敵対的 DNN 音声合成における F0・継続長の生成 *

◎齋藤 佑樹，高道 慎之介，猿渡 洋（東大院・情報理工）

1 はじめに

統計的パラメトリック音声合成により生成される合成音声の音質は，声のなりすましを防ぐ anti-spoofing に敵対することで改善される．これまでに我々は，anti-spoofing を詐称するように音響モデルを学習する Deep Neural Network（DNN）音声合成のための学習アルゴリズム（敵対的 DNN 音声合成） [1, 2] を提案し，テキスト音声合成のためのスペクトル特徴量の生成においてその音質改善効果を確認している．

本稿では，敵対的 DNN 音声合成の枠組みを F_0 及び継続長の生成に拡張し，合成音声のさらなる音質改善を目指す．実験的評価では，提案アルゴリズムによる合成音声の音質改善効果が得られることを示す．

2 敵対的 DNN 音声合成

敵対的 DNN 音声合成 [1] では，以降に示す損失関数を最小化するように，音声合成の DNN 音響モデルと anti-spoofing の DNN 識別モデルを交互に学習する．以降では，自然音声の特徴量系列を $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$，合成音声の特徴量系列を $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ とする．ただし，t はフレームインデックス，T はフレーム数である．

2.1 Anti-spoofing 学習の損失関数

Anti-spoofing [3] では，自然音声と合成音声を識別する識別モデルを学習する．本稿では，各フレームにおける音声特徴量を識別に利用し，次式の損失関数 $L_{\mathrm{D}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を最小化するように識別モデルを学習する．

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

ここで，$L_{\mathrm{D},1}(\boldsymbol{y})$ 及び $L_{\mathrm{D},0}(\hat{\boldsymbol{y}})$ 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．

$$L_{\mathrm{D},1}(\boldsymbol{y}) = -\frac{1}{T}\sum_{t=1}^{T} \log D(\boldsymbol{y}_t) \tag{2}$$

$$L_{\mathrm{D},0}(\hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T} \log\left(1 - D(\hat{\boldsymbol{y}}_t)\right) \tag{3}$$

2.2 音響モデル学習の損失関数

次式の損失関数 $L(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を最小化するように音響モデルを更新する．

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}} \frac{E_{L_{\mathrm{G}}}}{E_{L_{\mathrm{D}}}} L_{\mathrm{D},1}(\hat{\boldsymbol{y}}) \tag{4}$$

ここで，$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ はパラメータの生成誤差であり，$\boldsymbol{y}$ と最尤パラメータ生成 [4] 後の $\hat{\boldsymbol{y}}$ の二乗誤差 [5] として与えられる．$L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ は合成音声を自然音声と識別させるための損失であり，重み ω_{D} により重み付けされる．また，$E_{L_{\mathrm{G}}}$ と $E_{L_{\mathrm{D}}}$ はそれぞれ $L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ の期待値であり，$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ のスケールを調整する役割を持つ．音響モデルは，ω_{D} を 0 として初期化される．式 (4) の第二項は，$\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ の分布の差異を最小化するための損失 [6] であり，合成音声特徴量の分布を自然音声に近づけるための正則化項とみなされる．

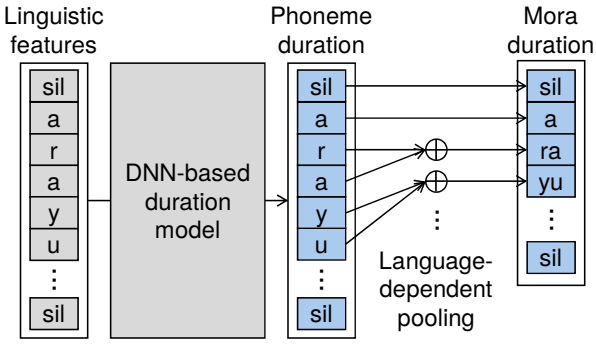

Fig. 1 等時性を考慮した継続長生成

3 敵対的 DNN 音声合成における F_0 と継続長の生成

スペクトル特徴量生成において有効である 2 節の手法 [1] を，F_0 と継続長の生成に拡張する．

3.1 F_0 の生成

連続 F_0 系列生成のための学習を，2 節と同様の枠組みで行う．ただし本稿では，スペクトルと連続 F_0 の結合ベクトルを生成・識別するモデルを構築する．異なる実装として有声フレームのみの連続 F_0 を識別に用いる手法も考えられるが，学習の簡単化のため，全フレームの連続 F_0 を用いる．

