深層学習による非可逆圧縮を施した音声の復元

松永 隆太郎 †　　中谷 俊介 ‡　　松崎 拓也 §

1　はじめに

低解像度の画像のみを入力として, より高解像度の画像を得る, いわゆる超解像 (Super-resolution) の手法は深層学習の応用によって大きく進展した. これを音声に応用し, 低サンプリングレートの音声データを入力として深層学習によって補間を行うことで, 例えば元の 2 倍のサンプリングレートの音声を得るという, 音声に対する超解像の技術が近年いくつか提案されている ([1] など). これを音声データ圧縮の技術と見た場合, 例えば n 倍の超解像を用いることで転送すべき音声データの量は $1/n$ になる. しかし, 低サンプリングレートの波形データにもなお冗長性があり, 音質を一部犠牲にして非可逆圧縮を行うことでより大幅にデータ量を削減できる.

そこで本研究では音声超解像の技術を応用し, 非可逆圧縮を施した音声を元音声へと復元することを試みた. 具体的には, MP3 および CELP によって非可逆圧縮された音声を入力とし, CNN をベースとする GAN によって圧縮前の音声品質に近づける. MP3 は人間の聴覚特性を利用した圧縮方式で, インターネット上での音声・音楽の配信に広く用いられている. 一方, CELP は音声用の符号化方式の一つで, 線形予測に基づく音声合成結果が入力に最も近くなるよう選んだ合成用パラメータを送信することでデータ圧縮を行う方式であり, VoIP などの用途に広く用いられている.

2　モデルの説明

2.1　ネットワークの構造

本研究で用いたモデルは音声データの超解像を GAN により実現した SRGAN [1] を基にしている. まず非可逆圧縮された音声に対して短時間フーリエ変換を行い振幅スペクトルと位相スペクトルを得る. ここでフレームサイズ 256, フレームシフト 64 とし, 窓関数としてハン窓を用いた. 次に, 対数振幅スペクトルを 32 ～ 128 フレームずつネットワークに入力し, 圧縮前の音声の対数振幅スペクトルを正解として GAN によって学習させる. 実行時には, ネットワークから得られた振幅スペクトルと, 短時間フーリエ変換を行ったときに得られた位相スペクトルを用いて逆短時間フーリエ変換により音声を復元する.

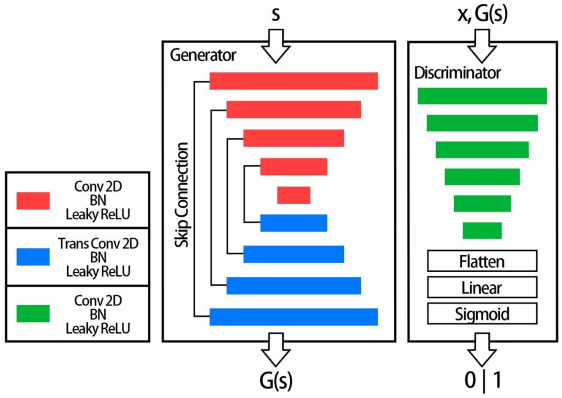

図 1：モデル全体の概要

ネットワークは Generator と Discriminator から成る. Generator は 4 層の畳み込み層と 4 層の脱畳み込み層で構成されており, それぞれに skip connection を適用している. Discriminator は 5 層の畳み込み層で構成されている. 活性化関数としては Discriminator の出力層にはシグモイド関数, それ以外の層には Leaky ReLU 関数を用いている. Generator にはすべての層に Batch Normalization (BN) を適用している. モデル全体の概要を図 1 に示す.

2.2　損失関数

学習の安定のため, 最初に Generator だけを平均二乗誤差 $\mathcal{L}_{MSE}$ を損失関数として, 数エポック学習させる. 以降は Discriminator は下記の $\mathcal{L}_{Dis}$, Generator は $\mathcal{L}_{Gen}$ を損失関数として学習する.

$$\mathcal{L}_{Dis} = \mathbb{E}_{x\sim\mathbb{P}}[\log D(x)] + \mathbb{E}_{s\sim\mathbb{Q}}[\log(1 - D(G(s)))]$$

$$\mathcal{L}_{Gen} = \alpha\mathbb{E}_{s\sim\mathbb{Q}}[-\log(D(G(s)))] + \beta\mathcal{L}_{MSE} + \gamma\mathcal{L}_{FM}$$

ここで, $D(\cdot)$ は Discriminator, $G(\cdot)$ は Generator を表し, $\mathbb{P}, \mathbb{Q}$ はそれぞれ正解データの分布, 入力データの分布を, α, β, γ は重み付け係数を表している.

$\mathcal{L}_{FM}$ は以下で定義される Feature Matching Loss である.

$$\mathcal{L}_{FM} = \sum_{i=1}^{T} \frac{1}{N_i} \|D^{(i)}(x) - D^{(i)}(G(s))\|_1$$

ここで $D^{(i)}$ は Discriminator の i 層目の状態, N_i は $D^{(i)}$ のユニット数を表す.

Recovering Lossy Compressed Speech Data by Deep Learning

† Matsunaga Ryutaro, Tokyo University of Science
‡ Nakaya Shunsuke, Tokyo University of Science
§ Matsuzaki Takuya, Tokyo University of Science

3 実験設定

実験には 109 人の話者による合計約 44 時間の音声からなる VCTK Corpus [2] を用いた. 同コーパスの音声は 48kHz, 16 ビットのリニア PCM として収録されている. 以下, この元音声を例えば 24kHz にダウンサンプリングしたものを「元音声 (24kHz)」と呼ぶ.

