

深層学習全学横断研究会

Road to Detective Conan

荒川 陸

東京大学大学院情報理工学系研究科
システム情報学専攻 稲見檜山研究室 修士1年

2019/10/16

自己紹介

- 荒川 陸 (@rikky0611)
- 東大情報理工修士1年
- 稲見檜山研究室
- 興味
 - 機械学習を活用した,
人間中心のインタラクションシステムの実装.



発表内容は, 2018年10月-2019年3月に東京大学工学部計数工学科システム情報工学コース 猿渡小山研究室で行った卒業研究の内容である.

名探偵コナンといえば



画像: <http://www.ytv.co.jp/conan/item/tai.html>

名探偵コナンといえば

- 蝶ネクタイ型変声器



名探偵コナンといえば

- 蝶ネクタイ型変声器
 - (1) リアルタイムのDNN音声変換
 - (2) 変換音声の品質向上
 - (3) デバイス化



コラボイベント展示



合同会社DMM.com との共同開発

出典: <https://prtimes.jp/main/html/rd/p/000003399.000002581.html>

動画: https://www.youtube.com/watch?v=vFSHxn_G2iQ

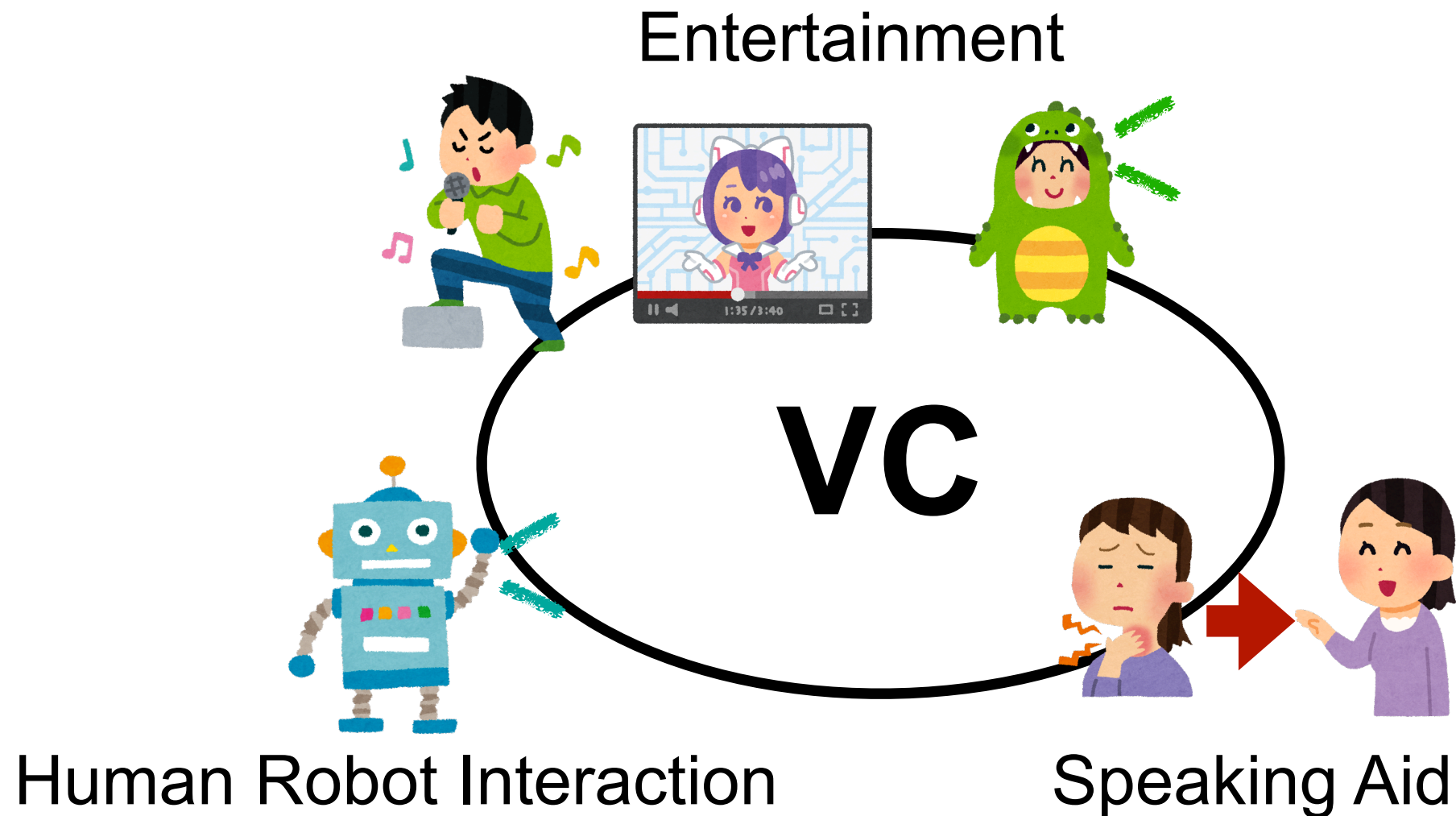
名探偵コナンといえば

- 蝶ネクタイ型変声器
 - (1) リアルタイムのDNN音声変換
 - (2) 変換音声の品質向上
 - (3) デバイス化



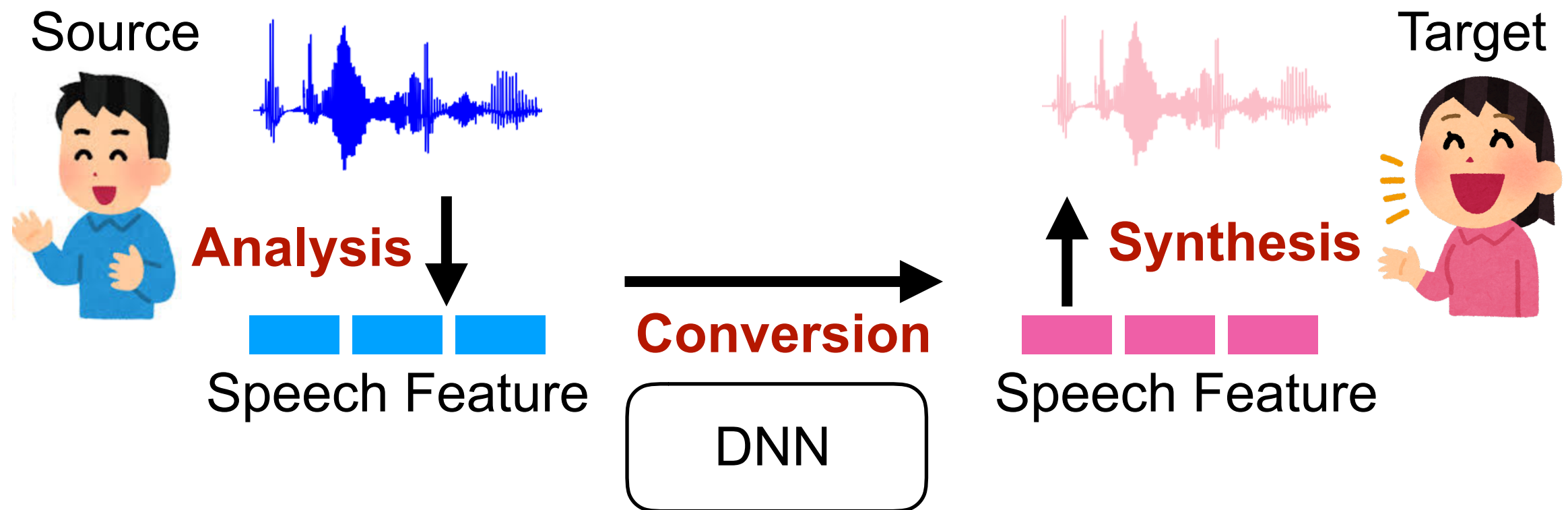
(1/3) リアルタイムDNN音声変換

研究テーマ: 音声変換 (VC)

