敵対的 DNN 音声合成におけるダイバージェンスの影響の調査*

◎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では，音響モデルから生成される合成音声パラメータ系列の過剰な平滑化により，音質の劣化が生じる．これまでに我々は，Deep Neural Network (DNN) 音声合成のための，Generative Adversairal Network (GAN) [1] の枠組みに基づく音響モデル学習 (敵対的 DNN 音声合成) [2] を提案し，その音質改善効果を確認している．GAN の学習では，自然音声パラメータと合成音声パラメータ分布間距離を最小化するため，合成音声パラメータの分布の縮小を緩和することが可能である．

これまでに，Goodfellow ら [1] によって提案された GAN を始めとして，数多くの GAN が提案されている．しかし，これらの GAN の有効性は主に画像生成においてのみ検証されており，同様の枠組みが敵対的 DNN 音声合成でも有効かは調査されていない．そこで，本稿では，各 GAN で最小化される距離規範の違いに着目し，それらが敵対的 DNN 音声合成の音質に与える影響を調査する．本稿では，音声信号処理との関連性が高い f ダイバージェンス最小化に基づく GAN (f-GAN) [3] と，画像生成において有効である Wasserstein GAN (W-GAN) [4]，Least Squares GAN (LS-GAN) [5] を採用した敵対的 DNN 音声合成を比較する．実験的評価では，種々の距離規範の中で，Earth Mover 距離最小化に基づく W-GAN が音質改善に最も有効であることを示す．

2 従来の枠組み

2.1 Minimum Generation Error (MGE) 学習による音響モデル学習

従来の MGE 学習 [6] による音響モデル学習の損失関数 $L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は，自然音声のパラメータ系列 $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ と，最尤パラメータ生成 [7] 後の合成音声のパラメータ系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ の間の二乗誤差として次式で与えられる．

$$
\begin{aligned}
L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{T}(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top (\hat{\boldsymbol{y}} - \boldsymbol{y}) \\
&= \frac{1}{T}\left(\boldsymbol{R}\hat{\boldsymbol{Y}} - \boldsymbol{y}\right)^\top \left(\boldsymbol{R}\hat{\boldsymbol{Y}} - \boldsymbol{y}\right) \qquad (1)
\end{aligned}
$$

ここで，t はフレームインデックス，T は総フレーム数，$\boldsymbol{y}_t$ はフレーム t における音声パラメータ，$\hat{\boldsymbol{Y}}$ は音響モデルの予測結果として得ら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・動的特徴量の制約行列 [7] 及び学習データを用いて別途推定される分散・共分散行列 $\boldsymbol{\Sigma}$ を用いて次式で計算される．

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

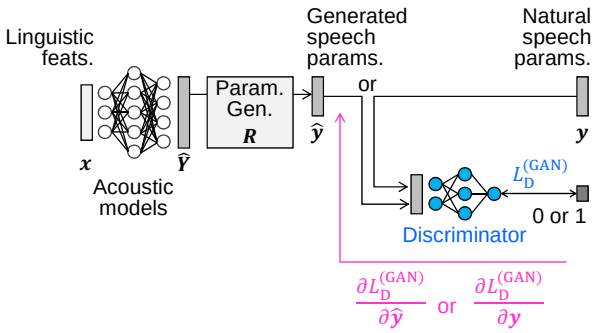

Fig. 1 識別モデルパラメータの更新時の計算フロー（Param. Gen. は最尤パラメータ生成 [7]）

以降，この音声パラメータ生成を $\hat{\boldsymbol{y}} = \boldsymbol{R}\hat{\boldsymbol{Y}} = \boldsymbol{G}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{G}})$ と定義する．$\boldsymbol{x}$ は，入力テキストのコンテキスト系列である．音響モデルのモデルパラメータ $\boldsymbol{\theta}_{\mathrm{G}}$ は，式 (1) の生成誤差の勾配 $\nabla_{\boldsymbol{\theta}_{\mathrm{G}}} L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を用いた backpropagation により更新される．勾配計算に必要な $\nabla_{\hat{\boldsymbol{Y}}} L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は $\boldsymbol{R}^\top(\hat{\boldsymbol{y}} - \boldsymbol{y})/T$ として計算される [6]．

