雑音環境下音声を用いた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[1,2,a)]　齋藤 佑樹[2,b)]　高道 慎之介[2,c)]　北村 大地[2,d)]　宮崎 亮一[1,e)]　猿渡 洋[2,f)]

概要：高品質な統計的パラメトリック音声合成システムの構築には，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，現存する膨大な音声データのうち，音声合成の学習に利用可能なもの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．従来，そのような音声を学習データとして用いる場合，spectral subtraction 等の雑音抑圧処理を施した後に，通常の音声合成の学習を行う．しかしながら，雑音スペクトルの生成分布をパラメトリックに定義する雑音抑圧法は処理後の音声を歪ませ，さらに，その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．そこで本稿では，敵対的学習アルゴリズムにより学習される雑音生成モデルを用いた，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．テキストから音声スペクトルを生成する音声合成モデルは，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，雑音スペクトルの生成分布を柔軟にモデル化でき，さらに，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，従来法において生じる品質低下を低減できる．実験的評価では，いくつかの雑音抑圧強度と SN 比において合成音声を作成し，提案法の知覚的音質が従来法を上回ることを示す．

Generative adversarial training of the noise generation model for speech synthesis using speech in noise

Masakazu Une[1,2,a)]　Yuki Saito[2,b)]　Shinnnosuke Takamichi[2,c)]　Daichi Kitamura[2,d)]
Ryoichi Miyazaki[1,e)]　Hiroshi Saruwatari[2,f)]

1. はじめに

統計的パラメトリック音声合成 [1] は統計モデルを使用してテキストから音声を合成する方法であり，音声合成の最終目標の 1 つは，人間の発話のように自然な音声を合成することである．音声品質は自然性の要素の 1 つであり，合成音声の品質向上のための様々な方法が提案されている [2], [3], [4]．特に，Deep Neural Network（DNN）に基づく音声合成 [5] は，合成音声の品質を著しく向上させている．しかしながら，高品質な統計的パラメトリック音声合成システムを構築するためには，スタジオ等の理想的な環境で収録された音声データを利用することが必須である．そのため，現存する膨大の音コーパス [6] や，地理的理由により劣悪環境で収録された音声コーパス [7] 等を利用することは，現状困難である．音声合成による音声コミュニケーションの拡張のためには，このような劣悪環境下の学習データからでも高品質な音声合成を構築する必要がある．劣悪環境の種類として，狭帯域 [8]，劣悪通信経路 [9] も挙げられるが，本稿では，CPJD (Crowdsourced speech corpora of Parallel Japanese Dialect) コーパス [7] を参考

[1] 徳山工業高等専門学校
National Institute of Technology, Tokuyama College, where?
[2] 東京大学　大学院情報理工学系研究科
The University of Tokyo, Engineering bldg. #6, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] i12une@tokuyama.ac.jp
[b)] yuuki_saito@ipc.i.u-tokyo.ac.jp
[c)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[d)] daichi_kitamura@ipc.i.u-tokyo.ac.jp
[e)] miyazaki@tokuyama.ac.jp
[f)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

にして，一般家庭環境において収録されたような，定常雑音の混入した音声を対象とする．

雑音環境下音声を統計的パラメトリック音声合成の学習データとして用いる場合，通常，その前処理として雑音抑圧を行う（Fig. 1 上）．ただし，音声合成のための雑音抑圧は，最終的にボコーダパラメータ（例えば，STRAIGHT [10] や WORLD [11] により抽出されたスペクトル包絡）を得る必要があるため，音声認識で用いられる一般的な雑音抑圧と異なる．音声合成のための雑音抑圧は，大きく二つに分けられる．ひとつは，雑音環境下音声からボコーダパラメータを直接的に推定する方法である [12]．この場合，雑音データベースを別途用意して，雑音環境下音声からボコーダパラメータを推定する統計モデルを事前に構築する．この手法は，DNN 等の利用により非線形変換を可能にするが，未知雑音に対する頑健性を保証しない．もうひとつの方法は，信号処理ベースの雑音抑圧を施した後に，通常の方法でボコーダパラメータを抽出する方法である．Spectral subtraction [13] などの教師なし雑音抑圧は，未知雑音に対しても頑健に動作するが，雑音抑圧後の音声波形に対するボコーダパラメータ抽出の頑健性を保証しない．一方で本稿では，ボコーダフリー DNN 音声合成方式を用いて，雑音環境下音声からの音声合成の構築を試みる．ボコーダフリー DNN 音声合成は，ボコーダパラメータではなく，スペクトルや音声波形を直接推定する枠組みである [14], [15], [16]．我々は，この方式の利用により，通常の雑音抑圧で用いられる音源モデルや雑音加算過程を考慮した音声合成学習が可能になると考える．