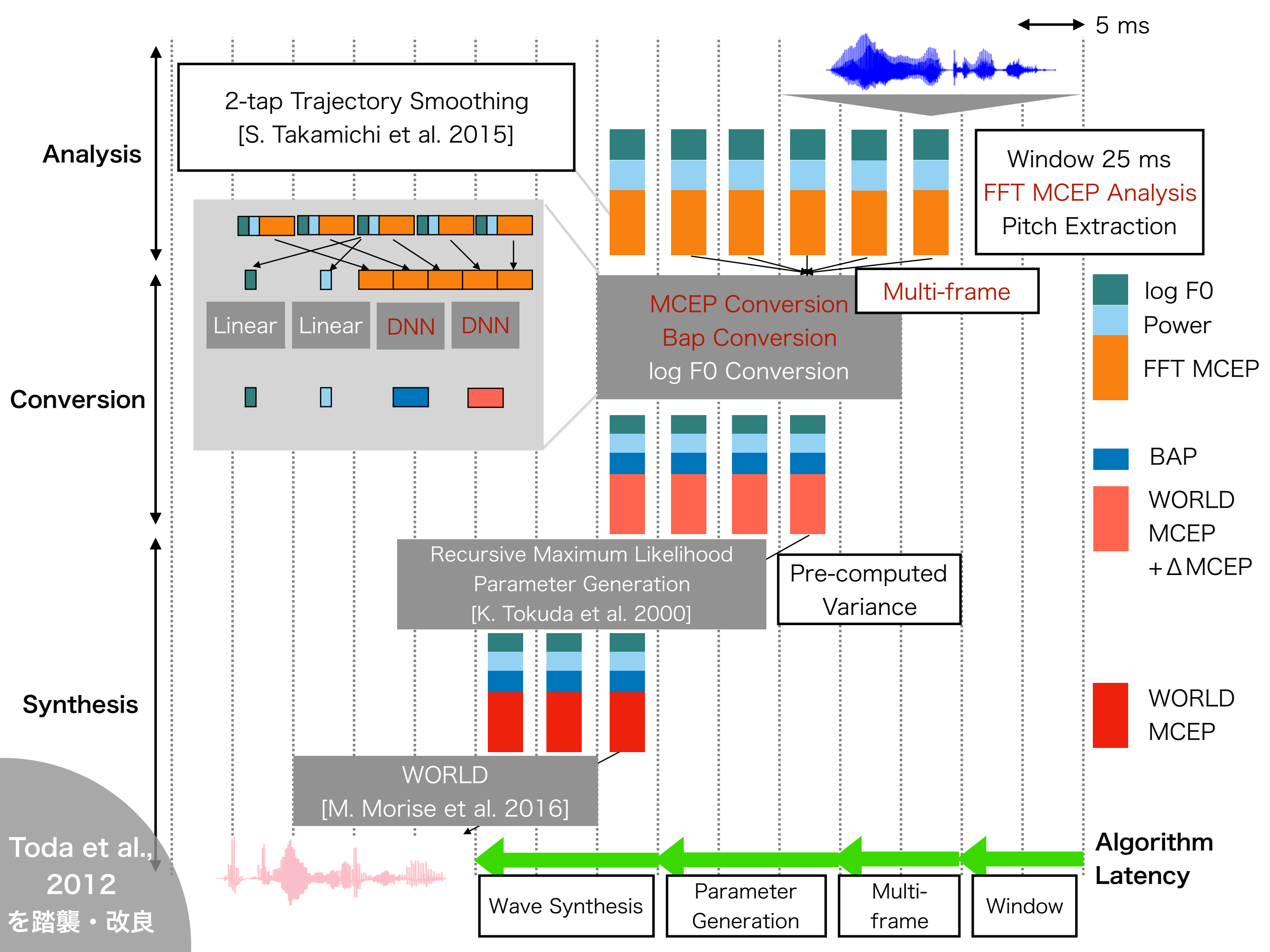


- ある話者の発話を別の話者が話したかのように変換.
- 言語情報を保持, パラ言語・非言語情報を変換.
- 発声器官の物理的制約を超えて, 多様な音声コミュニケーションを可能にするという点で期待.

DNN音声変換



- メルケプストラム係数(MCEP), 基本周波数(F0), 帯域平均化された非周期性指標(BAP)などの音声特徴量の抽出.
- 従来のDNN音声変換は, 分析・変換・合成にオフラインの処理を想定. (ex. Bidirectional Long Short-Term Memory [L. Sun 2015])
- リアルタイムDNN音声変換は未確立.
 - • • 遅延聴覚フィードバック~200 ms [Aubrey J. 1963]



遅延計測のための実験条件

データセット	声優統計コーパス [Y_Benjo and MagnesiumRibbon, 2017] (100発話, ~12分, 男性2名)
学習/評価	90 / 10 発話
サンプリング周波数	16 kHz
音声特徴量	F0 / パワー/ MCEP(40次元) / BAP(1次元)
DNN	3層のMulti-layer Perceptron Input-Output Residual 構造 [Y. Saito 2018]
CPU性能	Intel (R) Core (TM) i7-3770K CPU @ 3.50 GHz

遅延評価

- アルゴリズム遅延 **50 ms**
 - 分析部: **10 ms** (フレーム分析長: 2フレーム分)
 - 変換部: **10 ms** (マルチフレーム: 2フレーム分)
 - 合成部: **30 ms** (MLPG: 3フレーム分, WORLDバッファ: 3フレーム分)
- CPU処理遅延
 - 合計 Real Time Factor (RTF): **0.50 (<1.0)**

Table 1 処理遅延

	解析部	変換部	合成部
処理時間 [ms]	0.85±0.092	1.12±0.51	0.57±0.10
RTF	0.17	0.22	0.11

➡ リアルタイムで動作可能.

