WaveNetによる言語情報を含まない感情音声合成方式の検討

松本　剣斗（岡山大学　阿部研究室）

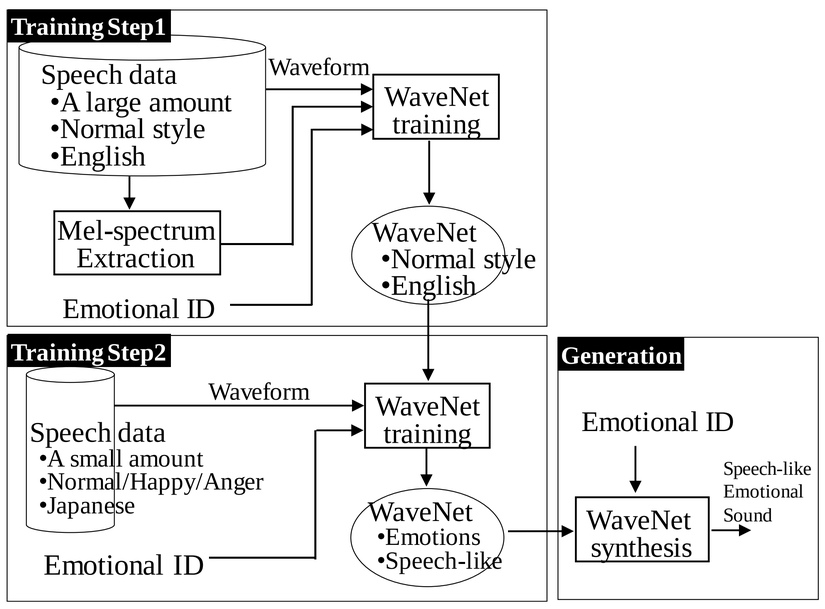
OKAYAMA UNIVERSITY

1. 研究背景・目的

- 音声は2つのチャネルにより感情情報を伝える
  - ✓ 言語チャネル
  - ✓ 非言語チャネル
- 2つのチャネルは依存or非依存
  - ✓ テキスト情報から一意に非言語情報が決まるわけではない
- 2つのチャネルを独立して扱いたい

言語情報を含まず感情情報のみを伝える音声を合成

2. 提案方式



- WaveNetを使用
  - ✓ Convolutional Neural Network
- 2つの学習ステップ
  - ✓ 必要な感情音声データの量を減らす
- Step 1: 音声の基本的な生成を学習
- Step 2: 感情表現を学習

3. 実験条件

学習データ	
学習データ	Step 1: The LJ Speech Dataset (24時間) Step 2: 声優統計コーパス (1時間)
サンプリング周波数	16 kHz
音声分析	
Window length	64 msec
Frame shift	16 msec
WaveNetの構成	
Iteration数	Step 1: 770,000 iterations Step 2: 40,000 iterations
Mini batch size	4
Residual block数	30
Dilations	$[2^0, 2^1, \ldots, 2^9]$ を3回繰り返し

4. 評価実験

- 評価データ
  - ✓ イタリア語とドイツ語の感情音声 30発話（ORIGINAL）
  - ✓ ORIGINALをWORLDで分析・合成した音声（WORLD）
  - ✓ ORIGINALをWORLDで分析し，MLSAフィルタで合成した音声（MLSA）
  - ✓ 提案方式により合成した音声　3×10 発話(WAVENET)
    - 学習データはORIGINALでない
- 実験参加者：11人の日本語母語話者
  - ✓ 音声はどこかの国の言葉と伝えた
  - ✓ イタリア語とドイツ語を聞き取れない

感情認識に関する実験

方式名	正解率		
	Normal	Angry	Happy
ORIGINAL	0.900	0.750	0.455
MLSA	0.909	0.782	0.368
WORLD	0.900	**0.827**	0.414
WAVENET	**0.927**	0.600	**0.723**

- WAVENET以外の方式
  - ✓ HappyがNormalに間違われやすい
- WAVENET
  - ✓ 他の方式と比べてHappyは良い
    - ⇒ Angryに比べて感情表現が学習された

ORIGINALの混同行列

正解感情 \ 知覚感情	Normal	Angry	Happy
Normal	**0.900**	0.073	0.027
Angry	0.232	**0.750**	0.027
Happy	0.523	0.023	**0.455**

WORLDの混同行列

正解感情 \ 知覚感情	Normal	Angry	Happy
Normal	**0.900**	0.082	0.018
Angry	0.150	**0.827**	0.023
Happy	0.564	0.023	**0.414**

MLSAの混同行列

正解感情 \ 知覚感情	Normal	Angry	Happy
Normal	**0.909**	0.082	0.009
Angry	0.195	**0.782**	0.023
Happy	0.582	0.045	**0.368**

