Anti-spoofing に敵対するDNN音声変換の評価

齋藤 佑樹, 高道 慎之介, 猿渡 洋 (東京大学大学院 情報理工学系研究科)

1. 本発表の概要

問題点: 統計的パラメトリック音声合成の音質劣化
生成される音声特徴量系列の過剰な平滑化が一因

テキスト音声合成における改善策: Anti-spoofing に敵対する音響モデル学習 (敵対的DNN音声合成)
声のなりすましを防ぐ anti-spoofing を詐称するような音声を生成 [Saito et al., 2017.]

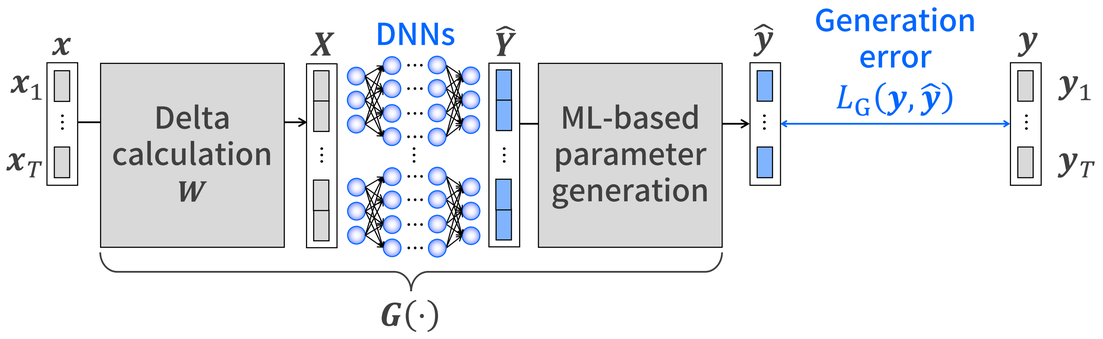
本発表:
(1) DNN音声変換のための anti-spoofing に敵対する音響モデル学習
(2) Highway network を用いた差分スペクトル推定

結果: 提案手法による音質改善効果を確認

音声サンプルはこちらです.
Speech samples are available.
http://sython.org/demo/sp201701advvc/demo.html

2. 従来手法

2.1 Minimum Generation Error (MGE) 学習 [Wu et al., 2016.]



$x, y, \hat{y}$: { input, output, converted } speech features
$X, \hat{Y}$: { input, converted } static-dynamic speech features

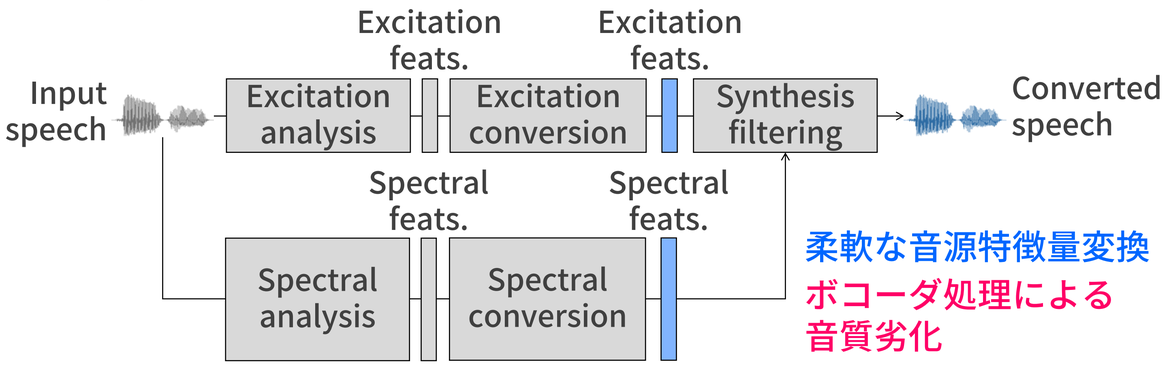
MGE学習の損失関数 (特徴量の生成誤差)

$$L_G(y, \hat{y}) = \frac{1}{T}(\hat{y} - y)^\top (\hat{y} - y) \rightarrow \text{Minimize}$$