デモ動画



動画: <https://www.youtube.com/watch?v=P9rGqoYnfCg>

(2/3) データ拡張法による 変換音声の品質向上

データ拡張法

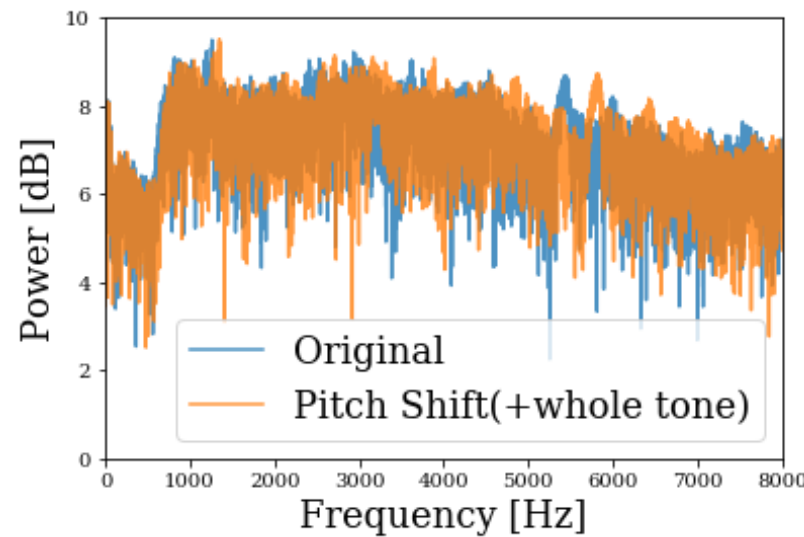
- DNNの汎化性向上のために使われる手法.
- 人工的に学習データを水増しする.



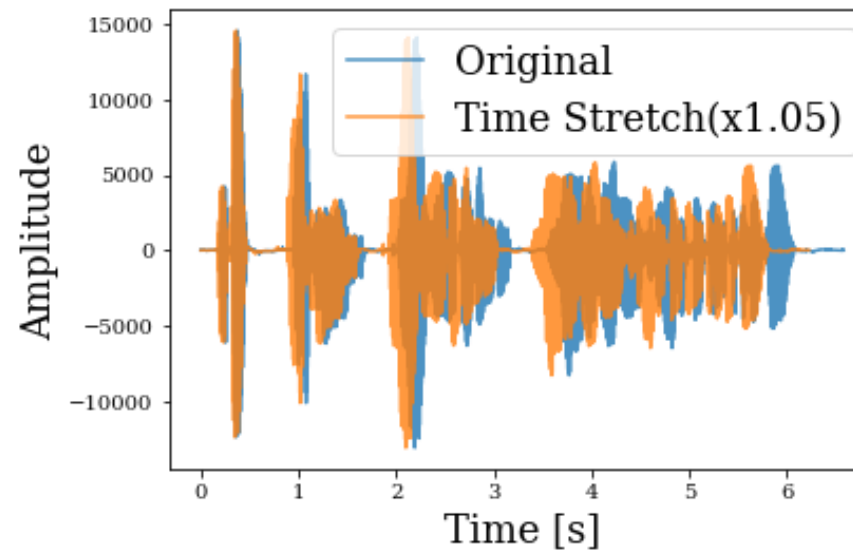
- 以下のタスクで有効性が確認されている.
 - 音声認識 [T. Ko et al. 2015]
 - 音響イベント分類 [B. Mcfee et al. 2015]
- 音声変換への適用は未調査.

提案するデータ拡張法

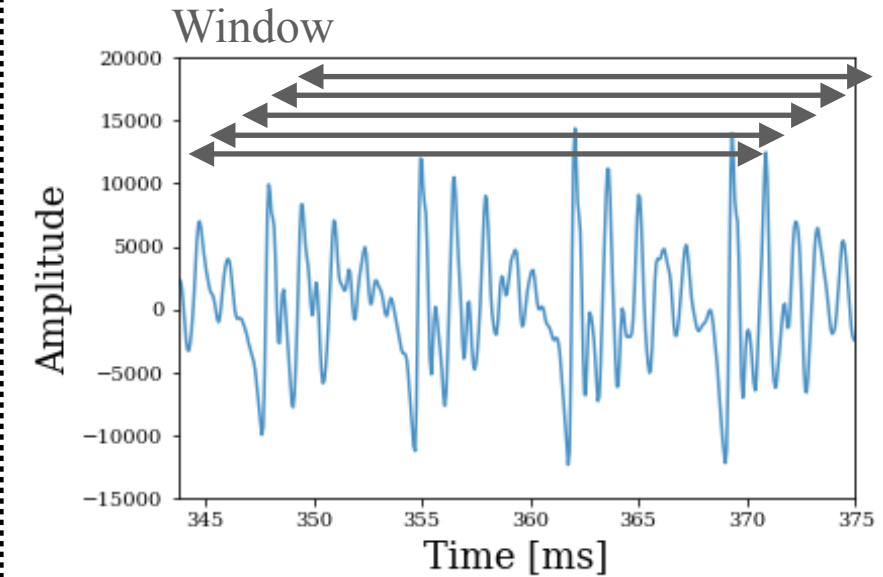
(1) Pitch Shift



(2) Time Stretch



(3) Time Shift



着想

人間は同じ文章でも全く同様の読み上げはしない。

{ピッチ / 話速}

- **Pitch Shift:** 話速を保ったままピッチを変化. [J Salmon 2017]
- **Time Stretch:** ピッチを保ったまま話速を変化. [J Salmon 2017]