3.2 継続長の生成

音素継続長の生成に対して，2 節の枠組みを直接的に適用することが可能である（ただし，動的特徴量は考慮しない）．しかしながら，この枠組みで生成した継続長が自然な等時性を持つ保証はない．そこで本稿ではさらに，言語依存の等時性を考慮した学習法を検討する．

等時性を考慮した継続長生成の枠組みを Fig. 1 に示す．モーラ等時性を持つ日本語の音声合成の場合，DNN 継続長モデルにより生成された音素継続長からモーラ継続長を計算し，anti-spoofing の識別モデルで識別する．つまり，最終的な音素継続長は，自然な継続長分布を持つモーラ継続長から再分配されることで決定される．

本稿では，生成された音素継続長と，自然音声の音素継続長の二乗誤差を最小化する学習により，継続長モデルを初期化する．

3.3 考察

3.1 節の手法では，スペクトルと F_0 の同時分布を補償する．故に，スペクトルと F_0 の分布を独立に補償する場合と比較して，各特徴量の相関性 [7] を考慮した学習が期待される．また，スペクトルと F_0 の特徴量次元数を考慮した学習 [8] も可能である．

3.2 節の等時性を考慮した手法は，階層モデル構造 [9] を用いた生成ではなく多重解像度に基づく敵対的学習 [10] に類似し，高い時間解像度における生成誤差最小化と低い時間解像度における敵対的学習の組み合わせに相当する．また，un-pooling を用いることで音素継続長をフレームレベルに展開し，スペクトル・F_0・継続長の同時分布の補償も可能である．

---

* F0 Contour and Duration Generation for Adversarial DNN-Based Speech Synthesis, by SAITO, Yuki, TAKAMICHI, Shinnosuke, and SARUWATARI, Hiroshi (The University of Tokyo).

4　実験的評価

4.1　実験条件

実験に用いるデータとして，ATR 音素バランス 503 文 [11] 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 [12] 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 [13] を用いる．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 442 次元ベクトルである．DNN 学習時には，音声特徴量，継続長，及び実数値をとるコンテキストラベル特徴量を平均 0，分散 1 に正規化する．音声合成の音響モデル，継続長予測モデル，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 512，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．継続長モデルの隠れ層数は 3，隠れ層の素子数は 256 であり，活性化関数は音声合成の音響モデルと同じである．Anti-spoofing の隠れ層数は 3，隠れ層の素子数は 256，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．最適化手法として学習率 0.01 の AdaGrad [14] を用いる．

まず，$\omega_{\mathrm{D}} = 0.0$ として，反復回数 25 回の MGE 学習により音響モデルと継続長モデルを初期化する．本稿では以下の手法を評価する．

MGE: 従来の Minimum Generation Error 学習 [5]
ADV (sp): 敵対的 DNN 音声合成（スペクトル）[1]
ADV (sp+F0): 同上（スペクトル+F_0）
ADV (phoneme): 同上（音素継続長）
ADV (mora): 同上（モーラ継続長）

ADV（sp）と ADV（sp+F0）においては，音響モデルを用いてスペクトル，連続対数 F_0，非周期性指標，有声/無声ラベルを生成した後，スペクトル特徴量のみ，またはスペクトル特徴量と F_0 を anti-spoofing に与えて学習を行う．それ以外の音声特徴量と継続長に対する損失関数は MGE と同様である．ADV（phoneme）と ADV（mora）においては，MGE の音響モデルを用いて音声特徴量を生成する．

