ニューラルボコーダ合成音声の品質向上を狙った勾配法による後処理 *

☆高津航輝, 矢田部浩平 (農工大), 小泉悠馬 (Google Research)

1 まえがき

ニューラルボコーダはテキスト音声変換や声質変換，音声強調などの多くの音声信号処理に使用されており，高品質なニューラルボコーダの開発は音声信号処理において重要である．

近年は，ニューラルボコーダの計算量と品質のトレードオフを改善する研究が行われている．例えば Multiband MelGAN [1] は MelGAN を基に，SpecGrad [2] は WaveGrad と PriorGrad を基に改良されたニューラルボコーダである．しかし，これらの改良はその基になったモデルのみを対象としているため，その他のニューラルボコーダにそのまま適用することはできない．そのため，ニューラルボコーダの信号生成過程に依存しない品質向上法の開発が望まれる．本研究では後処理によって任意のニューラルボコーダの生成音声の品質を向上させる手法を提案する．実験では，客観評価と主観評価で提案手法の有効性を確かめた．その結果，計算資源の限られた環境において，様々なニューラルボコーダ合成音声の品質を向上させることが示された．

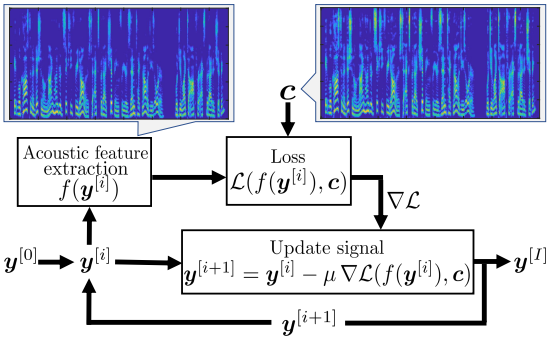
2 提案手法

ニューラルボコーダは，対数メルスペクトログラムに代表される音響特徴量 $\boldsymbol{c} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K) \in \mathbb{R}^{FK}$ から音声波形 $\boldsymbol{y} \in \mathbb{R}^D$ を生成する DNN である．ここで，F は時刻 k の特徴量ベクトルの次元を，K は時間フレームの総数を表す．本研究では，任意のニューラルボコーダと組み合わせることができる後処理手法を提案する．**図–1** に，提案手法の概要図を示す．

信号 $\boldsymbol{x}$ と信号 $\boldsymbol{y}$ の誤差を $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ とする．なお，本提案手法において誤差の算出方法に制約はない．i 回の反復で得られる音声信号を $\boldsymbol{y}^{[i]}$，入力信号から音響特徴量を得る写像を $f : \mathbb{R}^D \to \mathbb{R}^{FK}$，ニューラルボコーダで音声を合成する際に用いた音響特徴量を $\boldsymbol{c}$ とする．提案手法は誤差 $\mathcal{L}(f(\boldsymbol{y}^{[i]}), \boldsymbol{c})$ を減らすことを目標とする．誤差を減らす手法として，本研究では

$$\boldsymbol{y}^{[i+1]} = \boldsymbol{y}^{[i]} - \mu \nabla \mathcal{L}(f(\boldsymbol{y}^{[i]}), \boldsymbol{c}), \qquad (1)$$

で表される勾配法を用いる．なお，式 (1) 中の $\nabla \mathcal{L}(f(\cdot), \boldsymbol{c})$ は $\mathcal{L}(f(\cdot), \boldsymbol{c})$ の勾配，$\mu \in \mathbb{R}_{++}$ はステップサイズである．提案手法の誤差更新に必要なものはニューラルボコーダの出力 $\boldsymbol{y}^{[0]}$ のみであるため，前段に接続するニューラルボコーダの種類に制約はない．本提案手法では，損失関数 $\mathcal{L}(f(\cdot), \boldsymbol{c})$ を減少させることに主眼を置いている．従って，損失関数を減少させる最適化アルゴリズムとして式 (1) 以外の任意のものを適用することも可能である．



図–1 提案手法の概要図．入力信号の音響特徴量は入力信号 $\boldsymbol{y}^{[i]}$ から $f(\boldsymbol{y}^{[i]})$ で算出される．算出された音響特徴量は，与えられた音響特徴量 $\boldsymbol{c}$ と比較される．2 者間の誤差 $\mathcal{L}(f(\boldsymbol{y}^{[0]}), \boldsymbol{c})$ を勾配法によって減らす．