MP3 による圧縮には ffmpeg を利用し, ビットレートは 48k, 96k, 128k, 196kbps, サンプリングレートは全て 48kHz として圧縮した. 圧縮の際, 容量を抑えるために 48kbps では約 11kHz 以上, 96k, 128k, 196kbps では約 21kHz 以上の高周波成分がカットされる. MP3 を入力として復元する際の正解データは元音声 (48kHz) とする.

CELP による圧縮結果はまずローパスフィルタの適用後に, 元音声 (8kHz) を ITU-T V.729A [3] の参照実装に入力して 8kbps の圧縮結果を得た後, 再び同参照実装を用いて 8kHz の波形データに戻すことで得た. 正解データは元音声 (16kHz) とする. 復元する際は, まず CELP による圧縮結果を, 元音声 (8kHz) を正解とする GAN により復元し, さらにその結果を Cubic 補間により 16kHz にアップサンプリングした後, 元音声 (16kHz) を正解とする GAN により復元した.

4 実験結果

以下では, 復元した音声に対する主観的評価と客観的評価の結果を述べる. 主観的評価の実験では 2 つの音声 A, B を聞き, 音声が良い方を選ぶ AB テストを行った. 16 人の被験者がそれぞれ 20 問ずつ解答した. 実験に使用したデータは 48kbps で圧縮した MP3, それをモデルによって 48kHz の音声データとして復元したもの, および元音声 (48kHz および 24kHz) の 4 種類である. 表 1 に AB テストの結果を示す. 48kbps MP3 と元音声 (48kHz) の比較では元音声 (48kHz) が良いと解答した割合が 72.5% とある程度の差があったが, GAN により復元した音声と元音声 (48kHz) では元音声 (48kHz) が良いと解答した割合が 55.0% と差が小さいことから, GAN による復元により音質が向上したことがわかる.

客観的評価としては, 復元した音声と正解の差を以下で定義する LSD によって評価した.

$$\mathrm{LSD}(X, \hat{X}) = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(X(l,k) - \hat{X}(l,k)\right)^2}$$

X は圧縮前の音声から得た対数パワースペクトログラム, $\hat{X}$ は比較したい音声から得た対数パワースペクトログラムを表し, L は X, $\hat{X}$ の時間方向のサイズ, K は周波数方向のサイズを表す.

表 2 に, MP3 による圧縮結果およびモデルにより復元した音声と, 元音声 (48kHz) の間の LSD を示す.

表1：MP3からの復元結果に関するABテストの結果

A vs B	Aが良いと解答した割合
元音声(48kHz) vs 48kbps MP3	72.5%
元音声(48kHz) vs GAN(48kbps MP3)	55.0%
48kbps MP3 vs GAN(48kbps MP3)	40.0%
元音声(24kHz) vs GAN(48kbps MP3)	38.8%

表2：元音声(48kHz)に対するLSD

	LSD	LSD(LF)
48kbps MP3	5.35	2.09
96kbps MP3	3.19	1.84
128kbps MP3	2.92	1.17
196kbps MP3	2.73	0.65
GAN(48kbps MP3)	1.71	1.51
GAN(96kbps MP3)	1.27	1.11
GAN(128kbps MP3)	1.13	0.90
GAN(196kbps MP3)	0.93	0.58

表3：元音声(8kHz)に対するLSD

	LSD
8kHz CELP	2.44
8kHz CELP → GAN(8kHz)	2.38

表4：元音声(16kHz)に対するLSD

	LSD
8kHz CELP → Cubic補間	3.61
8kHz CELP → GAN(8kHz) → Cubic補間	3.48
8kHz CELP → GAN(8kHz) → Cubic補間 → GAN(16kHz)	3.10
元音声(8kHz) → Cubic補間 → GAN(16kHz)	2.50

LSD(LF) は MP3 の圧縮時にカットされた高周波成分を除いた低周波成分に対する LSD である. 表 2 からどのビットレートでもモデルによる復元で LSD が改善されたことがわかる. 音楽データなどは, 一般的に 128kbps で圧縮されることが多いため, 96kbps で圧縮した MP3 をモデルで復元した結果が 128kbps で圧縮した MP3 よりも LSD, LSD(LF) の両方の点で良くなるという結果から, 今後音楽データの圧縮にも役立つことが期待できる.

表 3, 表 4 に, CELP による圧縮結果およびモデルにより復元した音声と, 元音声 (8kHz および 16kHz) の間の LSD を示す. CELP で圧縮したデータから直接 16kHz の音声に復元するよりも, まず 8kHz の音声を復元し, その後 16kHz への超解像を行う方が良いことがわかる.

5 おわりに

本研究では, 非可逆圧縮を施した音声の復元を GAN を用いて行った. その結果, 主観的評価および客観的評価のいずれにおいても音質の向上が確認できた. 今後は音楽データに対して学習し, より効率の良い圧縮を実現することが課題である.

参考文献

[1] Sefik Emre Eskimez, Kazuhito Koishida, Adversarial Training for Speech Super-Resolution, IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 347-358, May 2019.

[2] VCTK Corpus, http://www.udialogue.org/ja/download-ja/cstr-vctk-corpus.html

[3] ITU-T recommendation V.729A, https://www.itu.int/rec/T-REC-G.729-199611-S!AnnA