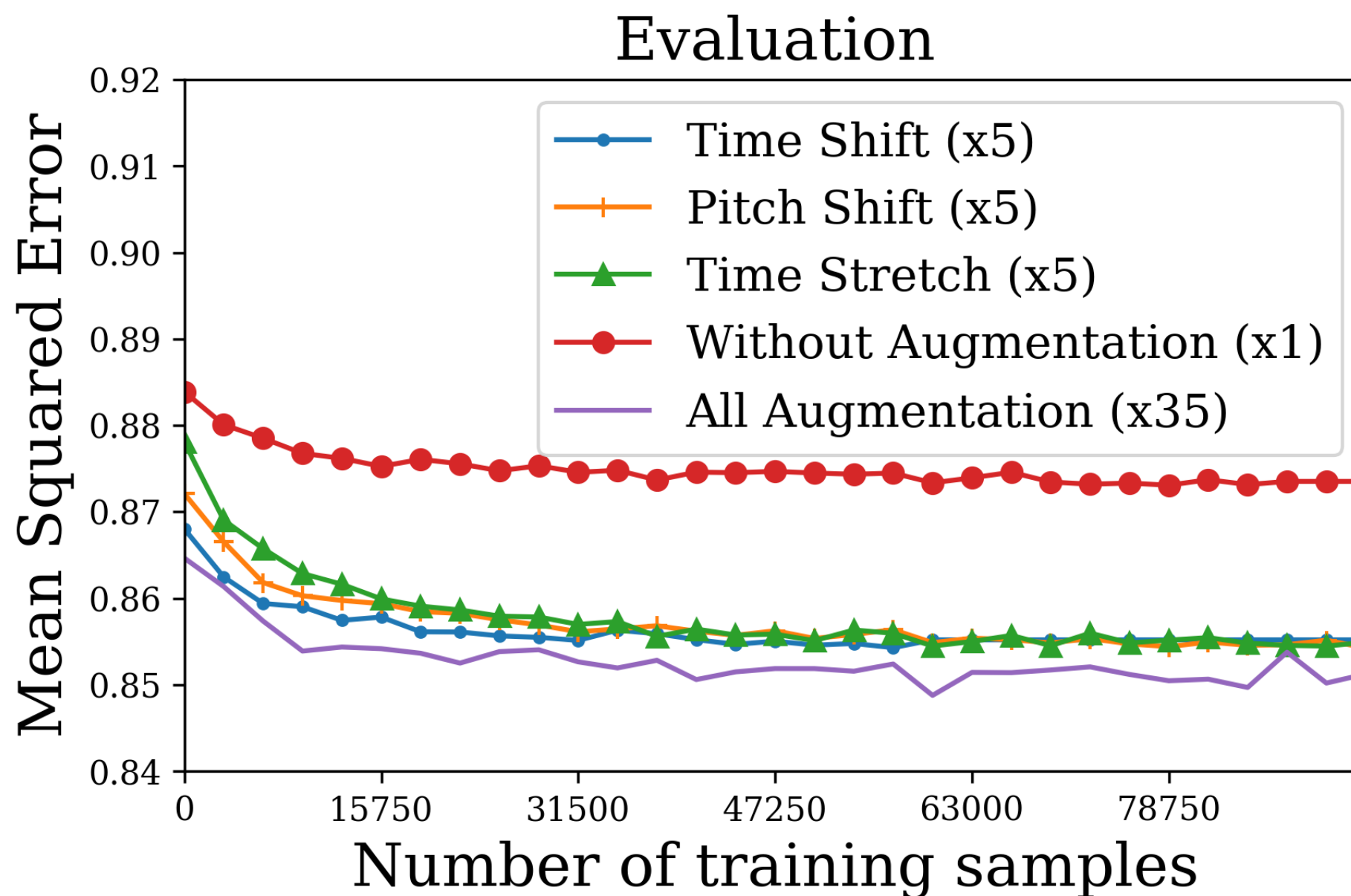
WSOLA アルゴリズム [W. Verhelst and M. Roelands, 1993]

- **Time Shift:** FFT分析の開始地点を摂動.