Algorithm 1 Proposed method (frame-wise)

Input: $\boldsymbol{y}^{[0]}$ (waveform generated by a neural vocoder), $\boldsymbol{c}$ (conditioning acoustic feature), μ (step size)
Output: $\boldsymbol{y}^{[I]}$ (refined waveform)

for $i = 0$ to $I - 1$ **do**
  for $k = 1$ to K **do**
    $\boldsymbol{y}_k^{[i]} = \boldsymbol{w} \odot \mathrm{GetKthSegment}(\boldsymbol{y}^{[i]})$
    $\boldsymbol{y}_k^{[i+1]} = \boldsymbol{y}_k^{[i]} - \mu \nabla \mathcal{L}(f_{\mathrm{seg}}(\boldsymbol{y}_k^{[i]}), \boldsymbol{c}_k)$
  end for
  $\boldsymbol{y}^{[i+1]} = \mathrm{OverLapAdd}(\boldsymbol{y}_1^{[i+1]}, \boldsymbol{y}_2^{[i+1]}, \ldots, \boldsymbol{y}_K^{[i+1]})$
end for

2.1 フレームごとの更新

提案手法をより効果的なものにするために，合成音声を時間フレームごとに更新する．1 フレームごとの音響特徴量抽出を $f_{\mathrm{seg}}(\cdot) : \mathbb{R}^L \to \mathbb{R}^F$ とする．信号 $\boldsymbol{y}^{[i]}$ から切り出した k 番目の時間フレームに窓をかけた信号 $\boldsymbol{y}_k^{[i]} = \boldsymbol{w} \odot \mathrm{GetKthSegment}(\boldsymbol{y}^{[i]})$ の更新式は，誤差 $\mathcal{L}(f_{\mathrm{seg}}(\boldsymbol{y}_k^{[i]}), \boldsymbol{c}_k)$ を用いて

$$\boldsymbol{y}_k^{[i+1]} = \boldsymbol{y}_k^{[i]} - \mu \nabla \mathcal{L}(f_{\mathrm{seg}}(\boldsymbol{y}_k^{[i]}), \boldsymbol{c}_k) \qquad (2)$$

のように表される．なお，この際に使用する窓 $\boldsymbol{w}$ は，オーバーラップ 1/4 のハン窓やハニング窓のような

* Post-processing method for improving sound quality of speech signals generated by neural vocoder. By Koki TAKATSU, Kohei YATABE (Tokyo University of Agriculture and Technology) and Yuma Koizumi (Google Research) .

表–1 WARP-Q の値と 95% 信頼区間．全ての信号は 16 kHz にダウンサンプリングして評価した．

Model (lr)3-5	Dataset	WARP-Q(↓) Generated	Prop (Alg. 1)	Prop (variant)
SpecGrad	Original	0.432 (±0.005)	**0.355** (±0.005)	0.418 (±0.007)
WaveGrad		0.476 (±0.006)	**0.358** (±0.005)	0.437 (±0.006)
PriorGrad		0.431 (±0.005)	**0.354** (±0.005)	0.415 (±0.005)
HiFi-GAN	LJSpeech	0.476 (±0.004)	**0.431** (±0.004)	0.472 (±0.004)
PWGAN		0.650 (±0.005)	**0.475** (±0.005)	0.582 (±0.005)
MelGAN		0.507 (±0.004)	**0.444** (±0.004)	0.491 (±0.006)
HiFi-GAN	LibriTTS	0.497 (±0.008)	**0.435** (±0.008)	0.483 (±0.008)
PWGAN		0.602 (±0.008)	**0.469** (±0.008)	0.545 (±0.005)
MelGAN		0.804 (±0.011)	**0.525** (±0.010)	0.638 (±0.009)

矩形の双対窓でなければならない．各フレームの更新ののち，$\boldsymbol{y}^{[i+1]} = \mathrm{OverLapAdd}(\boldsymbol{y}_1^{[i+1]}, \ldots, \boldsymbol{y}_K^{[i+1]})$ で表されるオーバーラップ処理を施す．時間フレームごとの更新を取り入れた提案手法の一連のアルゴリズムを Alg. 1 に示す．

3 実験

提案手法の妥当性を確かめるため，客観評価と主観評価を行った．実験では，3 つの DDPM ベースのニューラルボコーダ (WaveGrad, PriorGrad, SpecGrad) と 3 つの GAN ベースのニューラルボコーダ (HiFi-GAN, Parallel WaveGAN（PWGAN）, MelGAN) の出力に対して提案手法を適用した．

