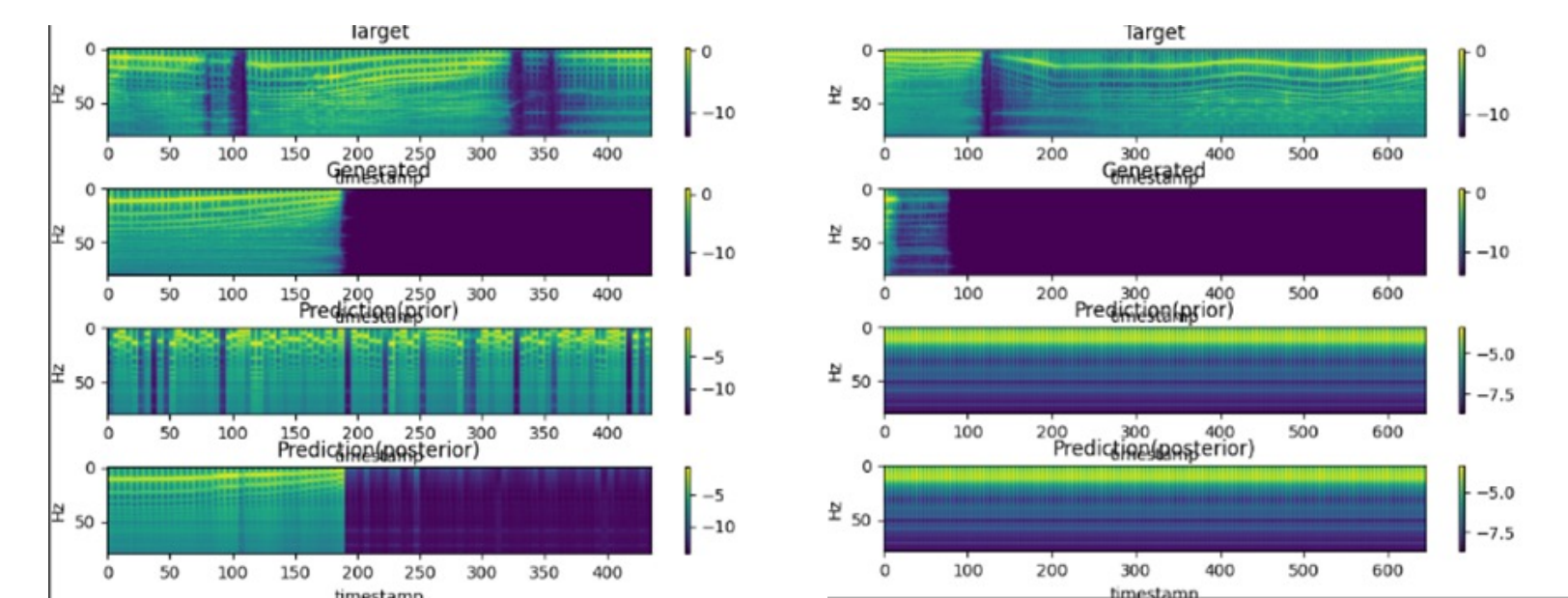
一方で、パラメータ数を増やすと右図に示すように  
コントローラモデルの損失は小さくなった。



パラメータ数ごとに、ターゲット音声と生成音声のメルスペクトログラムを可視化すると下図のようになる。

上から、ターゲット音声、生成音声、世界モデルのPosterior ( $q_E \rightarrow p_D$ )の予測、Prior ( $f \rightarrow p_S \rightarrow p_D$ )の予測

