雑音環境下音声を用いたDNN音声合成のための雑音生成モデルの敵対的学習*

☆宇根 昌和（徳山高専／東京大学），齋藤 佑樹，高道 慎之介，
北村 大地（東大院・情報理工），宮崎 亮一（徳山高専），猿渡 洋（東大院・情報理工）

1 はじめに

高品質な Deep Neural Network (DNN) 音声合成システム [1] の構築には，スタジオ等の理想的な環境で収録された音声データの利用が不可欠であるため，音声合成の学習に利用可能な音声データは非常に限定される．本稿では，雑音環境下音声から高品質な音声合成を構築する方法を提案する．通常，そのような音声を用いる場合，spectral subtraction (SS) 等のパラメトリック雑音抑圧を施した後に通常の音声合成の学習を行うが，その雑音抑圧による音声歪みは，音声合成の学習時に増幅されて合成音声品質を悪化させる．本稿では，敵対的学習アルゴリズム [2] により学習される雑音生成モデルを用いた，ボコーダフリー音声合成の学習法を提案する．雑音生成モデルは，観測雑音スペクトルの統計量を持つように学習され，雑音スペクトルを確率的に生成する．音響モデル（本稿では音声合成モデルと呼ぶ）は，生成雑音を加算した後のスペクトルが雑音環境下音声のスペクトルに一致するように学習される．提案法は雑音加算過程を考慮して音声合成モデルを学習するため，従来生じていた品質低下を低減できる．実験的評価により，提案法の有効性を示す．

2 従来法

SS [3] は，観測雑音のパワースペクトルの分布を期待値で近似して，雑音環境下音声のパワースペクトルから減算する手法である．ここで，観測雑音の対数振幅スペクトル系列を $\boldsymbol{y}_{\mathrm{n}} = [\boldsymbol{y}_{\mathrm{n},1}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{n},t}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{n},T_{\mathrm{n}}}^{\top}]^{\top}$，雑音環境下音声の対数振幅スペクトル系列を $\boldsymbol{y}_{\mathrm{ns}} = [\boldsymbol{y}_{\mathrm{ns},1}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{ns},t}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{ns},T}^{\top}]^{\top}$ とする．T_{n} と T はそれぞれ，観測雑音のフレーム数と雑音環境下音声のフレーム数である．$\boldsymbol{y}_{\mathrm{n},t} = [y_{\mathrm{n},t}(1), \cdots, y_{\mathrm{n},t}(f), \cdots, y_{\mathrm{n},t}(F)]^{\top}$ と $\boldsymbol{y}_{\mathrm{ns},t} = [y_{\mathrm{ns},t}(1), \cdots, y_{\mathrm{ns},t}(f), \cdots, y_{\mathrm{ns},t}(F)]^{\top}$ は，フレーム t における観測雑音及び雑音環境下音声の対数振幅スペクトルである．f は周波数ビンのインデックス，F は周波数ビン数である．ただし，$\boldsymbol{y}_{\mathrm{n}}$ は，$\boldsymbol{y}_{\mathrm{ns}}$ の非音声区間に対応する．Spectral subtraction 後の対数振幅スペクトル $\boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}$ は，次式で与えられる．

$$\exp\{y_{\mathrm{ns},t}^{(\mathrm{SS})}(f)\} = \begin{cases} \sqrt{\exp\{y_{\mathrm{ns},t}(f)\}^2 - \beta\bar{y}_{\mathrm{n},t}(f)} \\ \quad \text{if} \quad \exp\{y_{\mathrm{ns},t}(f)\}^2 > \beta\bar{y}_{\mathrm{n},t}(f) \\ 0 \quad \text{otherwise} \end{cases} \tag{1}$$

$$\bar{y}_{\mathrm{n},t}(f) = \frac{1}{T_{\mathrm{n}}} \sum_{t=1}^{T_{\mathrm{n}}} \exp\{y_{\mathrm{n},t}(f)\}^2 \tag{2}$$