2.2 GAN [1] の学習

GAN は DNN を用いた生成モデルを学習する手法であり，生成モデルと識別モデル $D(\boldsymbol{y}; \boldsymbol{\theta}_{\mathrm{D}})$ の 2 つの DNN を学習する．ここで，$\boldsymbol{\theta}_{\mathrm{D}}$ は識別モデルのモデルパラメータである．識別モデルの出力 $D(\boldsymbol{y})$ に sigmoid 関数を適用した値 $1/(1 + \exp(-D(\boldsymbol{y})))$ は，入力が真のデータである事後確率を表す．識別モデルは，真のデータに対して事後確率が 1，生成モデルから生成されたデータに対して事後確率が 0 となるように学習される．一方で，生成モデルは識別モデルを詐称するように学習される．すなわち，生成データに対して識別モデルの事後確率が 1 となるように学習される．

GAN の学習では，識別モデルと生成モデルが交互に更新される．まず，真のデータ $\boldsymbol{y}$ と生成データ $\hat{\boldsymbol{y}}$ を用いて，識別モデルの更新に用いる識別損失（cross-entropy 関数）$L_{\mathrm{D}}^{(\mathrm{GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を計算する．

$$
\begin{aligned}
L_{\mathrm{D}}^{(\mathrm{GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = &-\frac{1}{T}\sum_{t=1}^{T} \log \frac{1}{1 + \exp(-D(\boldsymbol{y}_t))} \\
&- \frac{1}{T}\sum_{t=1}^{T} \log \frac{\exp(-D(\hat{\boldsymbol{y}}_t))}{1 + \exp(-D(\hat{\boldsymbol{y}}_t))} \qquad (3)
\end{aligned}
$$

識別モデルのモデルパラメータ $\boldsymbol{\theta}_{\mathrm{D}}$ は，式 (3) の識別損失の勾配 $\nabla_{\boldsymbol{\theta}_{\mathrm{D}}} L_{\mathrm{D}}^{(\mathrm{GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を用いた backpropagation により更新される．Figure 1 に $\boldsymbol{\theta}_{\mathrm{D}}$ の更新時の計算フローを示す．識別モデルの更新後，生成モデ

---

*Experimental Investigation of Divergences in Adversarial DNN-Based Speech Synthesis, by SAITO, Yuki, TAKAMICHI, Shinnosuke, and SARUWATARI, Hiroshi (The University of Tokyo).

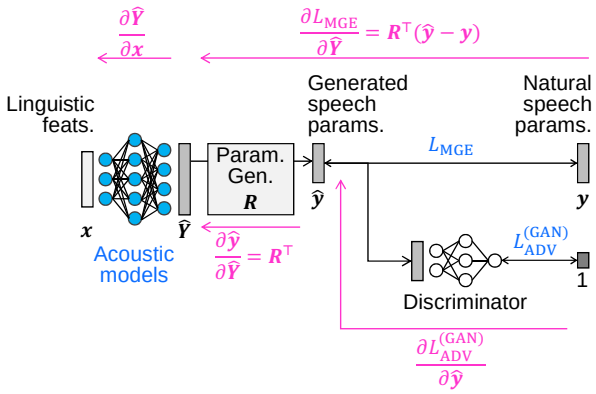

Fig. 2　音響モデルパラメータの更新時の計算フロー

ルの更新に用いる敵対損失 $L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}})$ を計算する．

$$L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T}\log\frac{1}{1+\exp\left(-D\left(\hat{\boldsymbol{y}}_t\right)\right)} \tag{4}$$

生成モデルのモデルパラメータ $\boldsymbol{\theta}_{\mathrm{G}}$ は，式（4）の敵対損失の勾配 $\nabla_{\boldsymbol{\theta}_{\mathrm{G}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}})$ を用いた backpropagation により更新される．この枠組みにより，真のデータの分布と生成されたデータの分布間の近似 Jensen–Shannon (JS) ダイバージェンスが最小化される [1]．

2.3　Goodfellow らの GAN [1] を用いた敵対的 DNN 音声合成 [2]

Goodfellow らの GAN [1] を用いた敵対的 DNN 音声合成 [2] において，音響モデル（生成モデル）の損失関数 $L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は次式で与えられる．

$$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}}\frac{E_{L_{\mathrm{MGE}}}}{E_{L_{\mathrm{ADV}}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) \tag{5}$$