3.1 客観評価

客観評価では，ニューラルボコーダの合成音声の音質を予測する際に用いられる指標（WARP-Q）と音声品質指標（SQuId）を使用した．

WARP-Q による評価では，提案手法（Prop. (Alg.1)）と併せて，フレームごとの更新を行わない派生手法（Prop. (variant)）も評価した．WARP-Q の値を**表–1** に示す．提案手法を適用した音声 (Prop. (Alg.1)) が最良の値を示し，派生手法を適用した音声 (Prop. (variant)) はニューラルボコーダ合成音声よりも良好な値を示した．WARP-Q はテスト信号と参照信号のスペクトログラムの差に基づく指標であるため，提案手法及び派生手法は誤差 $\mathcal{L}(\cdot, \boldsymbol{c})$ を減らすことが分かった．

続いて，WARP-Q の比較でより良い値を示した提案手法 (Prop(Alg.1)) の有効性をさらに確かめるため，ニューラルボコーダ合成音声の自然さを予測する指標 SQuId で評価した．SQuId の値を**表–2** に示す．提案手法は，PriorGrad を除く全ての条件で元のニューラルボコーダ合成音声より良い値を示した．また，PWGAN と MelGAN の合成音声に対しては大幅な値の向上を確認できた．これは，人間の話し声に比べて不自然な合成音声に対して，提案手法が効果的に作用したことを示している．

表–2 SQuId の値と 95% 信頼区間．全ての信号は 16 kHz にダウンサンプリングして評価した．

Model (lr)3-4	Dataset	SQuId(↑) Generated	Prop (Alg. 1)
GroundTruth	Original	4.214 (±0.022)	
SpecGrad	(F_s = 16 kHz)	4.152 (±0.025)	**4.154** (±0.026)
WaveGrad		4.100 (±0.026)	**4.128** (±0.026)
PriorGrad		**4.117** (±0.028)	4.112 (±0.029)
GroundTruth	LJSpeech	4.078 (±0.016)	
HiFi-GAN	(F_s = 16 kHz)	3.992 (±0.017)	**3.997** (±0.017)
PWGAN		3.844 (±0.023)	**3.922** (±0.021)
MelGAN		3.776 (±0.025)	**3.878** (±0.022)
GroundTruth	LibriTTS	3.979 (±0.038)	
HiFi-GAN	(F_s = 16 kHz)	3.878 (±0.047)	**3.896** (±0.046)
PWGAN		3.722 (±0.055)	**3.813** (±0.050)
MelGAN		3.530 (±0.060)	**3.701** (±0.055)

表–3 提案手法を適用する前後の MOS の値．表中の F_s はサンプリング周波数．

Model (lr)3-4	Dataset	MOS(↑) Generated	Prop (Alg. 1)
GroundTruth	Original	4.48 (±0.15)	
SpecGrad	(F_s = 24 kHz)	**4.32** (±0.16)	4.25 (±0.19)
WaveGrad		3.97 (±0.22)	**4.03** (±021)
PriorGrad		3.81 (±0.21)	**4.01** (±0.23)
GroundTruth	Original	4.35 (±0.15)	
SpecGrad	(F_s = 16 kHz)	**4.22** (±0.19)	3.96 (±0.22)
WaveGrad		3.95 (±0.21)	**4.02** (±0.20)
PriorGrad		3.81 (±0.23)	**3.96** (±0.23)

3.2 主観評価

SQuId のスコアに明確な差が見られなかったため，追加の品質評価として DDPM ベースモデルを対象に主観評価を行った．主観評価では，Mean Opinion Score（MOS）を使用した．MOS の結果を**表–3** に示す．結果は GAN ベースモデルの SQuId の値と同じ傾向で，提案手法が MOS の低いモデルの出力信号の自然度を改善したことを示した．以上より，提案手法は限られた計算機資源の条件下で様々なニューラルボコーダの品質を向上させるのに適していると言える．

4 まとめ

本研究ではニューラルボコーダの生成音声の品質を向上させる後処理手法を提案した．実験では，計算量が少なく品質の低いニューラルボコーダに対しての妥当性が確認された．今後は，高品質なニューラルボコーダに対しても有効な後処理手法の検討を行う．

参考文献

[1] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in Proc. SLT, 2021, pp. 492–498.

[2] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," in Proc. Interspeech, 2022, pp. 803-807.