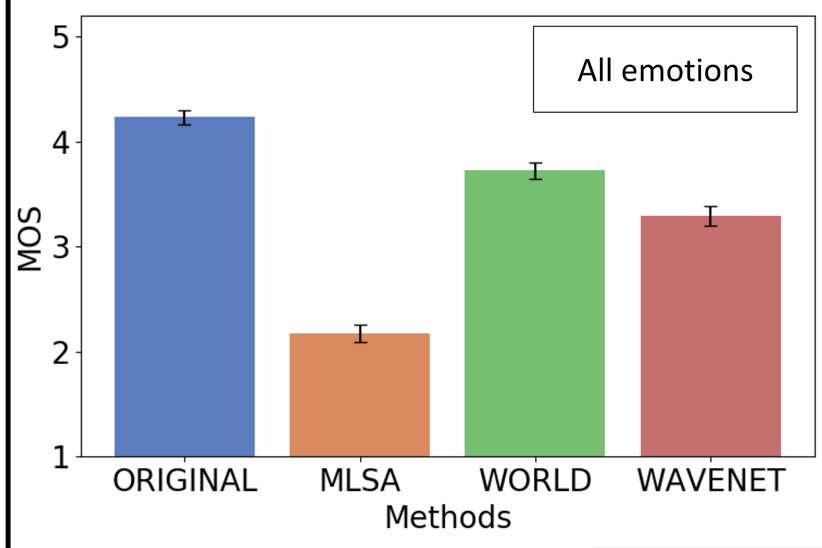
WAVENETの混同行列

正解感情 \ 知覚感情	Normal	Angry	Happy
Normal	**0.927**	0.050	0.023
Angry	0.300	**0.600**	0.100
Happy	0.200	0.077	**0.723**

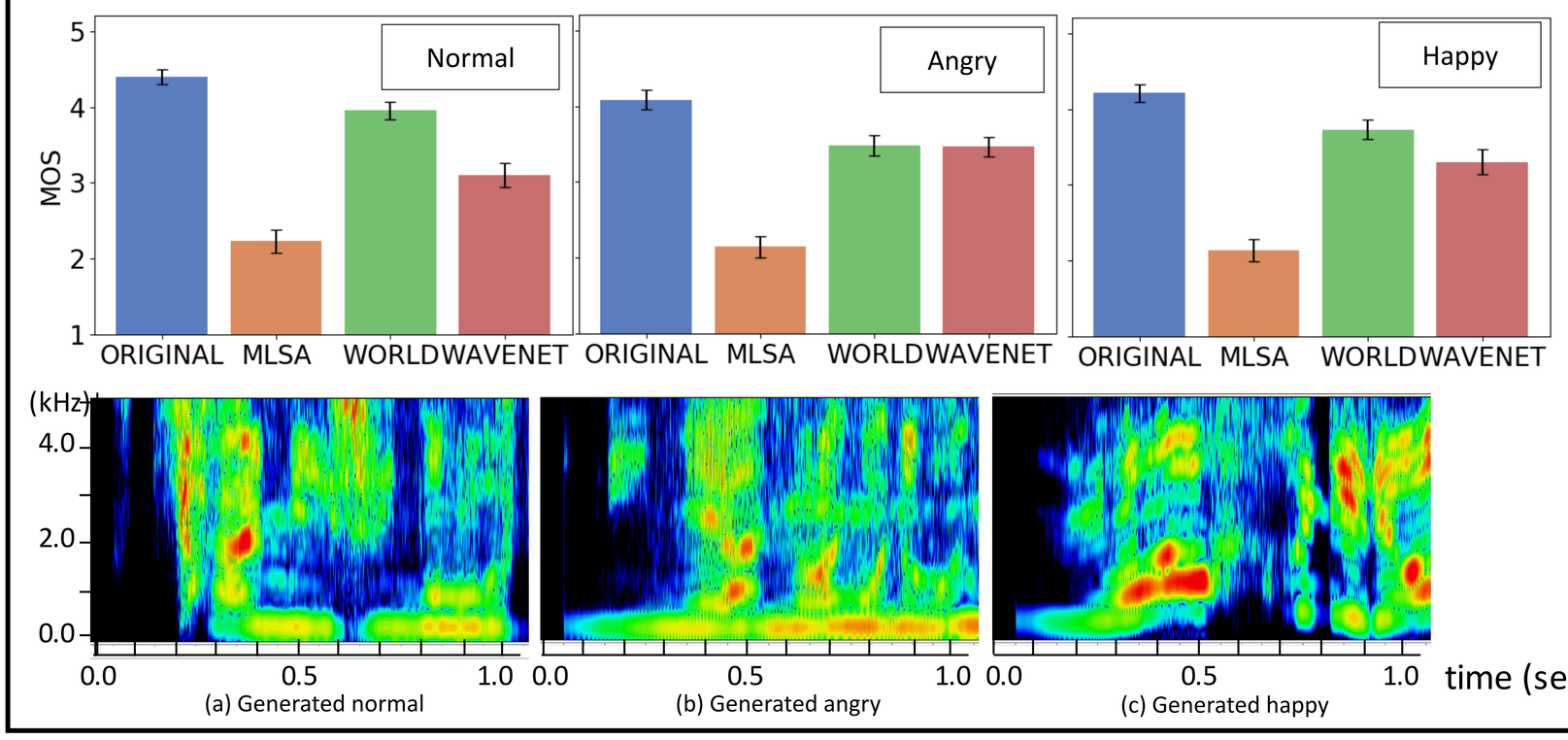
学習データの混同行列

正解感情 \ 知覚感情	Normal	Angry	Happy
Normal	**1.000**	0.000	0.000
Angry	0.000	**0.994**	0.006
Happy	0.000	0.011	**0.989**

自然性に関する実験



- WAVENET
  - ✓ MLSAよりも高く，WORLDに近い
    - ⇒ 感情ラベルのみにより合成だが，差はわずか
  - ✓ AngryとHappyではWAVENETとWORLDに大きな差はない
    - ⇒ スペクトルや基本周波数の急な変化があるためWORLDの分析合成が困難になったため



5. まとめと今後の課題

- WaveNetを用いた言語情報を含まないが感情情報は伝える音を生成する方式を検討
- 提案方式による合成音声は感情情報を含んでいることがわかった
- 今後の課題:
  - ✓ 学習データ量と合成品質との関係の調査
  - ✓ 笑い声やため息等の非言語的な音声合成の検討