本稿では，テキストから音声スペクトルを生成する音声合成モデル（通常，このモデルは音響モデルと呼ばれるが，後述の雑音生成モデルと対比させるため音声合成モデルと定義する）と，定常雑音を確率的に生成する雑音生成モデルを用いて，雑音環境下音声から高品質音声合成を構築する方法を提案する．提案法で導入される雑音生成モデルは，敵対的学習 [17] の枠組みを用いて，学習データに含まれる定常雑音スペクトルの統計量を推定する．音声合成モデルは，雑音生成モデルから確率的に生成される雑音スペクトルと音声合成モデルから生成されるスペクトルの和が，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．雑音成分の分布を期待値で近似する従来の spectral subtraction にくらべ，提案法は，雑音環境下音声から確率分布をデータドリブンに推定するため，より精微な雑音モデリングが可能である．また，雑音加算過程を考慮して音声合成モデルを学習する（Fig. 1 下）ため，音声スペクトルの歪みを減らし，より高品質な音声合成の構築が可能となる．実験的評価では，いくつかの雑音抑圧強度と SN 比において合成音声を作成し，提案法の知覚的音質が従来法を上回ることを示す．

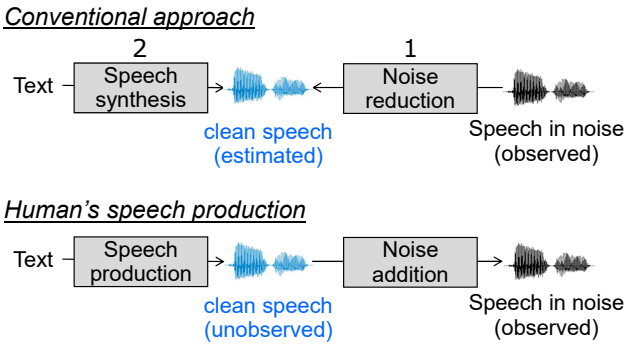

図 1　雑音環境下音声を用いた音声合成の学習手順．従来は，観測された雑音環境下音声に対して雑音抑圧処理を施した後，その推定されたクリーン音声を生成するように音声合成モデルの学習を行う．一方で提案法は，雑音加算過程を考慮して雑音環境下音声を直接的に生成するよう，音声合成モデルを学習する．

Fig. 1　Procedures of speech synthesis training using noisy speech. In the conventional way, noise reduction is first performed, then, the speech generator (i.e., acoustic model) is trained to predict the noise-reduced speech parameters. Our method directly predicts the noisy speech parameters, considering noise addition process.

2. Spectral subtraction による雑音抑圧と mean squared error 最小化による音声合成モデル学習

雑音環境下音声に対して spectral subtraction による雑音抑圧処理を施した後，mean squared error 最小化による音声合成モデルを行う．

2.1 Spectral subtraction による雑音抑圧

Spectral subtraction [13] 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 $\boldsymbol{y}_{\mathrm{ns}} = [\boldsymbol{y}_{\mathrm{ns},t}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{ns},t}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{ns},T}^{\top}]^{\top}$ 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．

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

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

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．

2.2 Mean squared error 最小化による音声合成モデル学習

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．$\boldsymbol{G}_{\mathrm{s}}(\cdot)$ は neural network で記述される [5], [16]．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 $\boldsymbol{y}^{(\mathrm{SS})}_{\mathrm{ns}}$ 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．

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

2.3 問題点

Spectral subtraction は，確率的に加算される雑音の分布を期待値で近似するため，処理後の音声の分布を大きく歪ませる．また，musical noise と呼ばれる聴覚的に不快な音 [18] を生成する．更に，この推定誤差は，後段の音声合成モデルの学習時に，その推定値を大きく歪ませる．

