DNN 音声合成のための anti-spoofing を考慮した学習アルゴリズム *

☆齋藤 佑樹, 高道 慎之介, 猿渡 洋 (東大院・情報理工)

1 はじめに

自然音声と合成音声を識別できる特徴量は, 高音質音声合成のための音響モデル構築に有効である. これまでに, 変調スペクトル [1] などの解析的特徴量を補償する学習アルゴリズムが提案され, その音質改善効果が確認されている. この枠組みを一般化することにより, 合成音声の更なる音質改善を期待できる.

本稿では, 合成音声による声のなりすましを検出する識別器である anti-spoofing を考慮した学習アルゴリズムを提案する. 音声合成の音響モデルは anti-spoofing に敵対するように学習されるため, 識別可能な特徴量を補償した高音質音声合成が可能となる. また, 音声合成の音響モデルと anti-spoofing の両者を Deep Neural Network (DNN) として記述することにより, DNN の恩恵を受けた学習・補償が可能となる. 実験的評価では, 客観・主観評価により提案アルゴリズムの有効性を示す.

2 従来の学習アルゴリズム

DNN 音声合成の学習部では, 自然音声パラメータ $\boldsymbol{c}$ と合成音声パラメータ $\hat{\boldsymbol{c}}$ から計算される損失関数 $L_{\mathrm{G}}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ を最小化する. 本稿で採用する Minimum Generation Error (MGE) 学習 [2] の損失関数は, 自然音声のパラメータ系列 $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ と最尤パラメータ生成 [3] 後の合成音声のパラメータ系列 $\hat{\boldsymbol{c}} = \left[\hat{\boldsymbol{c}}_1^\top, \cdots, \hat{\boldsymbol{c}}_t^\top, \cdots, \hat{\boldsymbol{c}}_T^\top\right]^\top$ の二乗誤差として次式で与えられる.

$$\begin{aligned} L_{\mathrm{G}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) &= \frac{1}{T}(\hat{\boldsymbol{c}} - \boldsymbol{c})^\top (\hat{\boldsymbol{c}} - \boldsymbol{c}) \\ &= \frac{1}{T}(\boldsymbol{R}\hat{\boldsymbol{o}} - \boldsymbol{c})^\top (\boldsymbol{R}\hat{\boldsymbol{o}} - \boldsymbol{c}) \end{aligned} \tag{1}$$

ここで, t はフレームインデックス, T はフレーム数, $\boldsymbol{c}_t$ はフレーム t における音声パラメータ, $\hat{\boldsymbol{o}}$ は DNN から出力される音声パラメータの静的・動的特徴量系列である. 行列 $\boldsymbol{R}$ は, 静的・動的特徴量の対応関係を示す行列 $\boldsymbol{W}$ [3] および学習データを用いて別途推定される共分散行列 $\boldsymbol{\Sigma}$ を用いて次式で計算される.

$$\boldsymbol{R} = \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \tag{2}$$

本稿では, 静的・動的特徴量の二乗誤差を最小化する学習 [4] を施した DNN を初期値として, MGE 学習を実施する.

3 提案する学習アルゴリズム

3.1 Anti-spoofing

Anti-spoofing [5] では, 合成音声による声のなりすましを防ぐために, 当該音声特徴量 (もしくは音声波形) を用いて自然音声と合成音声を識別する. DNN に基づく anti-spoofing (例えば [6]) では, 音声特徴量に対して素性関数 $\boldsymbol{\phi}(\cdot)$ を適用した後に, 当該音声特徴量が自然音声である事後確率 $D(\boldsymbol{\phi}(\cdot))$ を出力する. 本稿では, 素性関数を $\boldsymbol{\phi}(\boldsymbol{c}_t) = \boldsymbol{c}_t$ と定義し, 各フレームにおける音声パラメータを識別に用いる. 学習時に最小化される損失関数 $L_{\mathrm{D}}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ は, 次式で与えられる.

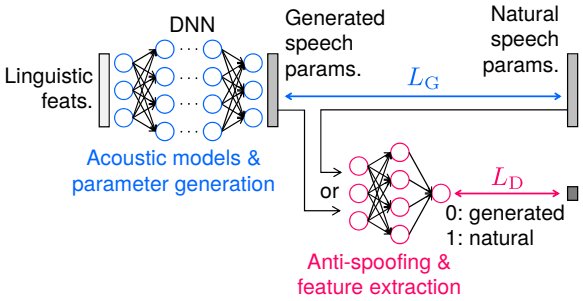


Fig. 1 Calculation of the loss function in the proposed algorithm.

$$L_{\mathrm{D}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = L_{\mathrm{D},1}(\boldsymbol{c}) + L_{\mathrm{D},0}(\hat{\boldsymbol{c}}) \tag{3}$$

ここで, $L_{\mathrm{D},1}(\boldsymbol{c})$ 及び $L_{\mathrm{D},0}(\hat{\boldsymbol{c}})$ は自然音声, 合成音声に対する損失であり, それぞれ次式で計算される.

$$L_{\mathrm{D},1}(\boldsymbol{c}) = -\frac{1}{T}\sum_{t=1}^{T} \log D(\boldsymbol{c}_t) \tag{4}$$

$$L_{\mathrm{D},0}(\hat{\boldsymbol{c}}) = -\frac{1}{T}\sum_{t=1}^{T} \log\left(1 - D(\hat{\boldsymbol{c}}_t)\right) \tag{5}$$

3.2 Anti-spoofing を考慮した音響モデル学習

ここでは, anti-spoofing に敵対するように音響モデルを学習する手法を提案する. 提案アルゴリズムにおける損失関数の計算手順を Fig. 1 に示す.