ただし，β は減算係数であり，観測信号から観測雑音をどの程度減算するかを決めるパラメータである．
入力コンテキストから音声の対数振幅スペクトルを予測する音声合成モデルを $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ とする．ここで，入力コンテキスト系列を $\boldsymbol{x} = [\boldsymbol{x}_1^{\top}, \cdots, \boldsymbol{x}_t^{\top}, \cdots, \boldsymbol{x}_T^{\top}]^{\top}$ とする．$\boldsymbol{G}_{\mathrm{s}}(\cdot)$ のモデルパラメータは，生成される対数振幅スペクトル $\hat{\boldsymbol{y}}_{\mathrm{s}} = \boldsymbol{G}_{\mathrm{s}}(\boldsymbol{x})$ と $\boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}$ の平均二乗誤差（MSE: Mean Squaed Error）を最小化するように学習される．その損失関数は，次式で示される．

$$L_{\mathrm{MSE}}\left(\hat{\boldsymbol{y}}_{\mathrm{s}}, \boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}\right) = \frac{1}{T}\left(\hat{\boldsymbol{y}}_{\mathrm{s}} - \boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}\right)^{\top}\left(\hat{\boldsymbol{y}}_{\mathrm{s}} - \boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}\right) \tag{3}$$

この学習手順は，雑音抑圧により生じた音声歪み（例えば musical noise [4]）を増幅させてしまう．

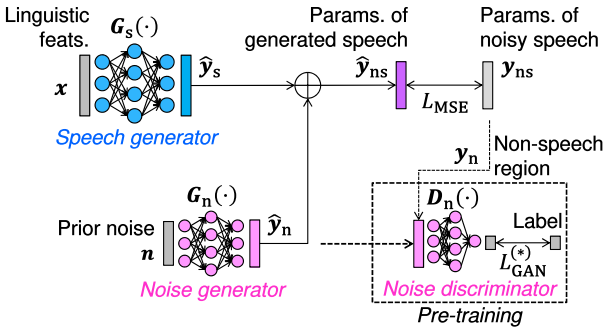

Fig. 1 提案法の DNN アーキテクチャ．雑音生成モデル $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ は，観測雑音を確率的に生成する．

3 提案法

提案法の DNN アーキテクチャを Fig. 1 に示す．音声合成モデル $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ に加え，雑音生成モデル $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ を導入する．$\boldsymbol{G}_{\mathrm{n}}(\cdot)$ は，既知の事前分布を観測雑音の分布に変形する役割を持ち，雑音スペクトルを確率的に生成する．$\boldsymbol{G}_{\mathrm{s}}(\cdot)$ は，その雑音スペクトルを加算した後のスペクトルが雑音環境下音声のスペクトルに一致するように学習される．予備実験において，雑音環境下音声を用いた $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ と $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ の同時学習を試みたが，雑音抑圧効果が低かった．故に本稿では，まず，$\boldsymbol{y}_{\mathrm{n}}$ を用いて，その分布を表現する $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ を事前学習し，その後，$\boldsymbol{G}_{\mathrm{n}}(\cdot)$ のモデルパラメータを固定し，雑音環境下音声を用いて $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ の学習を行う．$\boldsymbol{G}_{\mathrm{n}}(\cdot)$ の学習には，敵対的学習アルゴリズム [2] を使用する．

3.1 敵対的学習による雑音生成モデルの学習

$\boldsymbol{G}_{\mathrm{n}}(\cdot)$ の入力は，既知の事前分布からランダム生成された変数 $\boldsymbol{n} = [\boldsymbol{n}_1^{\top}, \cdots, \boldsymbol{n}_t^{\top}, \cdots, \boldsymbol{n}_{T_{\mathrm{n}}}^{\top}]^{\top}$ である．$\boldsymbol{n}_t$ は，フレーム t において，事前分布からランダム生成されたベクトルである．$\boldsymbol{G}_{\mathrm{n}}(\cdot)$ は，観測雑音スペクトル $\boldsymbol{y}_{\mathrm{n}}$ と生成雑音スペクトル $\hat{\boldsymbol{y}}_{\mathrm{n}} = \boldsymbol{G}_{\mathrm{n}}(\boldsymbol{n})$ を識別する雑音識別モデル $\boldsymbol{D}_{\mathrm{n}}(\cdot)$ と交互に更新される．$\boldsymbol{G}_{\mathrm{n}}(\cdot)$ の損失関数 $L_{\mathrm{GAN}}^{(G)}(\cdot)$ と，$\boldsymbol{D}_{\mathrm{n}}(\cdot)$ の損失関数 $L_{\mathrm{GAN}}^{(D)}(\cdot)$