3. 提案法：雑音生成モデルを利用した音声合成モデル学習

提案法の DNN アーキテクチャを Fig. 2 に示す．従来法の音声合成モデル $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ 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．音声合成モデル $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ 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，観測雑音の対数振幅スペクトル $\boldsymbol{y}_{\mathrm{n}}$ を用いて，その分布を表現する雑音生成モデル $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ 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音声合成モデル $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ 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 [17] を使用する．

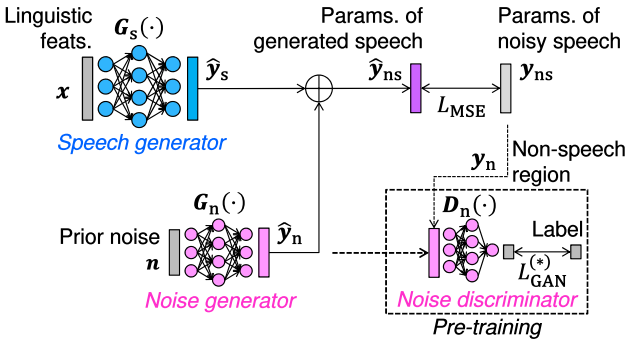

図 2　提案法の DNN アーキテクチャ．雑音生成モデル $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ は，観測雑音を確率的に生成する．

Fig. 2　Architectures of the proposed method. The noise generation model $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ randomly samples the noise.

3.1 敵対的学習による雑音生成モデルの学習

敵対的学習アルゴリズムにより雑音生成モデル $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ を学習する．$\boldsymbol{G}_{\mathrm{n}}(\cdot)$ 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 $\boldsymbol{y}_{\mathrm{n}}$ と生成雑音 $\hat{\boldsymbol{y}}_{\mathrm{n}} = \boldsymbol{G}_{\mathrm{n}}(\boldsymbol{n})$ 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 $L^{(G)}_{\mathrm{GAN}}(\cdot)$ と，$\boldsymbol{D}_{\mathrm{n}}(\cdot)$ の損失関数 $L^{(D)}_{\mathrm{GAN}}(\cdot)$ は，それぞれ次式で示される．

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

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

敵対的学習は，$\boldsymbol{y}_{\mathrm{n}}$ と $\hat{\boldsymbol{y}}_{\mathrm{n}}$ の分布間の近似 Jensen-Shannon divergence を最小化する．学習後の $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ 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つ．

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．

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

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

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．合成音声波形は，Griffin-Lim の位相復元アルゴリズム [19] を用いて生成する．

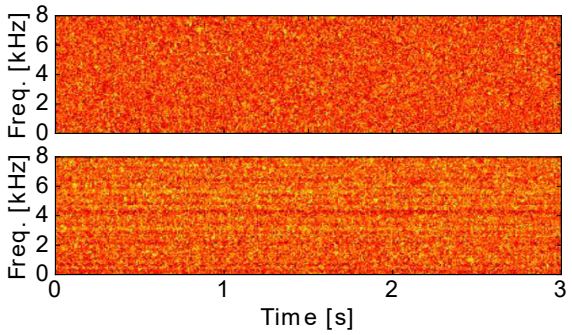

図 3 観測雑音（上）と生成雑音（下）のスペクトログラム．生成雑音は，各フレームごとに独立に生成している．

Fig. 3 Spectrograms of observed noise (above) and generated noise (below). The generated noise is sampled frame by frame independently.

3.3 考察

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 (GAN) の枠組みを用いて表現する．故に，musical noise などの歪みを低減し，また，Fig. 3 に示すように，部分的に誤りを観測できるものの観測雑音を効果的に表現できる．雑音生成モデルは，各フレームごとに独立な定常雑音スペクトルを生成するが，条件付き GAN [20] やリカレント構造を持った neural network 生成モデルの導入により，コンテキスト依存性・時間構造の考慮が可能である．

