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

目的

- 非可逆圧縮した音声の音質をなるべく元音声に近づける.
  → 圧縮技術に応用することでデータ量を大幅に削減する.

例

圧縮した音声(CELP 8kHz)

復元した音声(16kHz)

概要

- 超解像の手法を応用し
  - MP3 圧縮音声を復元
  - CELP 圧縮音声を復元

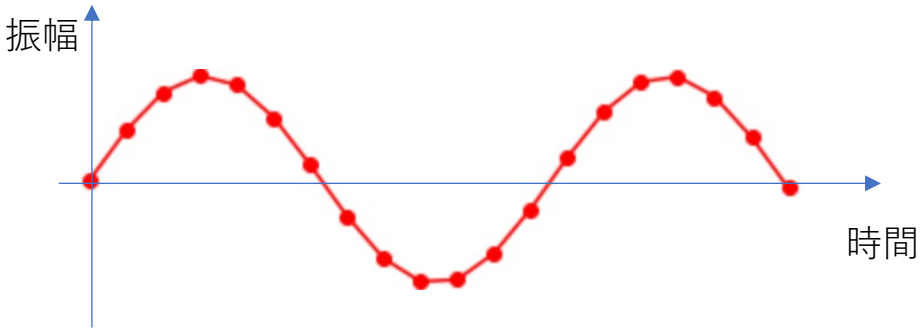
背景：超解像とは

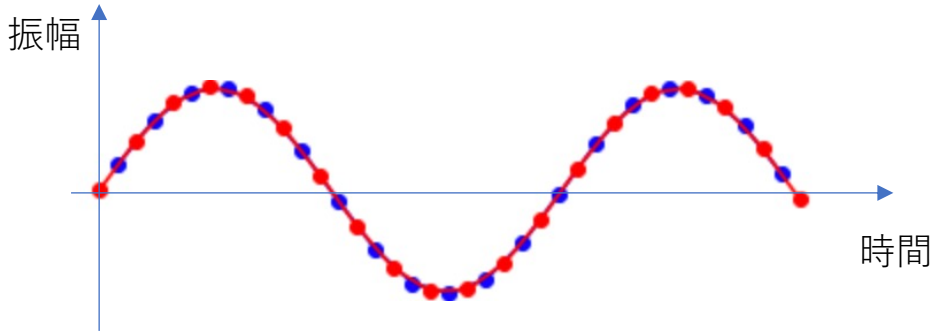
- 低解像度の画像や音声を入力として, 深層学習などによって補間を行うことにより高解像度の画像, 音声を得ること.