式（5）の第二項 $L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}})$ は，合成音声パラメータと自然音声パラメータの分布間距離（近似 JS ダイバージェンス）を最小化する．故に，敵対的 DNN 音声合成は，パラメータの生成誤差 $L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を最小化するだけでなく，合成音声パラメータの分布を自然音声のものに近づける．$E_{L_{\mathrm{MGE}}}$ と $E_{L_{\mathrm{ADV}}}$ はそれぞれ $L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}})$, の期待値であり，2 つの損失のスケールを調整する役割を持つ．ω_{D} は，敵対損失の影響を調整するハイパーパラメータである．音響モデルのモデルパラメータ $\boldsymbol{\theta}_{\mathrm{G}}$ は，式（5）の勾配 $\nabla_{\boldsymbol{\theta}_{\mathrm{G}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}})$ を用いた backpropagation により更新される．Figure 2 に $\boldsymbol{\theta}_{\mathrm{G}}$ の更新時の計算フローを示す．GAN の学習と同様に，音響モデルと識別モデルは交互に更新され，一方の更新中に他方のモデルパラメータは固定される．

3　敵対的 DNN 音声合成に用いる種々の距離規範

GAN の枠組みは真のデータと生成データの分布間の距離規範最小化とみなされる．2.2 節で述べたように，Goodfellow らの GAN [1] は近似 JS ダイバージェンスを最小化する．この分布間距離最小化の観点に基づき，様々な GAN を敵対的 DNN 音声合成に導入する．本稿では，非負値行列因子分解 [12, 13] における損失関数としてしばしば採用される Kullback–Leibler (KL) ダイバージェンスを含む f ダイバージェンスの最小化に基づく GAN (f-GAN) [3] と，画像生成において有効である Wasserstein GAN (W-GAN) [4] 及び Least Squares GAN (LS-GAN) [5] を導入する．以降で導入する識別モデルの識別損失 $L_{\mathrm{D}}^{(*\text{-GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と，音響モデルの敵対損失 $L_{\mathrm{ADV}}^{(*\text{-GAN})}(\hat{\boldsymbol{y}})$ は，それぞれ式（3）と式（4）の代わりに用いられる．

3.1　f-GAN [3]

f-GAN は，Goodfellow らの GAN [1] を包含する統一的な枠組みを記述する．分布間の距離は，KL ダイバージェンスや JS ダイバージェンスを含む f-ダイバージェンス $\mathcal{D}_f(\boldsymbol{y}\|\hat{\boldsymbol{y}})$ により次式で定義される．

$$\mathcal{D}_f(\boldsymbol{y}\|\hat{\boldsymbol{y}}) = \int q(\hat{\boldsymbol{y}})\, f\left(\frac{p(\boldsymbol{y})}{q(\hat{\boldsymbol{y}})}\right) d\boldsymbol{y} \tag{6}$$

ここで，$p(\cdot)$ と $q(\cdot)$ はそれぞれ $\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ の確率密度である．$f(\cdot)$ は $f(1) = 0$ を満たす凸関数である．$f(\cdot)$ の選択により様々なダイバージェンスを表すことが可能だが，本稿では音声信号処理との関連性が高いものを採用する．

3.1.1　KL-GAN

KL ダイバージェンスは，$f(r) = r\log r$ と定義することで次式で与えられる．

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{y}\|\hat{\boldsymbol{y}}) = \int p(\boldsymbol{y})\log\frac{p(\boldsymbol{y})}{q(\hat{\boldsymbol{y}})} d\boldsymbol{y} \tag{7}$$

識別モデルの識別損失 $L_{\mathrm{D}}^{(\mathrm{KL\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は

$$\begin{aligned} L_{\mathrm{D}}^{(\mathrm{KL\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = &-\frac{1}{T}\sum_{t=1}^{T} D\left(\boldsymbol{y}_t\right) \\ &+ \frac{1}{T}\sum_{t=1}^{T}\exp\left(D\left(\hat{\boldsymbol{y}}_t\right) - 1\right) \end{aligned} \tag{8}$$

であり，音響モデルの敵対損失 $L_{\mathrm{ADV}}^{(\mathrm{KL\text{-}GAN})}(\hat{\boldsymbol{y}})$ は

$$L_{\mathrm{ADV}}^{(\mathrm{KL\text{-}GAN})}(\hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T} D\left(\hat{\boldsymbol{y}}_t\right) \tag{9}$$