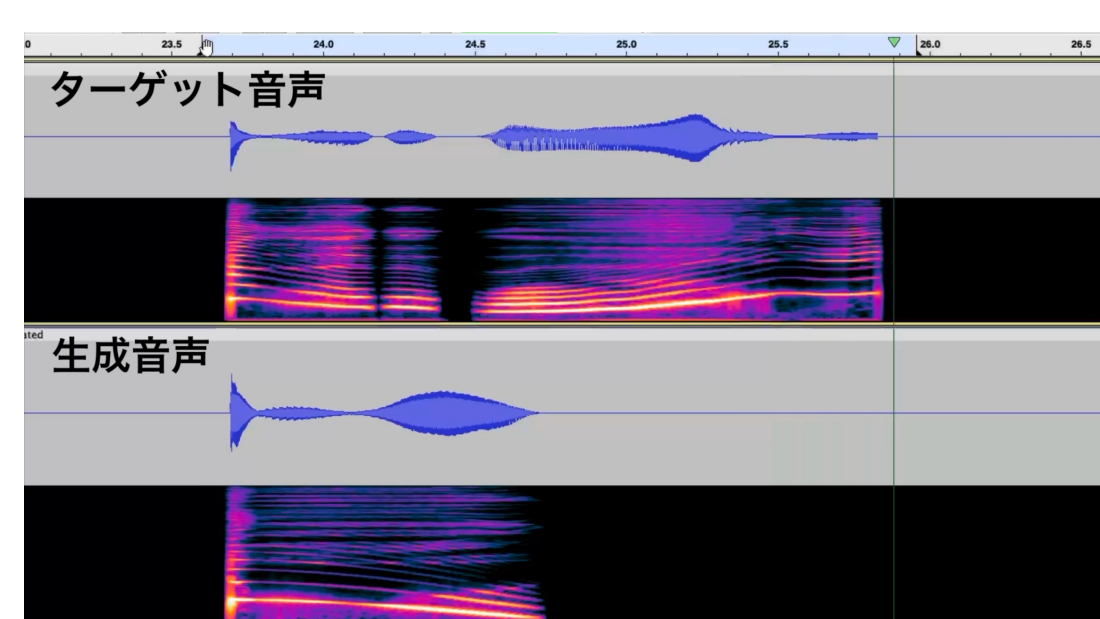
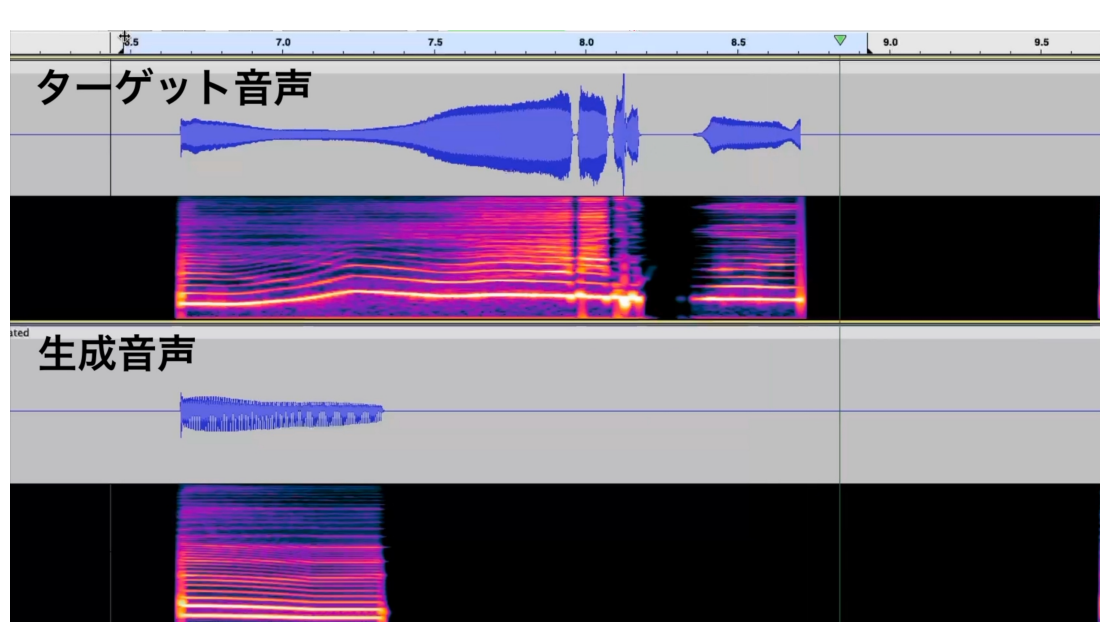
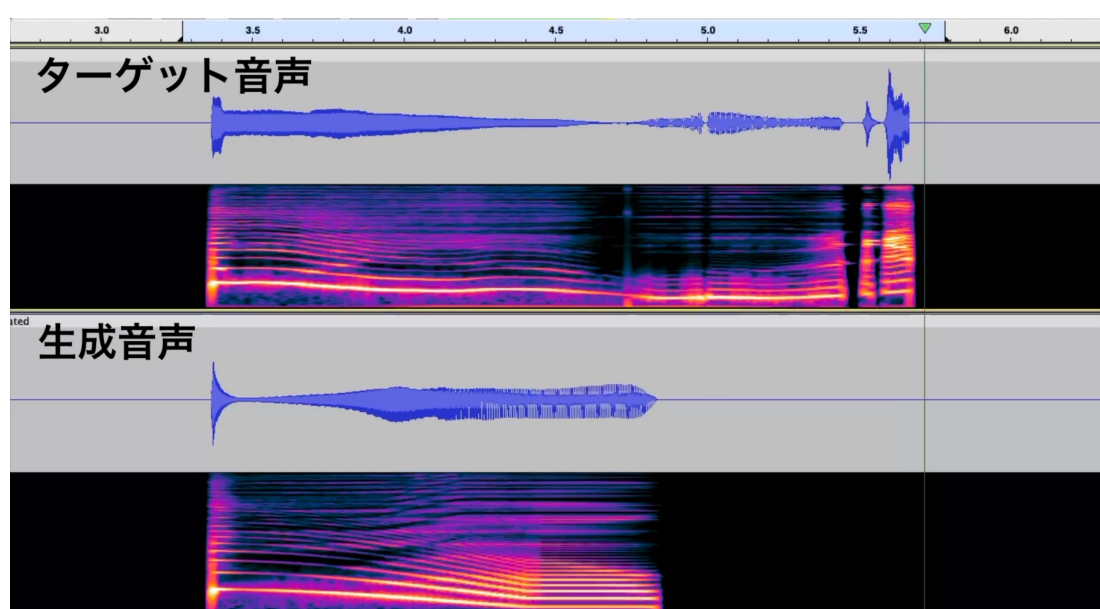
ターゲット音声に対して生成音声が生途中で消え、無音を生成した。