提案アルゴリズムでは, 次式の損失関数 $L(\boldsymbol{c}, \hat{\boldsymbol{c}})$ を最小化するように音響モデルを更新する.

$$L(\boldsymbol{c}, \hat{\boldsymbol{c}}) = L_{\mathrm{G}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) + \omega_{\mathrm{D}} \frac{E_{L_{\mathrm{G}}}}{E_{L_{\mathrm{D}}}} L_{\mathrm{D},1}(\hat{\boldsymbol{c}}) \tag{6}$$

ここで, $E_{L_{\mathrm{G}}}$ と $E_{L_{\mathrm{D}}}$ はそれぞれ $L_{\mathrm{G}}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{c}})$ の期待値を表す. 式 (6) の第 2 項にこれらの比の値をかけることで, $L_{\mathrm{G}}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{c}})$ のスケールを調整する. また, ω_{D} は anti-spoofing の損失に対する重みを表す. $\omega_{\mathrm{D}} = 0$ のときに, この損失関数は従来の MGE 学習と等価になり, $\omega_{\mathrm{D}} = 1$ のときに, $L_{\mathrm{G}}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{c}})$ は等重みをもつ. $L_{\mathrm{D},1}(\hat{\boldsymbol{c}})$ は合成音声パラメータを自然音声と識別させるための損失であり, 自然音声パラメータと合成音声パラメータの従う確率分布間の距離 (厳密には Jensen-Shannon divergence) を最小化させる [7]. 故に, 提案アルゴリズムにおける損失関数は, 生成誤差を最小化させ, かつ, 合成音声パラメータの従う確率分布を自然音声パラメータの従う確率分布と等しくさせる効果を持つ.

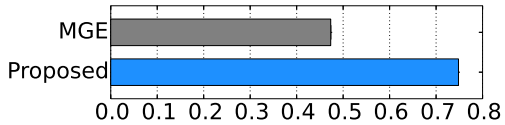
音響モデルの学習後には anti-spoofing を再学習し, 以降, これらの処理を繰り返して最終的な音響モデルを構築する.

3.3 提案アルゴリズムの特徴

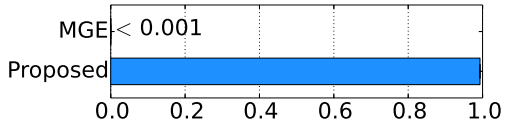
提案アルゴリズムでは, 素性関数 $\boldsymbol{\phi}(\cdot)$ として変調スペクトルなどの, 音声合成において既知の解析的特徴量のみならず, anti-spoofing において有効な解析的特徴量も, 直接的に利用可能である. また, 素性関数

---

*Training Algorithm Considering Anti-spoofing Verification for DNN-Based Speech Synthesis, by SAITO, Yuki, TAKAMICHI, Shinnosuke, and SARUWATARI, Hiroshi (The University of Tokyo).



(a) Generation loss



(b) Spoofing rates via synthetic speech

Fig. 2　Results of objective evaluations.

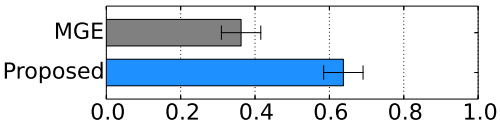


Fig. 3　Results of subjective evaluations (error bar denotes 95% confidence interval).

の設計そのものを DNN に組み込むことにより, 自動設計された特徴量も利用可能である.

提案アルゴリズムにおける学習手順は, 敵対的学習 [7] と, 識別器を含むマルチタスク学習 [8] の組合せとみなすことができる. DNN を用いた敵対的学習により, 自然音声パラメータの従う確率分布として, 従来の正規分布 [1] などよりも複雑な分布を利用できる.

音響モデルと anti-spoofing の学習は全て backpropagation の枠組みで完結するため, 任意の DNN アーキテクチャを利用できる. ただし, anti-spoofing において解析的特徴量を利用する場合, 素性関数 $\phi(\cdot)$ は微分可能である必要がある.