品質評価のための実験条件

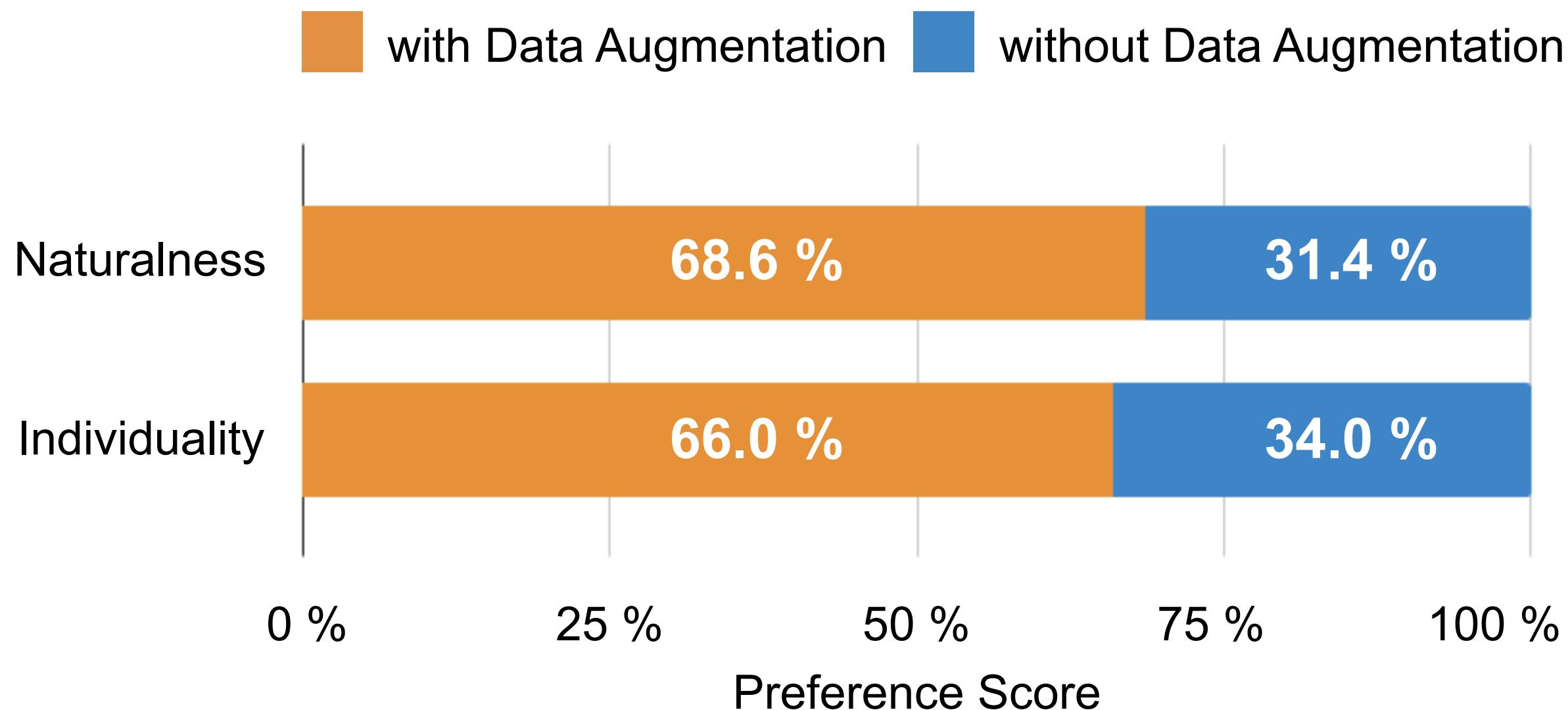
データセット	声優統計コーパス [Y_Benjo and MagnesiumRibbon, 2017] (100発話, ~12分, 男性2名)
学習/評価	90 / 10 発話
サンプリング周波数	16 kHz
音声特徴量	F0 / パワー/ MCEP(40次元) / BAP(1次元)
DNN	3層のMulti-layer Perceptron Input-Output Residual 構造 [Y. Saito 2018]
CPU性能	Intel (R) Core (TM) i7-3770K CPU @ 3.50 GHz

データ拡張法の客観評価



- 評価データの平均二乗誤差の減少を確認.
- 提案手法間の差異は確認できなかった.

データ拡張法の主観評価



- 自然性: Preference ABテスト(35人の被験者)
- 話者性: XABテスト
- **有意な音質改善効果を確認.**

サンプル音声

1. 入力話者音声
2. 目標話者音声
3. データ拡張なし (output)
4. データ拡張あり (output)

(3/3) マスク型デバイス

背景1: 雑音混入の防止



- 入力音声に雑音が混入.
 - 変換性能の低下.
 - マスク型デバイスを用いて雑音混入の防止を狙う.