Size small のメルスペクトログラム

Size large のメルスペクトログラム

## 4. 1. 生成サンプル



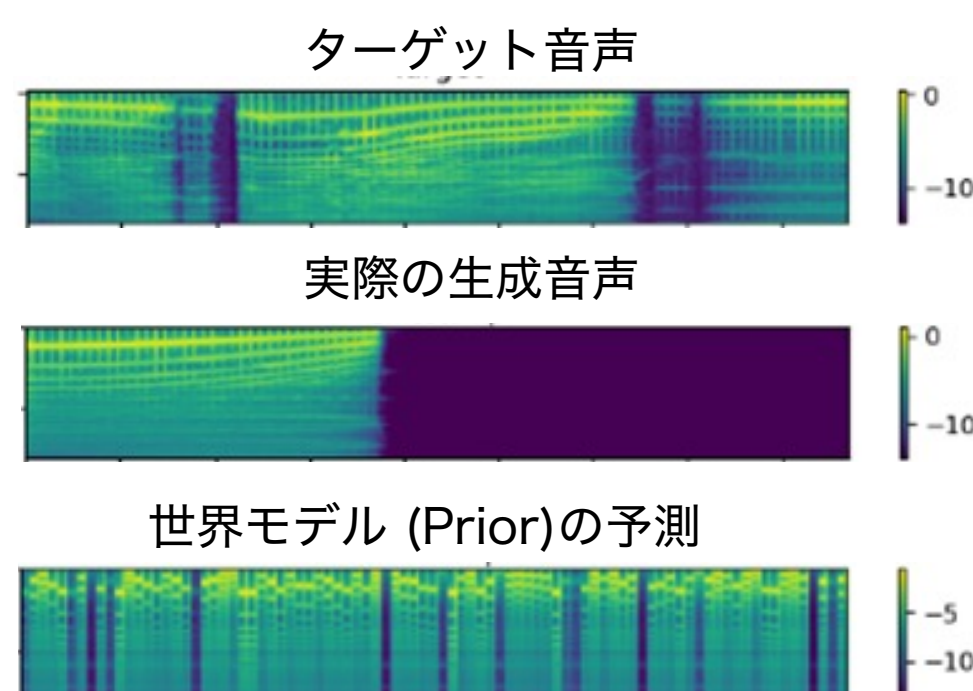
## 5. 考察

### 世界モデルの学習

実際の生成過程と比べて  
予測誤差が大きい

生成過程のモデル化に失敗

→ 行動から発声器官に影響が出るまでのステップ数が大きく、  
長期時系列をモデル化することが困難  
→ モデルの表現力が足りない



### コントローラモデルの学習

世界モデルの学習に失敗

→ コントローラモデルの学習に失敗

無音を生成するように収束

→ スペクトログラム上の誤差は無音の方が小さくなりやすい。  
音響特性をうまく表現しきれていない

### スケールに失敗

パラメータ数の増加に対し、  
訓練誤差が増加

→ モデルに残差接続のようなスケール性を  
持たせる仕組みがないためか。

## 6. 改善案 と 今後の課題

### 改善案

- 残差接続を用いてモデルのスケール性を高め、世界モデルの表現力を向上させる
- 長期系列のモデル化に強い手法を採用する
- 深層モデルの潜在空間といった、より音響特性が表現されている特徴量空間で誤差をとる

### 今後の課題

本研究は先行研究からの差分が大きく、結局の何が主要な失敗の原因であったのか結論づけることが困難であった。

先行研究の再現実装や、それに対して徐々に世界モデル的アプローチを増やしていくことによって着実に手法がうまくいくように努めていきたい。