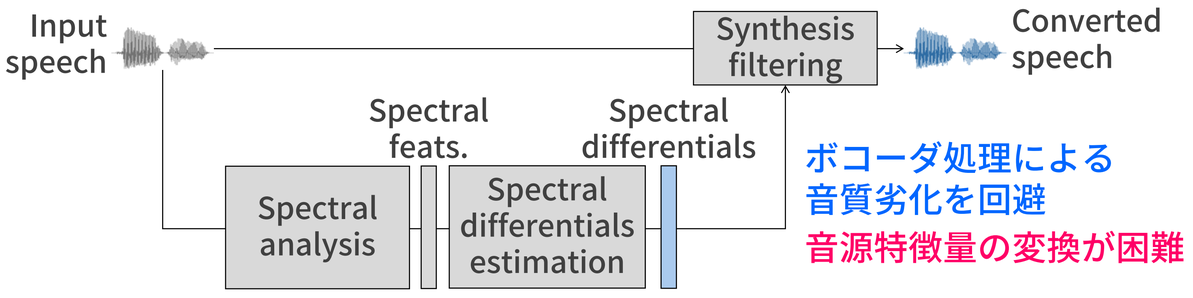
問題点: 自然音声と異なる特徴量分布
自然音声と比較して分布が縮小

2.2 音声変換方式

(1) 特徴量変換 [Toda et al., 2007.]

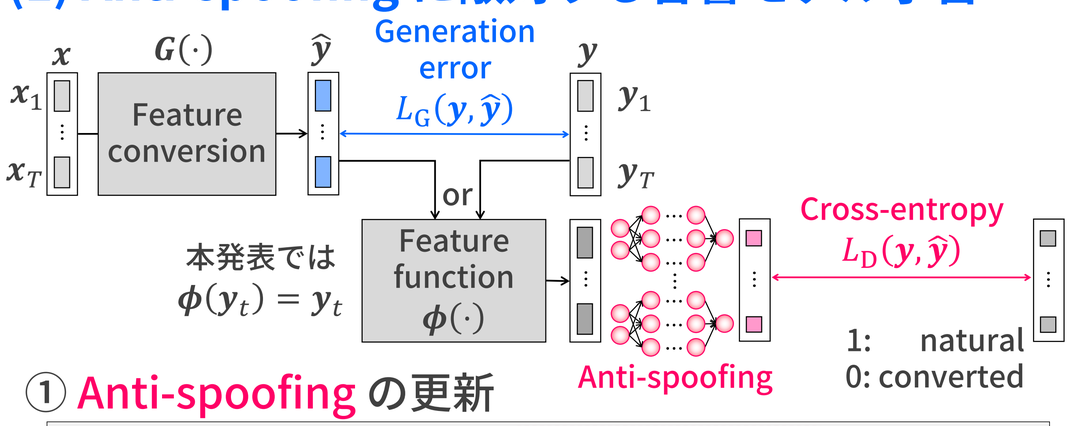


(2) 差分スペクトル推定 [Kobayashi et al., 2014.]



3. 提案手法

(1) Anti-spoofing に敵対する音響モデル学習



① Anti-spoofing の更新

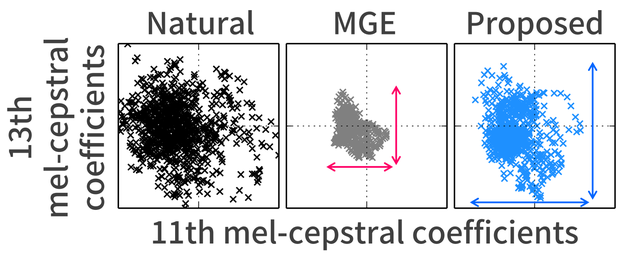
$$L_D(y, \hat{y}) = \underbrace{-\frac{1}{T}\sum_{t=1}^{T} \log D(y_t)}_{L_{D,1}(y)} \underbrace{-\frac{1}{T}\sum_{t=1}^{T} \log\left(1 - D(\hat{y}_t)\right)}_{L_{D,0}(\hat{y})} \rightarrow \text{Minimize}$$

② 音響モデルの更新

Anti-spoofing を詐称するための損失

$$L(y, \hat{y}) = L_G(y, \hat{y}) + \omega_D \frac{E_{L_G}}{E_{L_D}} L_{D,1}(\hat{y}) \rightarrow \text{Minimize}$$

