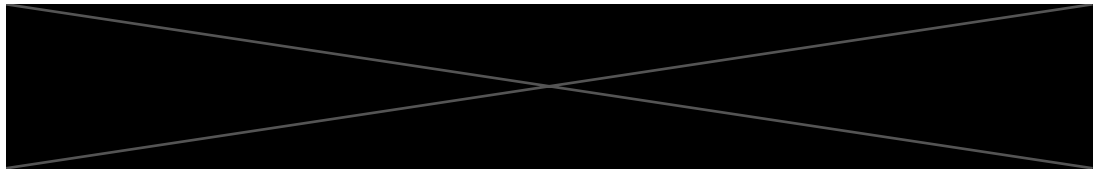
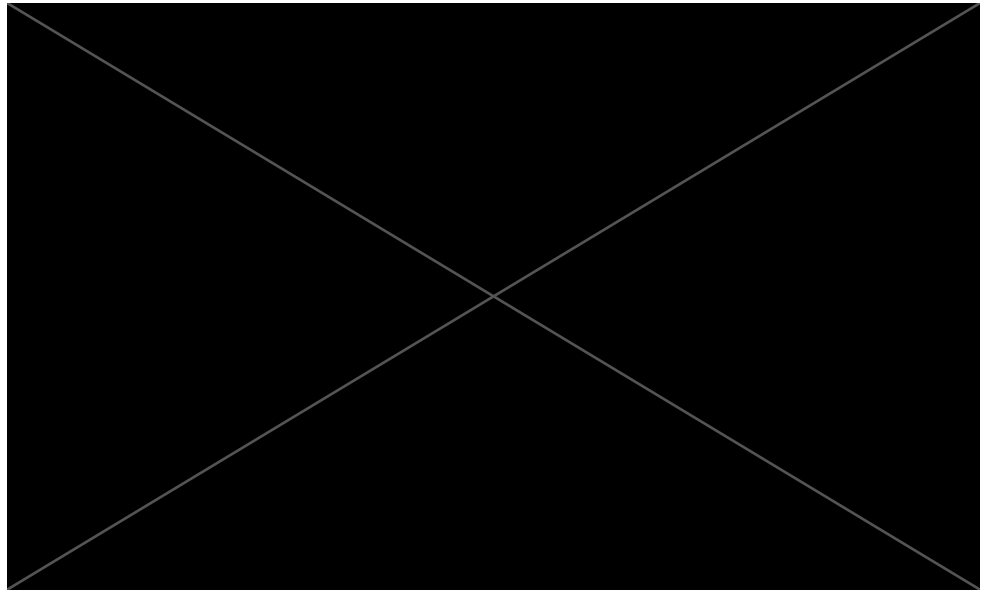


CycleGANを用いた
リアルタイム声質変換システムの開発

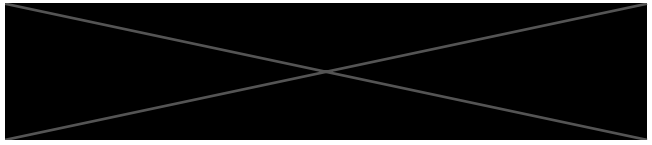


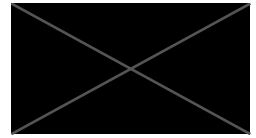
1．はじめに – 背景と目的 –

- 声質変換とは，ある話者の声を目的の話者の声に変換する技術．
  - 使用例）変換した音声を会話に利用する
- 求められることは，容易に声質変換が使用できること，リアルタイム性．
- 容易性は，学習データの収集が容易なCycleGANを用いた．
  - 機械学習を用いた多くの声質変換では，パラレルデータを用いる．
  - パラレルデータとは，入力話者と目的話者が全く同じ特徴とタイミングで読み上げたデータのこと

発表の流れ

1. はじめに – 背景と目的 –
2. システム概要
3. 学習概要
4. リアルタイム性の評価と結果
5. 声質変換の評価と結果
6. むすび





2．システム概要

- 振幅スペクトグラムを学習済みの生成器に入力.
- 声質変換された振幅スペクトログラムを位相復元し音声へ復元.