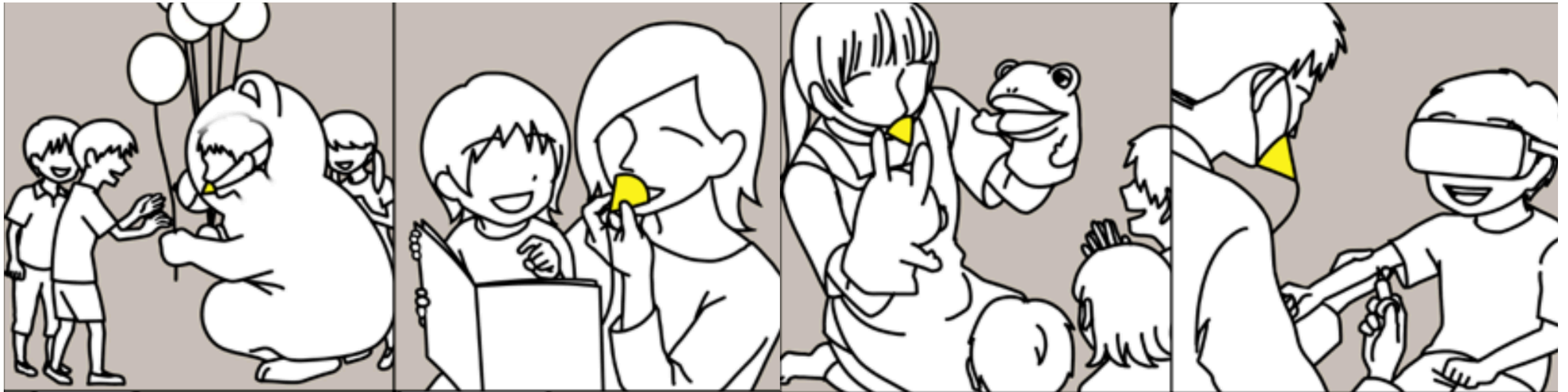
背景2: 元音声の封じ込め



画像: https://gifmagazine.net/post_images/536635

- 元話者の音声が周囲に聞こえてしまう.
 - 変換音声への没入感の低下.
 - マスク型デバイスを用いて元音声の封じ込めを狙う.

背景2: 元音声の封じ込め

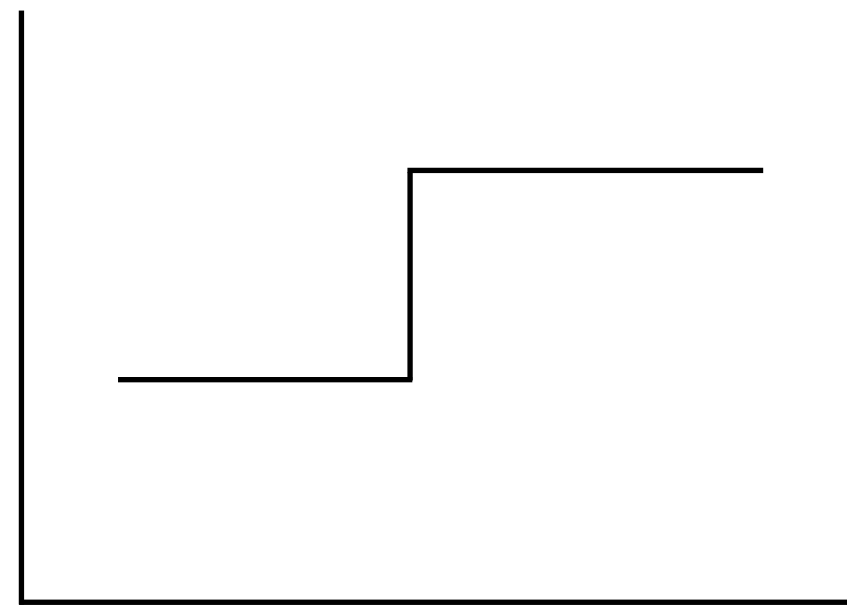


- 空間を共有する近接場コミュニケーションで重要.
→ 様々な応用事例を検討.

提案法: マスク型デバイス + フィルタ



Amplitude



Frequency [Hz]

- マスクによってこもった音声の低周波領域を弱めた音声を入力とするDNNを学習.

実験

- 音声の品質と防音効果を評価.
- 音声の品質
 - 評価データの平均二乗誤差を比較.
- 防音効果
 - 入力マイクと, 少し離れたところに置いたマイク (聞き手を想定) での音量を比較.

結果

- 音声の品質

Table 1. MSE of trained DNNs.

(a) without a mask	(b) with a mask without a filter	(c) with a mask with a filter
$0.918 \pm 6.8 \times 10^{-3}$	$0.960 \pm 4.3 \times 10^{-3}$	$0.928 \pm 3.0 \times 10^{-3}$

→ マスクによる精度悪化をフィルタで改善.

