

DNN型パワースペクトル復元によるボコーダ音質改善

○岡本拓磨¹, 橘健太郎¹, 戸田智基^{2,1}, 志賀芳則¹, 河井恒¹, (¹情報通信研究機構, ²名古屋大学)

1. 研究背景と目的

■ 研究背景：統計的音声合成(SPSS)の高音質化

■ SPSSの処理の流れと音質劣化要因



■ 高品質ボコーダ

■ STRAIGHT(Kawahara et al., 1999), WORLD(Morise et al., 2016)

※ 高品質な合成には多数のパラメータが必要

→SPSS：メルケプストラムのように少数のパラメータを使用→音質劣化

■ ニューラルネットワークを用いたコーパス依存型ボコーダ

■ GlottDNN(Airaksinen et al., 2016)

※ 位相回復も考慮した手法, 声門閉鎖点(GCI)の推定精度に依存

■ パワースペクトルによる合成(高木ら, 2016)

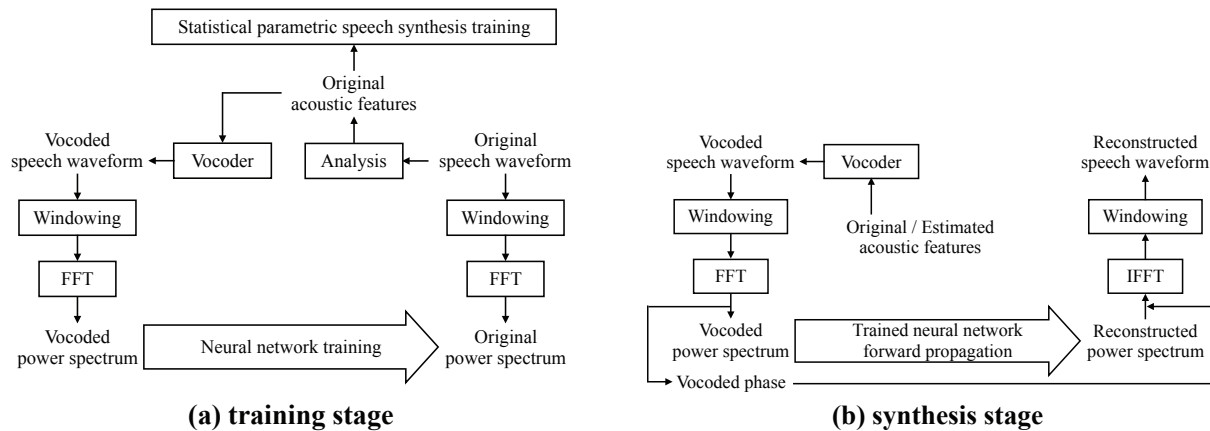
※ パワースペクトル+無矛盾反復法位相付与による波形生成, 位相復元が課題

■ 目的：任意のボコーダ音声の高品質化

■ 劣化音声をニューラルネットワークで原音声へと回復

2. 提案法

■ ニューラルネットワークを用いたボコーダ音声のパワースペクトル復元



■ スペクトル歪み(SD)を最小化するモデルパラメータの学習

$$\hat{\lambda} = \arg \min_{\lambda} \frac{1}{2} \sum_t \|P_{\text{org}} - P_v\|^2 \quad \text{SD} = 10 \log_{10} \frac{1}{FT} \sum_{f=1}^F \sum_{t=1}^T |P_{\text{org}} - P_v|$$

■ 位相：ボコーダ音声の位相をそのまま使用

3. 分析合成音を用いた実験

■ 音声コーパス：日本語女性音声, サンプル周波数16 kHz

■ 学習セット：7000 文(約 4.7 時間), テストセット：20文

■ パワースペクトル分析

■ 窓関数：Hann窓256サンプル, フレームシフト：1 ms

■ ニューラルネットワークパラメータ

■ 入力層：ボコーダ音声パワースペクトル129次元

■ 出力層：原音声パワースペクトル129次元

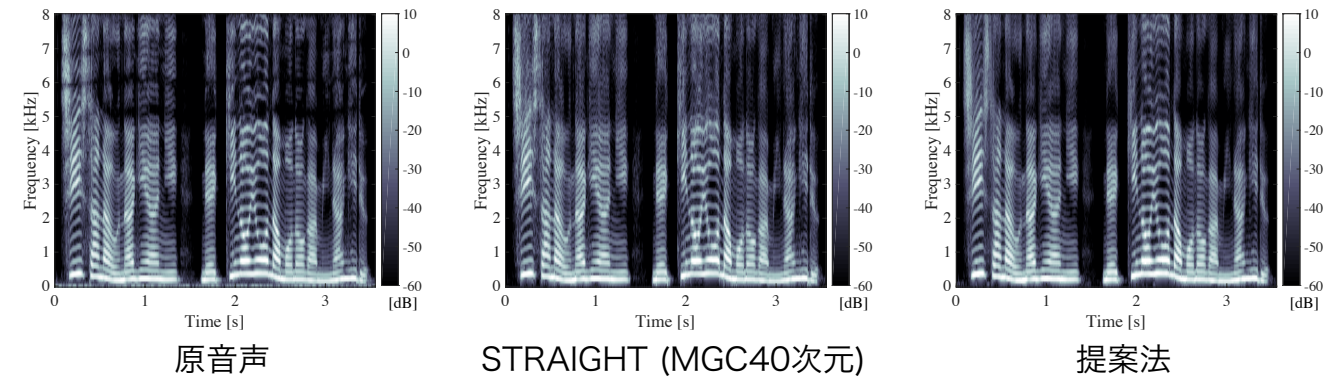
■ 隠れ層：3, 1024ノード, 活性化関数：シグモイド, 最適化手法：Adam
ミニバッチ数256, 20 epoch, 正規化：平均0, 分散1

■ ボコーダ：F0はSTRAIGHT分析による取得

■ MLSA：メルケプストラム25次(SPTK分析) + パルス励振

■ STRAIGHT：1024次元→40次元メルケプストラム(MGC) + 非周期成分5帯域

■ 客観評価



■ 主観評価

■ 対比較聴取実験：被験者7名

S	P	N	p-value	Z-score
22	99	19	< 10 ⁻¹⁰	-7.0

S : STRAIGHT, P : Proposed, N : Neutral

客観評価結果(スペクトル歪み)

	SD (dB)
MLSA	-24.5
MLSA with proposal	-27.1
STRAIGHT	-33.6
STRAIGHT with proposal	-32.6

4. DNN音声合成音を用いた実験

■ DNN音声合成コーパス：学習セット6000文

■ SPSSニューラルネットワークパラメータ

■ 入力層：コンテキストラベル483次元

■ 出力層：音響特徴量

※ MLSA：79次元, STRAIGHT：139次元

主観評価結果(スペクトル歪み)

	SD (dB)
MLSA	-21.8
MLSA with proposal	-22.4
STRAIGHT	-23.4
STRAIGHT with proposal	-23.2