4. 実験的評価

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 [21] J セット 53 文である．学習データのサンプリング周波数は 16 kHz である．フレーム分析の窓長，シフト長，FFT 長は，それぞれ，400 サンプル (25 ms), 80 サンプル (5 ms), 512 サンプルとする．窓関数はハミング窓とする．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．合成音声波形は，予測した対数振幅スペクトルに対して Griffin-Lim による位相復元 [19] を施し生成する．ただし，予備実験より，従来法と提案法ともに合成音声に残留雑音が含まれることが確認されたため，従来法と提案法の生成した振幅スペクトル系列に対して，音声成分を知覚的に歪ませない程度の spectral subtraction を適用した．ケプストラム [22]，系列内変動 [23]，変調スペクトル [3] に基づく強調処理は行わない．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．実応用時にこの継続長特徴量，連続対数 F_0，及び有声無声ラベルは雑音環境下音声から抽出されるが，この特徴抽出による音声品質の低下 [24] をさけるため，本稿では，これらの特徴量を雑音加算前の音声から抽出する．学習時には，コンテキスト $\boldsymbol{x}$ 及び雑音環境下音声の対数振幅スペクトル $\boldsymbol{y}_{\mathrm{ns}}$ を，0 平均 1 分散に正規化する．生成時には，$\hat{\boldsymbol{y}}_{\mathrm{s}} = \boldsymbol{G}_{\mathrm{s}}(\boldsymbol{x})$ を生成した後，$\boldsymbol{y}_{\mathrm{ns}}$ の統計量を用いて元のスケールに戻す．この処理は本来，不良設定問題であるため（$\boldsymbol{y}_{\mathrm{ns}}$ のスケーリングのみが既知で，その構成要素である $\boldsymbol{y}_{\mathrm{n}}$ と $\boldsymbol{y}_{\mathrm{s}}$ をスケーリングするため），この正規化処理・スケーリング処理は，今後改善する必要がある．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．音声合成モデルの学習時には，非音声区間の 90% を除外する．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 [25] 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．DNN のモデルパラメータは乱数で初期化する．最適化アルゴリズムには AdaGrad [26] を使用する．

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．本評価は，ボコーダフリー音声合成の枠組みにおける比較を目的とするため，ボコーダを用いる合成法を対象から除外する．

- **SS+MSE**: spectral subtraction 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

SN 比は，CPJD コーパス [7] において多く含まれる 0dB, 5dB，10dB とする．ただし，音声認識のための雑音抑圧において音声歪み（または残留雑音量）と音声認識精度の関係性が知られており [27] ，同様の議論が音声合成においても必要であると思われる．そこで，spectral subtraction における 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，従来法と提案法の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．評価は我々のクラウドソーシング評価システム上で実施し，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. 4 から Fig. 6 にそれぞれ，0dB，5dB，10dB の SN 比における結果を示す．図より，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っていることが分かる．また，全設定において，従来法と提案法のスコア間の p 値が 10^{-6} を下回っているため，提案法の有効性が示された．

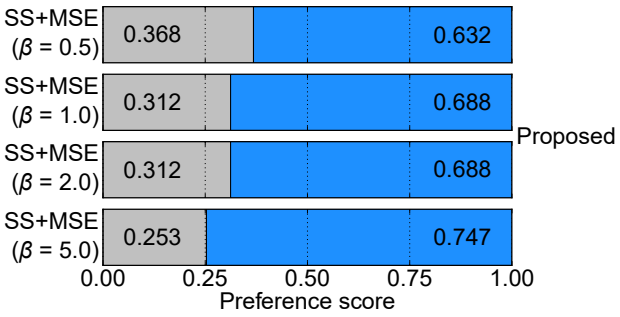

図 4　合成音声品質に関するプリファレンススコア (SNR = 0 dB)
Fig. 4　Preference scores on synthetic speech quality (SNR = 0 dB).

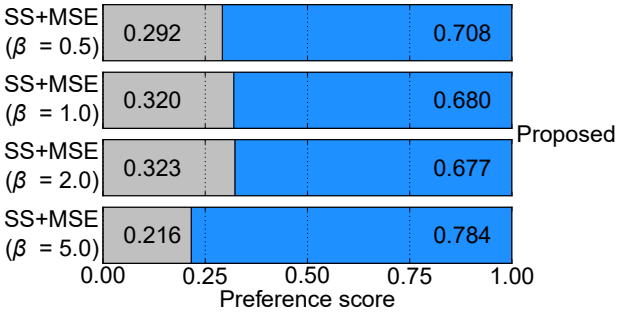

図 5　合成音声品質に関するプリファレンススコア (SNR = 5 dB)
Fig. 5　Preference scores on synthetic speech quality (SNR = 5 dB).