---

*Generative Approach Using the Noise Generation Models for DNN-based Speech Synthesis Trained from Noisy Speech, by UNE Masakazu (NIT Tokuyama College and Univ. of Tokyo), SAITO Yuki, TAKAMICHI Shinnosuke, KITAMURA Daichi (Univ. of Tokyo), MIYAZAKI Ryoichi (NIT Tokuyama College), and SARUWATARI Hiroshi (Univ. of Tokyo)

は，それぞれ次式で示される．

$$L_{\mathrm{GAN}}^{(G)}\left(\hat{\boldsymbol{y}}_{\mathrm{n}}\right) = -\frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}}\log D_{\mathrm{n}}(\hat{\boldsymbol{y}}_{\mathrm{n},t}) \tag{4}$$

$$\begin{aligned} L_{\mathrm{GAN}}^{(D)}\left(\boldsymbol{y}_{\mathrm{n}}, \hat{\boldsymbol{y}}_{\mathrm{n}}\right) = &-\frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}}\log D_{\mathrm{n}}(\boldsymbol{y}_{\mathrm{n},t}) \\ &-\frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}}\log\left(1 - D_{\mathrm{n}}(\hat{\boldsymbol{y}}_{\mathrm{n},t})\right) \end{aligned} \tag{5}$$

この学習は，$\boldsymbol{y}_{\mathrm{n}}$ と $\hat{\boldsymbol{y}}_{\mathrm{n}}$ の分布間の近似 Jensen-Shannon divergence を最小化する．

3.2 雑音生成モデルを用いた音声合成モデル学習

音声と雑音の位相情報を無視して，振幅ドメインにおける加法性が成り立つと仮定する．学習済みの $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ を用いて，次式の損失関数を最小化するように，音声合成モデル $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ を学習する．

$$L_{\mathrm{MSE}}\left(\hat{\boldsymbol{y}}_{\mathrm{ns}}, \boldsymbol{y}_{\mathrm{ns}}\right) = \frac{1}{T}\left(\hat{\boldsymbol{y}}_{\mathrm{ns}} - \boldsymbol{y}_{\mathrm{ns}}\right)^{\top}\left(\hat{\boldsymbol{y}}_{\mathrm{ns}} - \boldsymbol{y}_{\mathrm{ns}}\right) \tag{6}$$

$$\hat{\boldsymbol{y}}_{\mathrm{ns}} = \log\left(\exp\hat{\boldsymbol{y}}_{\mathrm{s}} + \exp\hat{\boldsymbol{y}}_{\mathrm{n}}\right) \tag{7}$$

ただし，ここでの $\hat{\boldsymbol{y}}_{\mathrm{n}}$ の系列長は T であることに注意する．生成時には，$\hat{\boldsymbol{y}}_{\mathrm{s}} = \boldsymbol{G}_{\mathrm{s}}(\boldsymbol{x})$ を，合成音声の対数振幅スペクトルとする．

提案法は，明示的な確率分布を定義せず，その経験分布を Generative Adversarial Network の枠組みで表現するため，従来の音声歪みを軽減できる．

4 主観評価実験

4.1 実験条件

利用する音声データは，無響室にて収録された，日本人女性 1 名による約 3000 文である．雑音環境下音声は，この収録音声データに対して白色雑音を人工的に加算したものとする．評価データは ATR 音素バランス 503 文，J セット 53 文である．音声合成モデル及び雑音生成モデルは，動的特徴量を含まない 257 次元の対数振幅スペクトルを予測する．コンテキスト特徴量は 444 次元のベクトルであり，439 次元の言語特徴量，3 次元の継続長特徴量，連続対数 F_0，及び有声無声ラベルである．雑音生成モデルに入力される $\boldsymbol{n}_t$ は各フレーム毎に 100 次元ベクトルであり，各次元の値は一様分布からランダムに生成される．音声合成モデル，雑音生成モデル，雑音識別モデルは，それぞれ Feed-Forward neural network で記述され，従来法と提案法で同様の音声合成モデルを使用する．各モデルの隠れ層数は 3，隠れ層の素子数は 512，隠れ層の活性化関数は，leaky ReLU である．音声合成モデルと雑音生成モデルの出力層の活性化関数は，線形関数である．雑音識別モデルの出力層の活性化関数は，sigmoid 関数である．