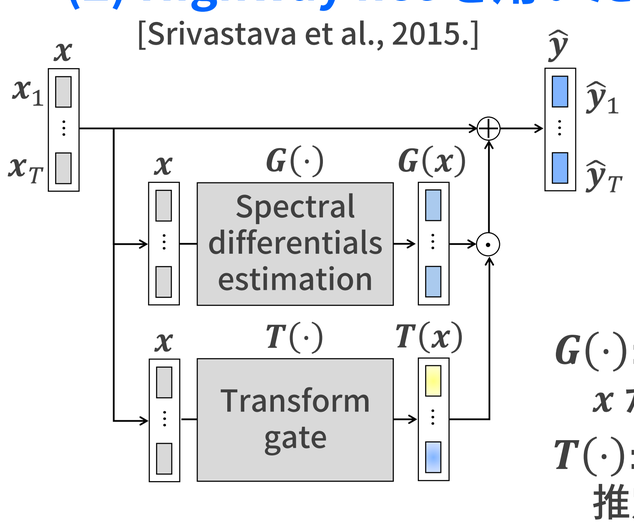
ω_D: 重み, E_{L_G}, E_{L_D}: $L_G(\mathbf{y}, \hat{y}), L_{D,1}(\hat{y})$ の期待値 (損失のスケールを調整)



提案手法により
分布の違いを補償！

(2) Highway net を用いた差分スペクトル推定

[Srivastava et al., 2015.]



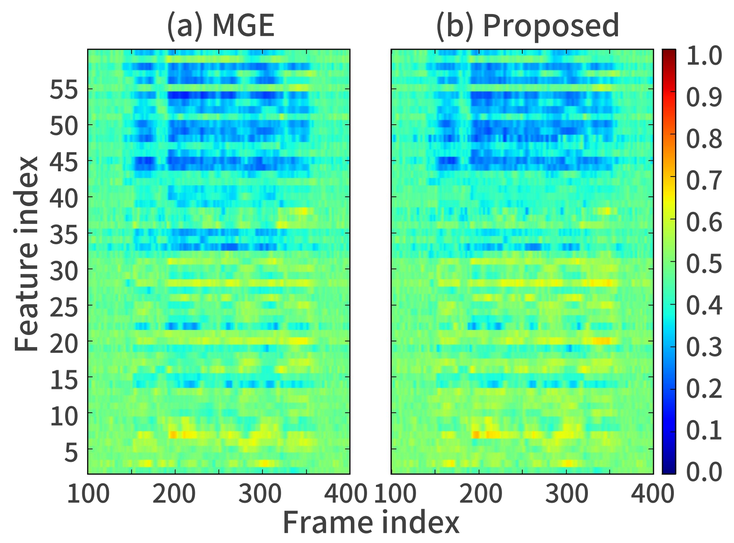
入力音声特徴量を直接的に利用

$$\hat{y} = x + G(x) \odot T(x)$$

入力音声特徴量に応じて推定結果を重み付け

$G(\cdot)$: 差分スペクトル推定器
x から $y - x$ を推定

$T(\cdot)$: Transform gate
推定結果に対する重み (0 ～ 1)

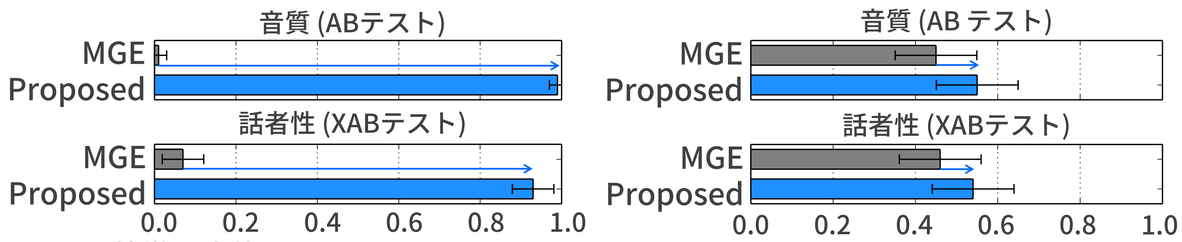


高次メルケプストラム
小さい重み
→ 入力音声特徴量を直接的に使用

低次メルケプストラム
大きい重み
→ 入力音声特徴量を重点的に変換

4. 実験的評価

データセット	ATR 音素バランス503文 (男性話者 2名)
学習 / 評価データ	A−I セット 450文 / Jセット 53文
サンプリング周波数	16 kHz
音声パラメータ	60次元のメルケプストラム, F_0, 5帯域の非周期性指標
提案法の適用 / 重み ω_D	メルケプストラム / 1.0
音響モデル & anti-spoofing	Feed-Forward



(a) 特徴量変換 w/ Feed-Forward (b) 差分スペクトル推定 w/ highway net
プリファレンススコア (エラーバーは95%信頼区間)