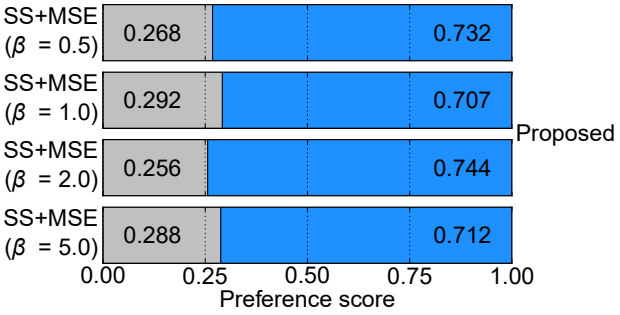

図 6　合成音声品質に関するプリファレンススコア (SNR = 10 dB)
Fig. 6　Preference scores on synthetic speech quality (SNR = 10 dB).

0dB の結果（Fig. 4）において，β を大きくすると従来法のプリファレンススコアが悪化していることが分かる．これに関して我々は，SN 比が低い場合に，spectral subtraction により生じた過剰な音声歪みが，音声合成品質を劣化させることを確認している．

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．雑音生成モデルは，敵対的学習を用いて，観測される定常雑音の確率分布を表現するように学習され，音声合成モデルは，その生成スペクトルと雑音生成モデルの生成したスペクトルの和が，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．実験的評価では，spectral subtraction による雑音抑圧と通常の音声合成モデル学習を組み合わせた従来法と比較して，提案法が有意に合成音声品質を改善させることを明らかにした．

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 [28] 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，JSPS 科研費 16H06681 及び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, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
[2] S. Takamichi, K. Tomoki, and H. Saruwatari, “Sampling-based speech parameter generation using moment-matching network,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.
[3] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, “Postfilters to modify the modulation spectrum for statistical parametric speech synthesis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.
[4] Y. Saito, S. Takamichi, and H. Saruwatari, “Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis,” in *Proc. ICASSP*, Orleans, U.S.A., Mar. 2017.
[5] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*, Vancouver, Canada, May 2013.
[6] S. A.-E.-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “YouTube-8M: A large-scale video classification benchmark,” vol. abs/1609.08675, 2016. [Online]. Available: https://arxiv.org/abs/1609.08675
[7] 高道慎之介 and 猿渡洋, “クラウドソーシングを利用した対訳方言音声コーパスの構築,” in 日本音響学会 *2017* 年秋季研究発表会講演論文集, 愛媛, Sep. 2017.
[8] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, “Statistical bandwidth extension for speech synthesis based on Gaussian mixture model with sub-band basis spectrum model,” *IEICE Transactions on Information and Systems*, vol. E99-D, no. 10, pp. 2481–2489, 2016.
[9] A. Saeb, R. Menon, H. Cameron, W. Kibira, J. Quinn, and T. Niesler, “Very low resource radio browsing for agile developmental and humanitarian monitoring,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 2118–2122.
[10] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and

an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[11] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[12] C. V.-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. INTERSPEECH*, Sep. 2016, pp. 352–356. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-159

[13] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," vol. abs/1609.03499, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[16] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proc. NIPS*, pp. 2672–2680, 2014.

[18] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2080–2094, Sep. 2012.

[19] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2015.

[21] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara, "A large-scale Japanese speech database," in *ICSLP90*, Kobe, Japan, Nov. 1990, pp. 1089–1092.

[22] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, Budapest, Hungary, Apr. 1999, pp. 2347–2350.

[23] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 816–824, 2007.

[24] P. Baljekar and A. W. Black, "Utterance selection techniques for TTS systems using found speech," in *Proc. SSW9*, Sunnyvale, CA, USA, Sep. 2016, pp. 199–204.

[25] L. A. Maas, Y. A. Hannun, and Y. A. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013.

[26] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 2121–2159, 2011.

[27] 藤本雅清, "Factored deep convolutional neural networks による雑音下音声認識," in 電子情報通信学会技術報告 *SP2017-18*, vol. 117, no. 160, 宮城, Jul. 2017.

[28] 坂東宜昭, 三村正人, 糸山克寿, 吉井和佳, and 河原達也, "深層生成モデルを事前分布に用いた教師なし音声強調," in 電子情報通信学会技術報告 *SP2017-20*, vol. 117, no. 189, 京都, Aug. 2017.