として与えられる．

3.1.2　Reversed KL (RKL)-GAN

KL ダイバージェンスは非対称な距離規範であるため，その双対な距離規範である reversed KL (RKL) ダイバージェンス $\mathcal{D}_{\mathrm{RKL}}(\boldsymbol{y}\|\hat{\boldsymbol{y}})$ は $\mathcal{D}_{\mathrm{KL}}(\boldsymbol{y}\|\hat{\boldsymbol{y}})$ と異なる．RKL ダイバージェンスは，$f(r) = -\log r$ と定義することで次式で与えられる．

$$\mathcal{D}_{\mathrm{RKL}}(\boldsymbol{y}\|\hat{\boldsymbol{y}}) = \int q(\hat{\boldsymbol{y}})\log\frac{q(\hat{\boldsymbol{y}})}{p(\boldsymbol{y})} d\boldsymbol{y} = \mathcal{D}_{\mathrm{KL}}(\hat{\boldsymbol{y}}\|\boldsymbol{y}) \tag{10}$$

識別モデルの識別損失 $L_{\mathrm{D}}^{(\mathrm{RKL\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は

$$L_{\mathrm{D}}^{(\mathrm{RKL\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T}\sum_{t=1}^{T}\exp\left(-D\left(\boldsymbol{y}_t\right)\right) + \frac{1}{T}\sum_{t=1}^{T}\left(-1 + D\left(\hat{\boldsymbol{y}}_t\right)\right) \tag{11}$$

であり，音響モデルの敵対損失 $L_{\mathrm{ADV}}^{(\mathrm{RKL\text{-}GAN})}(\hat{\boldsymbol{y}})$ は

$$L_{\mathrm{ADV}}^{(\mathrm{RKL\text{-}GAN})}(\hat{\boldsymbol{y}}) = \frac{1}{T}\sum_{t=1}^{T}\exp\left(-D\left(\hat{\boldsymbol{y}}_t\right)\right) \tag{12}$$

として与えられる．

3.1.3 JS-GAN

近似なしの JS ダイバージェンスの最小化は，f-GAN の枠組みで記述可能である．JS ダイバージェンスは，$f(r) = -(r+1)\log\frac{r+1}{2} + r\log r$ と定義することで次式で与えられる．

$$\mathcal{D}_{\mathrm{JS}}\left(\boldsymbol{y}\|\hat{\boldsymbol{y}}\right) = \frac{1}{2}\int p\left(\boldsymbol{y}\right)\log\frac{2p\left(\boldsymbol{y}\right)}{p\left(\boldsymbol{y}\right)+q\left(\hat{\boldsymbol{y}}\right)}d\boldsymbol{y} + \frac{1}{2}\int q\left(\hat{\boldsymbol{y}}\right)\log\frac{2q\left(\hat{\boldsymbol{y}}\right)}{p\left(\boldsymbol{y}\right)+q\left(\hat{\boldsymbol{y}}\right)}d\boldsymbol{y} \tag{13}$$

識別モデルの識別損失 $L_{\mathrm{D}}^{(\mathrm{JS\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は

$$L_{\mathrm{D}}^{(\mathrm{JS\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T}\log\frac{2}{1+\exp\left(-D\left(\boldsymbol{y}_t\right)\right)} - \frac{1}{T}\sum_{t=1}^{T}\log\frac{2\exp\left(-D\left(\hat{\boldsymbol{y}}_t\right)\right)}{1+\exp\left(-D\left(\hat{\boldsymbol{y}}_t\right)\right)} \tag{14}$$

であり，音響モデルの敵対損失 $L_{\mathrm{ADV}}^{(\mathrm{JS\text{-}GAN})}(\hat{\boldsymbol{y}})$ は

$$L_{\mathrm{ADV}}^{(\mathrm{JS\text{-}GAN})}(\hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T}\log\frac{2}{1+\exp\left(-D\left(\hat{\boldsymbol{y}}_t\right)\right)} \tag{15}$$

として与えられる．Goodfellow らの GAN [1] は，$2\mathcal{D}_{\mathrm{JS}}(\boldsymbol{y}\|\hat{\boldsymbol{y}}) - \log(4)$ を最小化する．

3.2 Wasserstein GAN (W-GAN) [4]

Goodfellow らの GAN [1] の学習の不安定さを改善させるために，W-GAN が提案されている．W-GAN の学習では，次式に示す Earth-Mover 距離 (Wasserstein-1) を最小化する．

$$\mathcal{D}_{\mathrm{EM}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = \inf_{\gamma}\mathbb{E}_{(\boldsymbol{y},\hat{\boldsymbol{y}})\sim\gamma}\left[\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|\right] \tag{16}$$

