ホワイトボックス攻撃による Neural Audio Codec の敵対的サンプル生成*

☆田中智己, 吉野夏樹, 矢田部浩平 (農工大), 中村友彦 (産総研)

1　はじめに

深層学習を用いた音響信号の圧縮手法である Neural Audio Codec（NAC）は，極めて低いビットレートで高品質な信号の再構成を可能とする．さらに NAC は入力信号の離散的な表現（離散トークン）を生成できることから，様々なタスクにおける基盤モデルとしての役割が期待されている．しかしながら，基盤モデルは様々な入力を安定して処理することが求められる一方，深層学習モデルは入力に悪意のある微小な変化（敵対的摂動）を加えた敵対的攻撃に脆弱であり，攻撃によって意図しない出力をする可能性がある．このような攻撃はシステムの信頼性低下を招く恐れがあるため，敵対的サンプルの性質からモデルの脆弱性を特定する試みが行われている [1]．本稿では，NAC による信号再構成と NAC の離散トークンを用いた自動音声認識（ASR）タスクを対象とし，敵対的サンプルに対する NAC の頑健性の調査を行う．具体的には，モデル構造が既知であるホワイトボックス設定で，タスクに応じた敵対的サンプルを射影勾配型アルゴリズム [2] により生成し，これを用いた攻撃がタスク結果に及ぼす影響を実験的に評価する．

2　射影勾配型アルゴリズムによる攻撃

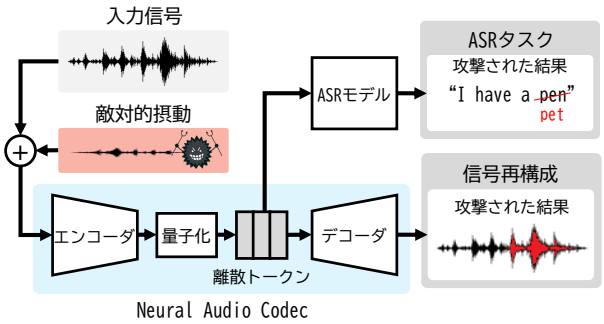
元信号 $\boldsymbol{x}$ による出力と，敵対的摂動 $\boldsymbol{\delta}$ を加えた敵対的サンプル $\widetilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}$ による出力について，攻撃はこの 2 出力間の誤差を大きくすることで達成される．ここで誤差を表す目的関数を $\mathcal{L}(\widetilde{\boldsymbol{x}}, \boldsymbol{x})$ とする．射影勾配型のアルゴリズムでは，ステップ t における敵対的摂動を $\boldsymbol{\delta}^{(t)}$ とすると，更新式は

$$\boldsymbol{\delta}^{(t+1)} = P_{\mathcal{C}}\left(\boldsymbol{\delta}^{(t)} + \alpha \operatorname{sgn}\left(\nabla_{\boldsymbol{\delta}^{(t)}} \mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}^{(t)}, \boldsymbol{x})\right)\right)$$

となる [2]．ここで，α はステップサイズ，$P_{\mathcal{C}}$ は $\boldsymbol{\delta}$ の制約集合 $\mathcal{C}$ への射影である．本稿では，$\mathcal{C}$ を敵対的サンプルの振幅の不等式制約とする．

3　NAC の下流タスクと攻撃の目的関数

図–1 に本稿で扱う NAC の構造及び攻撃の概要を示す．NAC は入力信号を圧縮するエンコーダ $E(\cdot)$ と，その圧縮結果を有限個のパターンの組み合わせに量子化し離散トークンへ変換する量子化器 $Q(\cdot)$，圧縮結果から信号を再構成するデコーダ $D(\cdot)$ から構成される．これらを入力信号と再構成信号の誤差を下げるよう学習することで，ビットレートを抑えつつ高品質な信号の再構成が可能となる．加えて，量子化器から得るトークン列は他の下流タスクを行うモデルの入力に用いることもできる．このように NAC は様々な下流タスクへ利用できるが，一方で，入力に敵対的摂動が与えられることでトークン列が変化すると，トークン列を用いる下流タスクの性能にも影響が及ぶ可能性がある．そのため NAC の攻撃に関しては，NAC を信号再構成に用いる場合と，トークン列を下流タスクの入力に用いる場合を検証する必要がある．