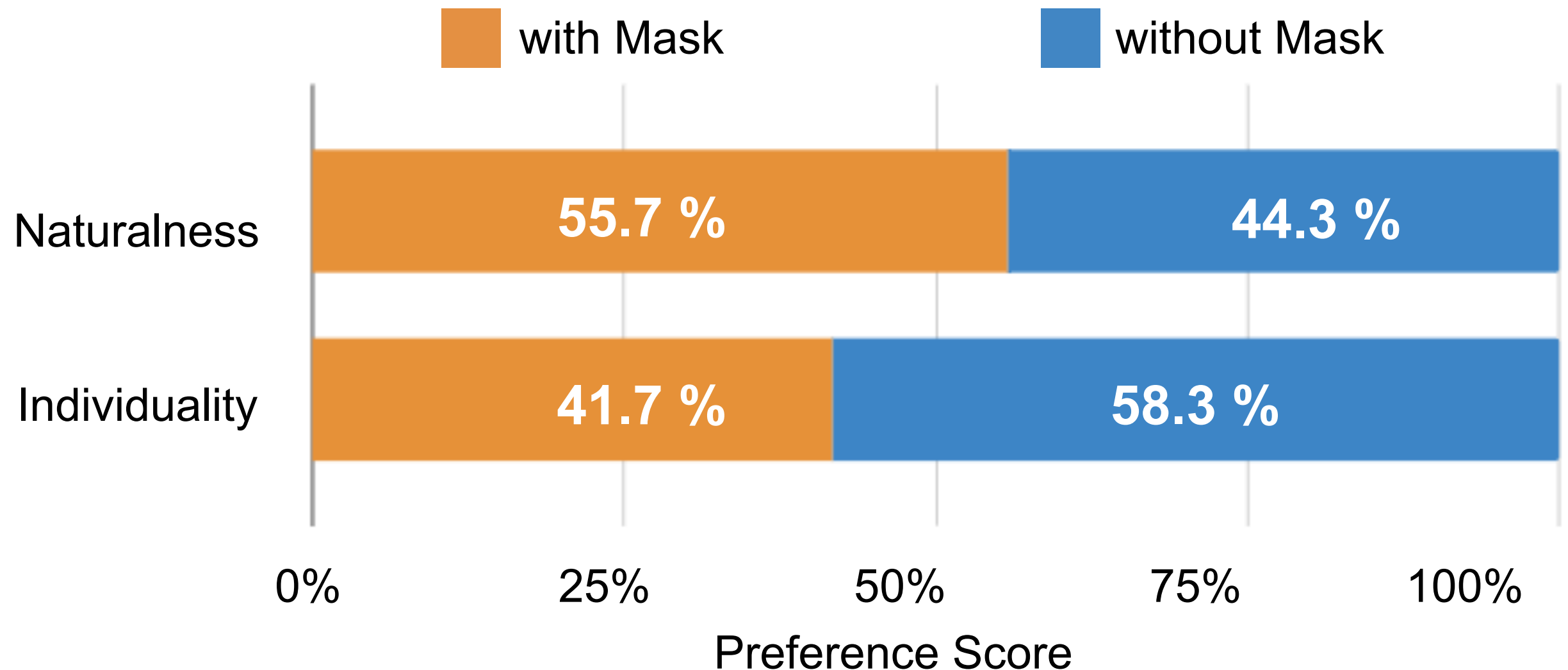
- 防音効果

Table 2. Average RMS of speaker's original speech.

	without a mask	with a mask
at a microphone	63.2	70.8
at a listener	38.3	14.6

→ マスクによる防音効果を確認.

雑音下音声変換の主観評価



- 自然性はマスクありの有効性を確認.
- 話者性ではマスクなしの方が評価が良い.
- 今後マスクの形, 素材の探求が必要.

おわりに

名探偵コナンといえば

- 蝶ネクタイ型変声器
 - (1) リアルタイムのDNN音声変換
 - (2) 変換音声の品質向上
 - (3) デバイス化



Real-Time VC Demo

今後

- 日常的にリアルタイム音声変換を使用することによる知覚体験.
 - “自分の声がタレントの声に変わって生活したら, 性格に変化は起こるのか.”
- 品質の向上.
- データ収集の簡便化.

アプリケーションアイデアがあったらぜひ議論しましょう！

まとめ

提案

- ・ リアルタイムDNN音声変換. [1, 2]
- ・ 品質改善のためのデータ拡張法. [1, 2]
- ・ 近接場コミュニケーションのためのマスク型デバイス. [3]

結果

- ・ リアルタイムでの動作を確認.
- ・ 有意な品質改善効果を確認.
- ・ 品質を保ちつつ、元音声の封じ込め効果の確認.

参照

1. 荒川 他, “リアルタイムDNN音声変換の実装とデータ拡張法による音質改善法”, 日本音響学会2019年春季研究発表会.
2. Arakawa et al., “Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device”, The 10th ISCA Speech Synthesis Workshop.
3. Arakawa et al., “TransVoice: Real-Time Voice Conversion for Augmenting Near-Field Speech Communication”, ACM UIST 2019 (to appear).