ここで，$\gamma(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は周辺分布がそれぞれ $\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ の分布となるような結合分布である．Kantorovich–Rubinstein の定理 [14] を用いると, 識別モデルの識別損失 $L_{\mathrm{D}}^{(\mathrm{W\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は

$$L_{\mathrm{D}}^{(\mathrm{W\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T}D\left(\boldsymbol{y}_t\right) + \frac{1}{T}\sum_{t=1}^{T}D\left(\hat{\boldsymbol{y}}_t\right) \tag{17}$$

であり，音響モデルの敵対損失 $L_{\mathrm{ADV}}^{(\mathrm{W\text{-}GAN})}(\hat{\boldsymbol{y}})$ は

$$L_{\mathrm{ADV}}^{(\mathrm{W\text{-}GAN})}(\hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T}D\left(\hat{\boldsymbol{y}}_t\right) \tag{18}$$

として与えられる．W-GAN では，識別モデルが K-Lipschits 関数である必要がある．そのため，識別モデルの更新後に，重みパラメータの値を $[-0.01, 0.01]$ のような一定区間内に収める必要がある．

3.3 Least Squares GAN (LS-GAN) [5]

Goodfellow らの GAN [1] では，sigmoid cross-entropy を用いるため，勾配消失が起きやすい．これを防ぐために，LS-GAN [5] では，目的関数を二乗誤差として定式化する．識別モデルの識別損失 $L_{\mathrm{D}}^{(\mathrm{LS\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は

$$L_{\mathrm{D}}^{(\mathrm{LS\text{-}GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{2T}\sum_{t=1}^{T}\left(D\left(\boldsymbol{y}_t\right)-b\right)^2 + \frac{1}{2T}\sum_{t=1}^{T}\left(D\left(\hat{\boldsymbol{y}}_t\right)-a\right)^2 \tag{19}$$

であり，音響モデルの敵対損失 $L_{\mathrm{ADV}}^{(\mathrm{LS\text{-}GAN})}(\hat{\boldsymbol{y}})$ は

$$L_{\mathrm{ADV}}^{(\mathrm{LS\text{-}GAN})}(\hat{\boldsymbol{y}}) = \frac{1}{2T}\sum_{t=1}^{T}\left(D\left(\hat{\boldsymbol{y}}_t\right)-c\right)^2 \tag{20}$$

として与えられる．ここで，a，b，c はそれぞれ，識別モデルに合成音声を合成音声と，自然音声を自然音声と，合成音声を自然音声として識別させるためのラベルである．これらのラベルが $b-c=1$ 及び $b-a=2$ を満たすとき，LS-GAN は $p(\boldsymbol{y})+q(\hat{\boldsymbol{y}})$ と $2q(\hat{\boldsymbol{y}})$ の間の Pearson $\mathcal{X}^2$ ダイバージェンスを最小化する．予備実験により，このダイバージェンスは合成音声の品質を劣化させることが確認されたため，本稿では [5] の式 (9) で提案されている $a=0$，$b=1$，$c=1$ を用いる．

4 実験的評価

4.1 実験条件

実験的に用いるデータとして，ATR 音素バランス 503 文 [15] 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 [16] 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 [17] 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 [18] を利用する．継続長は，自然音声のものを用いる．なお，0 次のメルケプストラム係数については，敵対的 DNN 音声合成による品質の劣化が確認されたため，MGE 学習後のものを用いる．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，音声特徴量及び実数値をとるコンテキストラベル特徴量を平均 0,

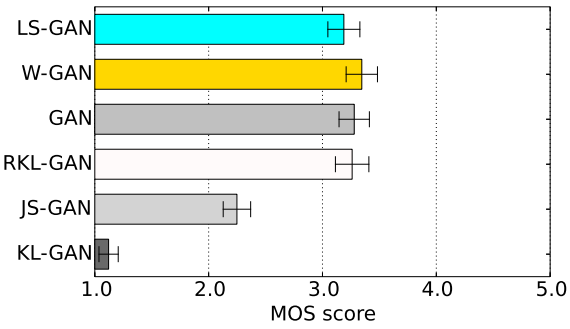

Fig. 3　音質に関する主観評価結果（エラーバーは95%信頼区間）

分散 1 に正規化する．音声合成の音響モデルと識別モデルのための 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．識別モデルの隠れ層数は 3，隠れ層の素子数は 256，隠れ層の活性化関数は ReLU である．識別モデルの出力層の活性化関数は，各 GAN により異なる．音響モデルはスペクトル特徴量（25 次元），連続対数 F_0（1 次元），非周期成分（5 次元）とそれらの動的特徴量 (Δ，ΔΔ)，U/V（1 次元）を結合させた 94 次元のベクトルをフレーム毎に予測し，識別モデルはスペクトル特徴量と連続対数 F_0 の静的特徴量を結合させた 26 次元のベクトルを用いて自然音声と合成音声をフレーム毎に識別する．最適化手法として学習率 0.001 の AdaGrad [19] を用いる．