図–1　本稿で扱う NAC の構造及び攻撃の概要を示した図．下流タスクの結果が敵対的サンプルの入力によって変化した様子を赤色で示す．信号再構成では再構成の結果が劣化したことを，ASR タスクでは離散トークンの変化を通じた文字起こし結果が悪化したことを表している．

これを踏まえ，本稿では NAC に対する 2 種類の攻撃を検証する．具体的には，「NAC による信号の再構成」及び「入力信号の離散トークンを用いた ASR タスク」に対する攻撃を考え，NAC のモデル構造やパラメータが既知の状態で射影勾配型アルゴリズムによる敵対的サンプルの生成を行う．

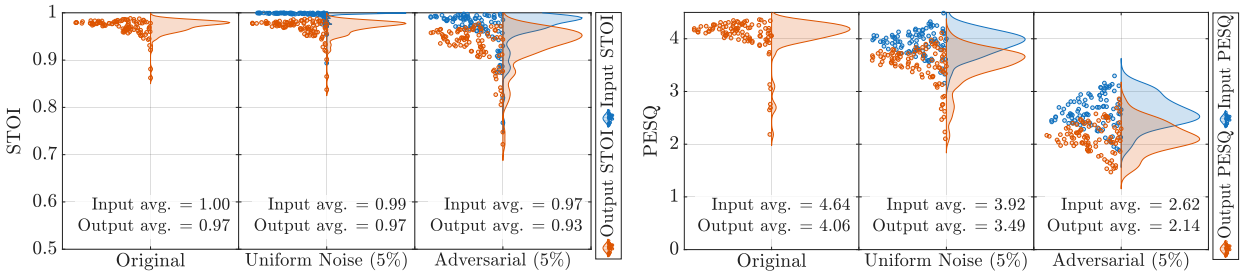
3.1　信号再構成における攻撃の目的関数

NAC による信号の再構成については，敵対的サンプル $\widetilde{\boldsymbol{x}}$ を NAC へ入力したときの再構成信号 $\widetilde{\boldsymbol{y}} = D(Q(E(\widetilde{\boldsymbol{x}})))$ が，元信号 $\boldsymbol{x}$ から離れるよう敵対的摂動を設計する．本稿では，時間周波数構造の差を，複数の設定の短時間フーリエ変換（STFT）により評価する多重解像度 STFT 損失に基づき，目的関数を

$$\mathcal{L}_{\mathrm{rec}}(\widetilde{\boldsymbol{x}}, \boldsymbol{x}) = \sum\nolimits_{r \in \mathcal{R}} \left\| |\mathrm{STFT}_r(\boldsymbol{x})| - |\mathrm{STFT}_r(\widetilde{\boldsymbol{y}})| \right\|_1$$

とし，これを反復的に増加させて敵対的摂動を生成する．ここで，$\mathcal{R}$ は STFT の設定，$\|\cdot\|_1$ は ℓ_1 ノルム，$|\cdot|$ は要素ごとの絶対値，$\mathrm{STFT}_r(\cdot)$ は設定 r におけ

*Generating adversarial samples for neural audio codec by white-box attack. By Tomoki TANAKA, Natsuki YOSHINO, Kohei YATABE (Tokyo University of Agriculture and Technology (TUAT)) and Tomohiko NAKAMURA (National Institute of Advanced Industrial Science and Technology (AIST)).



図–2　入力信号（青）と信号再構成に対する攻撃を行った際の再構成信号（赤）の STOI（左）及び PESQ（右）の分布．STOI・PESQ の両図で，左から元信号，元信号に対し RMS 包絡線の 5 %の一様ノイズを加えた信号，元信号に対し RMS 包絡線の 5 %の敵対的摂動を加えた信号を入力した場合である．各図の下に入出力の平均を示す．