4　実験的評価

4.1　実験条件

実験的評価に用いるデータとして, ATR 音素バランス 503 文 [9] を利用し, A–I セット 450 文を学習に, J セット 53 文を評価に用いる. 学習データのサンプリング周波数は 16 kHz, フレームシフトは 5 ms とする. スペクトル特徴量として STRAIGHT 分析 [10] による 0 次から 24 次のメルケプストラム係数, 音源特徴量として F_0, 5 周波数帯域における平均非周期成分 [11] を用いる. スペクトル特徴量に対する前処理として, 50 Hz のカットオフ変調周波数による trajectory smoothing [12] を利用する. コンテキストラベルは, 音素, モーラ位置, アクセント型, 音素内フレーム位置などから成る 274 次元ベクトルである. DNN 学習時には, スペクトル特徴量を, 平均 0, 分散 1 に正規化する. 音声合成の音響モデルと anti-spoofing のための DNN は, Feed-Forward 型とする. 音声合成の音響モデルの隠れ層数は 3, 隠れ層の素子数は 400, 隠れ層及び出力層の活性化関数は, それぞれ Rectified Linear Unit (ReLU) 及び線形関数である. Anti-spoofing の隠れ層数は 2, 隠れ層の素子数は 200, 隠れ層及び出力層の活性化関数は, それぞれ ReLU 及び sigmoid 関数である. Anti-spoofing はスペクトル特徴量 (25 次元), 音声合成の音響モデルはその静的・動的特徴量 (75 次元) のみをそれぞれ DNN の入力・出力特徴量として扱う. F_0, 非周期成分, 継続長については, 自然音声の特徴量を使用する. 最適化アルゴリズムとして, 学習率 0.01 の AdaGrad [13] を用いる.

まず, $\omega_{\mathrm{D}} = 0$ として, 反復回数 25 回の MGE 学習により音響モデルを初期化する. 次に, $\omega_{\mathrm{D}} = 0.5$ として, anti-spoofing 学習及び, 提案アルゴリズムによる音響モデル学習を交互に実施する. この際の反復回数は 25 回とする. ただし, anti-spoofing は, 自然音声パラメータと MGE 学習後の合成音声パラメータをある程度識別するように初期化する. この初期化の反復回数は 5 回とする. 提案アルゴリズムにおける期待値 $E_{L_{\mathrm{G}}}$ と $E_{L_{\mathrm{D}}}$ は, 反復毎に, その時点における anti-spoofing と音響モデルを用いて計算する.

従来の学習アルゴリズムと提案アルゴリズムを比較するため, まず, 合成音声パラメータの生成誤差 (式 (1)) と anti-spoofing における詐称率を計算する. 詐称率は合成音声パラメータを自然音声と誤識別した割合を表す. ただし, 詐称率を評価する anti-spoofing は, MGE 学習後の合成音声パラメータを用いて別途構築する. 次に, 提案アルゴリズムによる音質改善効果を確認するため, 8 名によるプリファレンス AB テストを実施する.

4.2　評価結果

客観評価結果として, 合成音声パラメータの生成誤差を Fig. 2(a) に, anti-spoofing における詐称率を Fig. 2(b) に示す. Figure 2(a) より, 提案アルゴリズムにより生成誤差は悪化することが確認できる. 一方で, Fig. 2(b) より, 詐称率は大幅に改善することが確認できる. Figure 2(b) の結果から, 従来の MGE 学習と比較して, 提案アルゴリズムにより anti-spoofing に敵対した学習及び特徴量補償が可能になることが明らかになった.

主観評価結果を Fig. 3 に示す. 提案アルゴリズムは従来の MGE 学習よりも高いスコアを獲得しているため, 提案アルゴリズムによる音質改善効果が確認された.

5　おわりに

本稿では, DNN 音声合成のための anti-spoofing を考慮した学習アルゴリズムを提案し, 実験的評価によりその有効性を示した. 今後は, 重み ω_{D} の自動調整, 及び素性関数の自動設計について検討する.

謝辞　本研究は, 総合科学技術・イノベーション会議による革新的研究開発推進プログラム（ImPACT）及びセコム科学技術振興財団の支援を受けた.

参考文献

[1] Takamichi et al., *Proc. INTERSPEECH*, pp. 1206–1210, 2015.
[2] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 7, pp. 1255–1265, 2016.
[3] Tokuda et al., *Proc. ICASSP*, pp. 1315–1318, 2000.
[4] Zen et al.，*Proc. ICASSP*, pp. 7962–7966, 2013.
[5] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 4, pp. 768–783, 2016.
[6] Chen et al., *Proc. INTERSPEECH*, pp. 2097–2101, 2015.
[7] Goodfellow et al.，*Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.
[8] Huang et al., *Proc. INTERSPEECH*, pp. 2464–2468, 2015.
[9] 阿部 他, ATR テクニカルレポート, TR-I-0166, 1990.
[10] Kawahara et al., *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[11] Ohtani et al., *Proc. INTERSPEECH*, pp. 2266–2269, 2006.
[12] Takamichi et al., *Blizzard Challenge Workshop*, 2015.
[13] Duchi et al., *The Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159, 2011.