まず，音響モデルの初期化として，反復回数 25 回の MGE 学習を行う．次に，識別モデルの初期化として，自然音声特徴量と MGE 学習後の合成音声特徴量を識別するような反復回数 5 回の学習を行う．その後，初期化された音響モデルと識別モデルを用いて，反復回数 25 回の敵対的 DNN 音声合成の学習を行う．

本稿では，以下の GAN を採用した敵対的 DNN 音声合成を比較する．ただし，ハイパーパラメータ ω_{D} は，全ての GAN で通常設定の 1.0 とする．

GAN: 式 (3)，(4)

KL-GAN: 式 (8)，(9)

RKL-GAN: 式 (11)，(12)

JS-GAN: 式 (14)，(15)

W-GAN: 式 (17)，(18)

LS-GAN: 式 (19)，(20)

主観評価として，我々のクラウドソーシングによる評価システムを用いて，55 名の被験者による 5 段階の MOS テストを実施する．評価指標は，合成音声の音質である．

4.2　評価結果

Figure 3 に評価結果を示す．評価結果より，KL-GAN と JS-GAN を除く種々の距離規範が敵対的 DNN 音声合成に有効であることが確認できる．興味深い点として，(1) KL ダイバージェンスを最小化する KL-GAN は有効でないが，その双対である RKL-GAN は有効であり，高いスコアを獲得していること，(2) JS ダイバージェンスを最小化する JS-GAN はあまり有効ではないが，その近似を最小化する GAN は有効であることが挙げられる．最も高いスコアを獲得した GAN は W-GAN であり，LS-GAN，JS-GAN，そして KL-GAN と比較して有意に高い音質であることが分かった．この要因として，特に無声音及び無音区間における音質の改善を確認した．

5　おわりに

本稿では，種々の距離規範を最小化する GAN を敵対的 DNN 音声合成に導入し，実験的評価によりその影響を調査した．評価結果より，Earth-Mover 距離最小化に基づく Wasserstein GAN が，合成音声の品質改善に最も有効であること示した．今後は，スペクトログラムの生成 [20] において有効な距離規範及び DNN アーキテクチャについて調査する．

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] Goodfellow et al., *Proc. NIPS*, pp. 2672–2680, 2014.
[2] Saito et al., *Proc. ICASSP*, pp. 4900–4904, 2017.
[3] Nowozin et al., *Proc. NIPS*, pp. 271–279, 2016.
[4] Arjovsky et al., *arXiv:1701.07875*, 2017.
[5] Mao et al., *arXiv:1611.04076*, 2017.
[6] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 7, pp. 1255–1265, 2016.
[7] Tokuda et al., *Proc. ICASSP*, pp. 1315-1318, 2000.
[8] Huang et al., *Proc. INTERSPEECH*, pp. 2464–2468, 2015.
[9] Reed et al., *Proc ICML*, pp. 1060–1069, 2016.
[10] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 4, pp. 768–783, 2016.
[11] Chen et al., *Proc. INTERSPEECH*, pp. 2097–2101, 2015.
[12] Lee et al., *Proc. NIPS*, pp. 556–562, 2000.
[13] Kompass, *Neural Computation*, Vol. 19, No. 3, pp. 780–891, 2007.
[14] Vilani, *Optimal Transport: Old and New*, Springer, 2009.
[15] 阿部 他, ATR テクニカルレポート, TR-I-0166, 1990.
[16] Kawahara et al., *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[17] Ohtani et al., *Proc. INTERSPEECH*, pp. 2266–2269, 2006.
[18] Takamichi et al., *Proc. Blizzard Challenge Workshop*, 2015.
[19] Duchi et al., *JMLR*, Vol. 12, pp. 2121–2159, 2011.
[20] Wang et al., *arXiv:1703.10135*, 2017.