主観評価として，被験者数 8 名による音質に関するプリファレンス AB テストを実施する．F_0 生成における評価では，MGE，ADV（sp），及び ADV（sp+F0）を用いる．継続長生成における評価では，MGE，ADV（phoneme），及び ADV（mora）を用いる．

4.2　評価結果

主観評価結果を Fig. 2 に示す．F_0 生成における評価結果（Fig. 2（a））より，提案アルゴリズムは MGE 学習よりも高いスコアを獲得していることが確認できる．また，提案アルゴリズム同士を比較すると，ADV（sp+F0）は ADV（sp）よりも高いスコアを獲得しているため，スペクトル特徴量生成だけでなく F_0 生成にも提案アルゴリズムを適用することで，さらなる音質改善効果が得られることを確認できる．一方で，継続長生成における評価結果（Fig. 2（b））においては，有意な差は見られない．これを分析するために，MGE 学習後の音声特徴量を用いて anti-spoofing を構築し，学習データに対する accuracy を計算した（Fig. 3）．この図より，スペクトル及び F_0 と比較して，継続長を用いた場合の accuracy が低いことがわかる．故に，提案アルゴリズムによる分布補償の影響が小さく，主観的に有意な差が得られなかったと思われる．

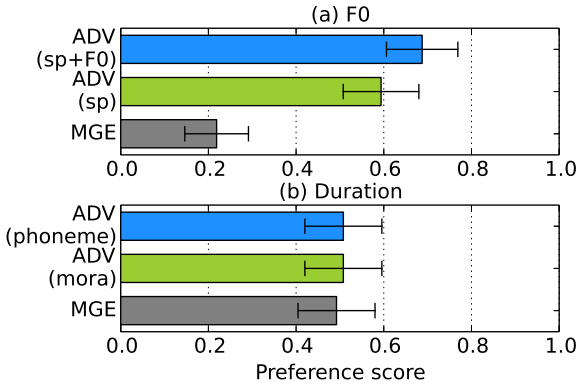

Fig. 2　主観評価結果 (エラーバーは 95%信頼区間)

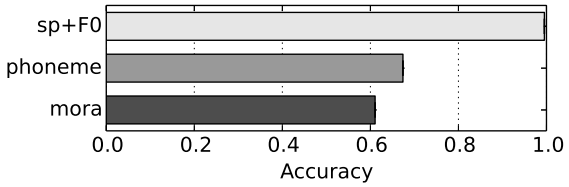

Fig. 3　Anti-spoofing の accuracy．縦軸のラベルはそれぞれ，スペクトルと F_0（sp+F0），音素継続長（phoneme），及びモーラ継続長（mora）を識別に用いたことを表す．

5　おわりに

本稿では，敵対的 DNN 音声合成の枠組みを F_0 及び継続長の生成に拡張し，F_0 生成に関して実験的評価によりその有効性を示した．今後は，他言語における提案アルゴリズムの性能について調査する．

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），セコム科学技術振興財団，及び JSPS 科研費 16H06681 の支援を受けた．

参考文献

[1] 齋藤 他, 音講論（秋）, 3-5-1, 2016.
[2] Saito et al., *Proc. ICASSP*, 2017. (to appear)
[3] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 4, pp. 768–783, 2016.
[4] Tokuda et al., *Proc. ICASSP*, pp. 1315–1318, 2000.
[5] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 7, pp. 1255–1265, 2016.
[6] Goodfellow et al., *Proc. NIPS*, pp. 2672–2680, 2014.
[7] Tanaka et al., *IEICE Trans. on Inf. and Syst.*, Vol. E97-D, No. 6, pp. 1429–1437, 2014.
[8] Kang et al., *Proc. INTERSPEECH*, pp. 1959–1963, 2014.
[9] Yin et al., *Speech Commun.*, Vol. 76, pp. 61–81, 2016.
[10] Zhang et al., *arXiv:1612.03242*, 2016.
[11] 阿部 他, ATR テクニカルレポート, TR-I-0166, 1990.
[12] Kawahara et al., *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[13] Ohtani et al., *Proc. INTERSPEECH*, pp. 2266–2269, 2006.
[14] Duchi et al., *The Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159, 2011.