4.2 主観評価結果

実験的評価では，以下の 2 手法を比較する．

- **SS+MSE**: SS を施した後，平均二乗誤差最小化により音声合成モデルを学習
- **Proposed**: 提案法

雑音環境下音声の SN 比は，0dB，5dB，10dB とし，SS における減算係数 β を，$0.5, 1.0, 2.0, 5.0$ に設定する．β の値が小さいほど音声歪みは小さく，β の値が大きいほど音声歪みは大きい．評価として，各 SN 比，各 β の設定において，合成音声の自然性に関するプリファレンス AB テストを実施する．評価者には，より不快でなく，かつ，より自然な音声を選択させた．評価人数は各評価に対して 25 人，計 300 人である．

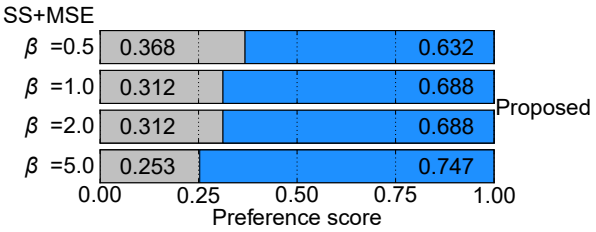

Fig. 2　音質に関する主観評価結果 (SNR = 0 dB)

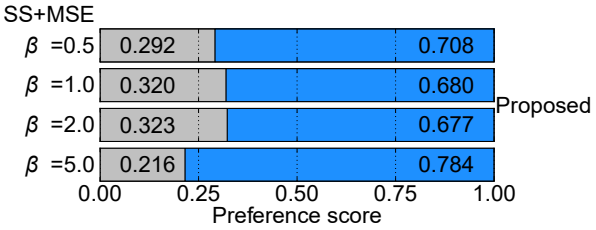

Fig. 3　音質に関する主観評価結果 (SNR = 5 dB)

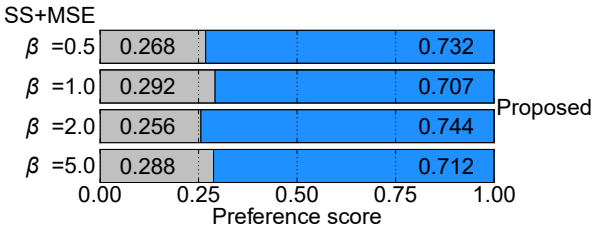

Fig. 4　音質に関する主観評価結果 (SNR = 10 dB)

Fig. 2 から Fig. 4 に評価結果を示す．全設定において提案法のスコアが従来法のスコアを上回り，その p 値は 10^{-6} より小さいため，提案法の有効性が示される．

5 まとめ

本稿では，雑音環境下音声を用いた高品質音声合成のために，雑音を確率的に生成する雑音生成モデルを導入し，雑音加算過程を考慮した音声合成モデル学習法を提案した．実験的評価では，　提案法による音質改善効果を示した．

今後の予定として，nonnegative matrix factorization のアクティベーション行列などによる時間変動のモデリングや，雑音混入強度の導入などが挙げられる．また，ボコーダを使用する合成方式との比較，クリーン音声を用いた適応学習を行う．

謝辞: 本研究の一部は，セコム科学技術支援財団の助成を受け実施した．

参考文献

[1] H. Zen *et al.*, *Proc. ICASSP*, vol. 51, no. 11, pp. 7962–7966, Vancouver, Canada, May, 2013.

[2] I. Goodfellow *et al.*, *Proc. NIPS*, pp. 2672–2680, 2014.

[3] S. F. Boll, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] R. Miyazaki *et al.*, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2080–2094, 2012.