る STFT を表す．ただし本稿では，互いに異なる窓幅を持つガウス窓での STFT 設定を $\mathcal{R}$ に含める．

3.2　ASR タスクにおける攻撃の目的関数

ASR タスクでは，NAC のエンコーダから得る離散トークンを ASR モデルの入力に用いる場合を考える．元信号 $\boldsymbol{x}$ から得るトークン列を $\boldsymbol{\tau} = E(\boldsymbol{x})$，敵対的サンプル $\widetilde{\boldsymbol{x}}$ から得るトークン列を $\widetilde{\boldsymbol{\tau}} = E(\widetilde{\boldsymbol{x}})$ とすると，$\boldsymbol{\tau}$ と $\widetilde{\boldsymbol{\tau}}$ が異なるほど，ASR モデルに入力される情報が変化し，ASR 結果が悪化すると考えられる．そこで本稿では，トークン表現の cos 類似度の絶対値に基づいて，目的関数を

$$\mathcal{L}_{\mathrm{ASR}}(\widetilde{\boldsymbol{x}}, \boldsymbol{x}) = -\left| \frac{\langle \boldsymbol{\tau}, \widetilde{\boldsymbol{\tau}} \rangle}{\|\boldsymbol{\tau}\|_2 \|\widetilde{\boldsymbol{\tau}}\|_2} \right|$$

とし，これを反復的に増加させて敵対的摂動を生成する．ここで，$\|\cdot\|_2$ は ℓ_2 ノルム，$\langle \cdot, \cdot \rangle$ は内積を表す．

4　実験

信号再構成に対する攻撃, 及び ASR タスクに対する攻撃について実験を行った．両実験とも，微分不可である量子化器の勾配は Straight-Through Estimator で置き換え，敵対的摂動に対する出力の勾配を計算した．敵対的摂動は振幅を元信号の二乗平均平方根（RMS）包絡線の 5 %となるように制約し，10 000 反復の更新により生成した．ただし $\alpha = 5.0 \times 10^{-3}$ とした．比較対象として，元信号に対して RMS 包絡線の 5 %の一様ノイズを加えた信号を用いた．

4.1　信号再構成に対する攻撃

まず，信号再構成に対する攻撃について実験を行った．NAC には標本化周波数 44.1 kHz で学習された Descript Audio Codec [3]（DAC）を用いた．敵対的サンプルは VCTK Corpus [4] の mic2 録音から 100 件を用いて生成し，攻撃の評価には再構成信号の STOI と PESQ を用いた．結果を図–2 に示す．敵対的サンプルは，一様ノイズを加えた信号よりも再構成信号の STOI 及び PESQ が低下した．従って，本稿で生成した敵対的サンプルは NAC の信号再構成に対し悪影響を及ぼすことが確認された．

表–1　ASR タスクにおける各信号の WER 及び CER

評価指標	元信号	一様ノイズ加算	敵対的サンプル
WER	41.4 %	54.1 %	70.7 %
CER	22.6 %	32.3 %	46.3 %

4.2　ASR タスクに対する攻撃

次に，ASR タスクに対する攻撃について実験を行った．NAC には標本化周波数 24 kHz で学習された DAC を用いた．ASR モデルは DAC から得るトークン列を入力とし，Conformer と CTC Loss を用い Whisper-tiny の語彙に基づき文字列を予測する構成とした．モデルのパラメータ数は約 50 M であった．学習には LibriTTS-R [5] の train-clean-360 を用いた．敵対的サンプルは LibriTTS-R の test から 100 件を用いて生成し，攻撃の評価には，単語誤り率（WER）と文字誤り率（CER）を用いた．結果を表–1 に示す．一様ノイズを加えた信号に比べ，敵対的サンプルでは WER が約 16 ポイント，CER が約 14 ポイント増加した．これにより，本稿で生成した敵対的サンプルは ASR モデルの性能を低下させることが示唆された．

5　むすび

本稿では，NAC に対するホワイトボックス攻撃により敵対的サンプルを生成する方法を検討した．今後は NAC の各構成要素について敵対的摂動に対する脆弱性を分析し，頑健な NAC の設計指針を検討する．

謝辞 本研究は科研費 23K28108 の助成を受けた．

参考文献

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *ICLR*, (2018).
[2] A. Kurakin, I. Goodfellow and S. Bengio, “Adversarial examples in the physical world,” *ICLR*, (2017).
[3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” *NeurIPS*, pp.27980-–27993 (2023).
[4] J. Yamagishi, C. Veaux and K. MacDonald, “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” *University of Edinburgh. The Centre for Speech Technology Research*, (2019).
[5] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han and A. Bapna, “LibriTTS-R: A restored multi-speaker text-to-speech corpus,” *Interspeech*, pp.5496–5500 (2023).