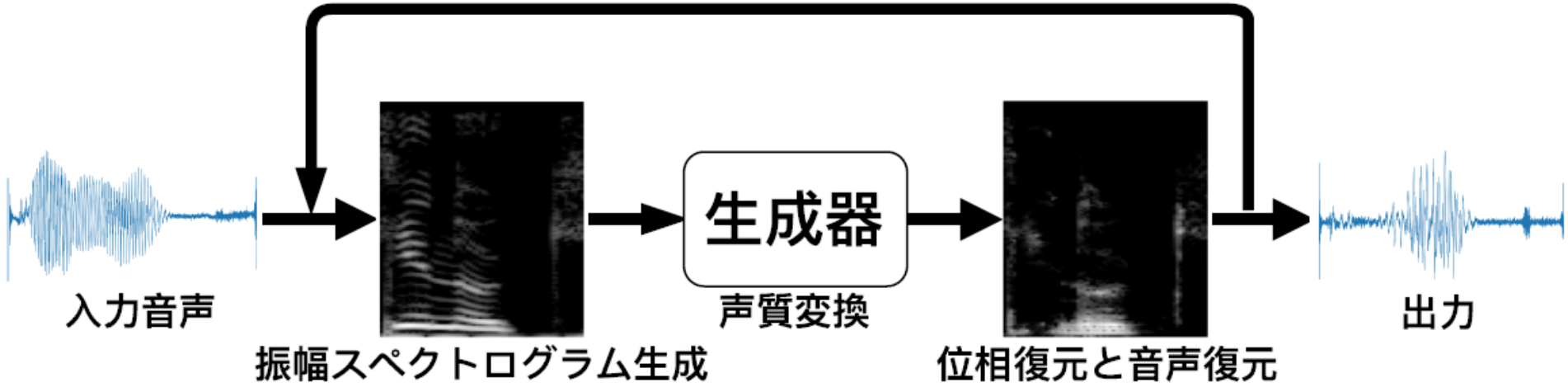


Figure 1: リアルタイム声質変換システムの概要

3．学習概要

- 全結合層で周波数の特徴を，畳み込み層で全体の特徴を学習．
- 畳み込み層で画像のサイズに出力．

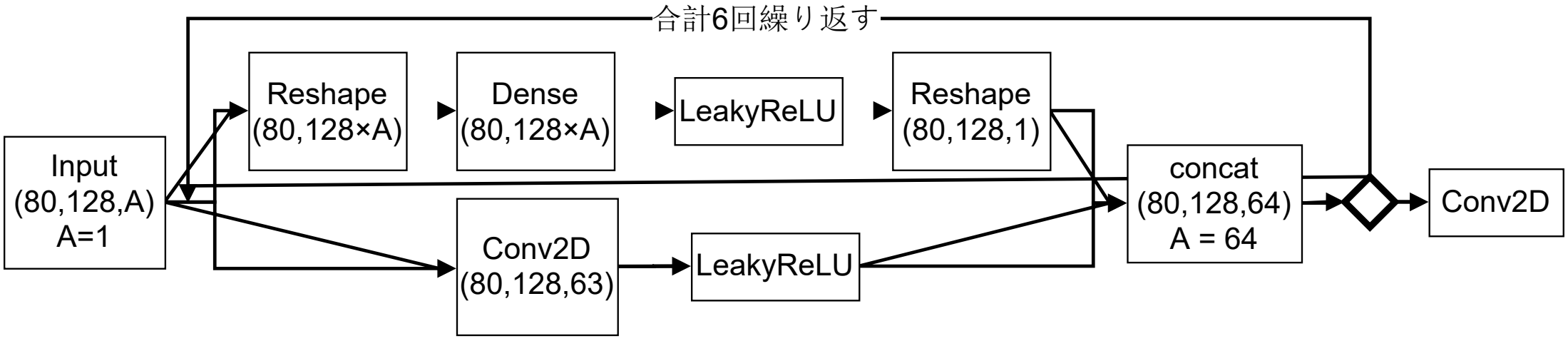


Figure 2: 生成器の構造

4．リアルタイム性の評価と結果

- リアルタイム声質変換システムは，入力音声が0.32秒のため，変換処理でそれ以上かかると音声が途切れる．
- 処理時間の大半は，位相復元法のGriffin-Lim法．他は約0.01秒．
- Griffin-Lim 法の反復回数とそれにかかる時間を実験によって求めた．
  - 実験は100×254pixelsの振幅スペクトログラムを100個用意．
  - 各反復回数で100回の位相復元を行い，平均秒数と最大秒数を計測．
- 反復回数70回が平均秒数，最大秒数ともに0.3秒以下であった．
- タイムラグは最大0.608秒であることがわかった．

5．声質変換の評価と結果

- リアルタイム声質変換システムの評価をアンケートによって行った．
  - 目的話者とリアルタイム声質変換システムを用いて小生（吹山）の声を目的話者の声に変換した音声を聴き比べ，アンケートを行う．
  - その際に，目的話者が男性と女性の場合で行う．
- 目的話者が男性と女性のどちらの場合も，半数以上の「やや似ている」以上の評価を得ることができた．

6．むすび

- CycleGANを用いたリアルタイム声質変換システムを開発した．
- リアルタイム性の評価は，最大0.608秒のタイムラグが発生することがわかった．
- 声質変換の評価は，アンケートを行った結果半数以上の「やや似ている」以上の評価を得ることができた．
- 読み込む音声の長さを短くすることや処理時間を短縮することで更にリアルタイムに変換を行うことができる．