画像

https://sorabatake.jp/11913/ より引用

音声

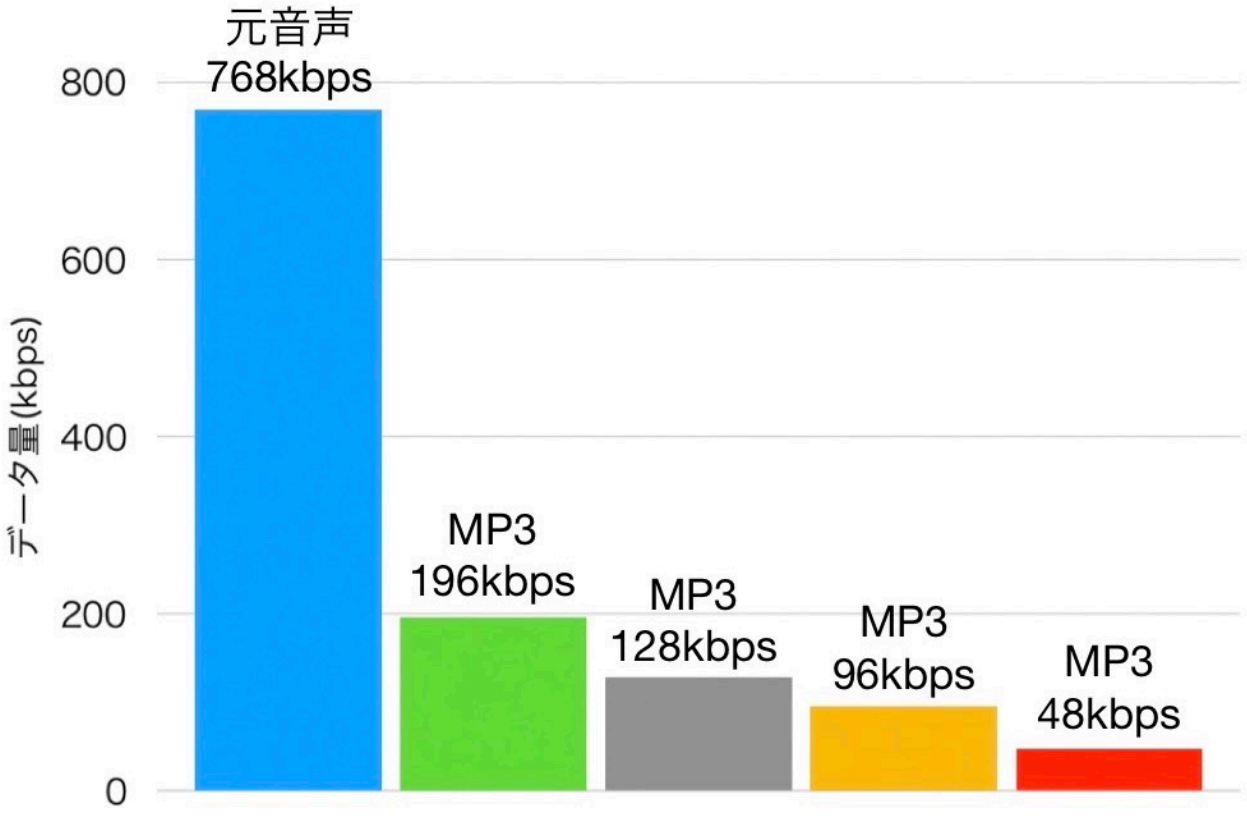




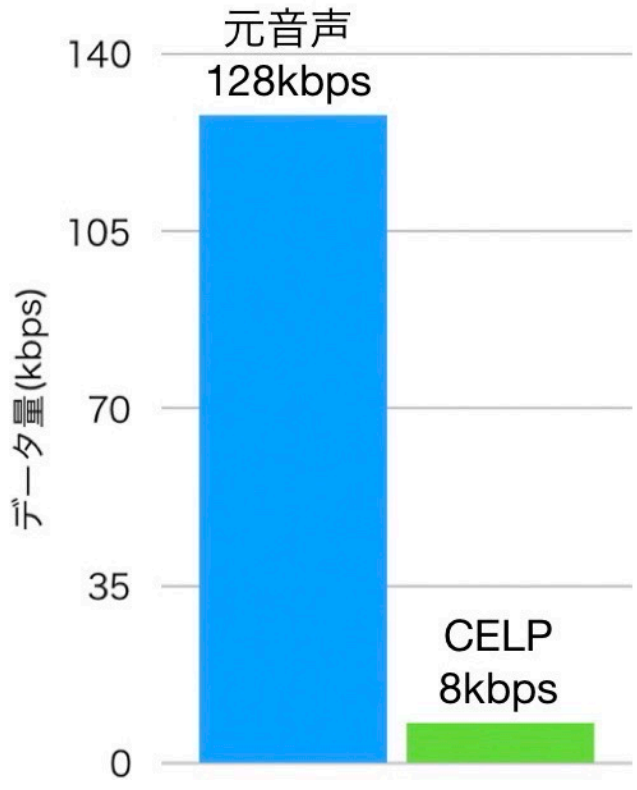
- 超解像をデータ圧縮の技術と見た場合, n 倍の超解像により転送すべき音声データの量は $1/n$, 画像データの量は $1/n^2$ になる.

背景：非可逆圧縮

- 音質を一部犠牲にして大幅にデータ量を削減できる.



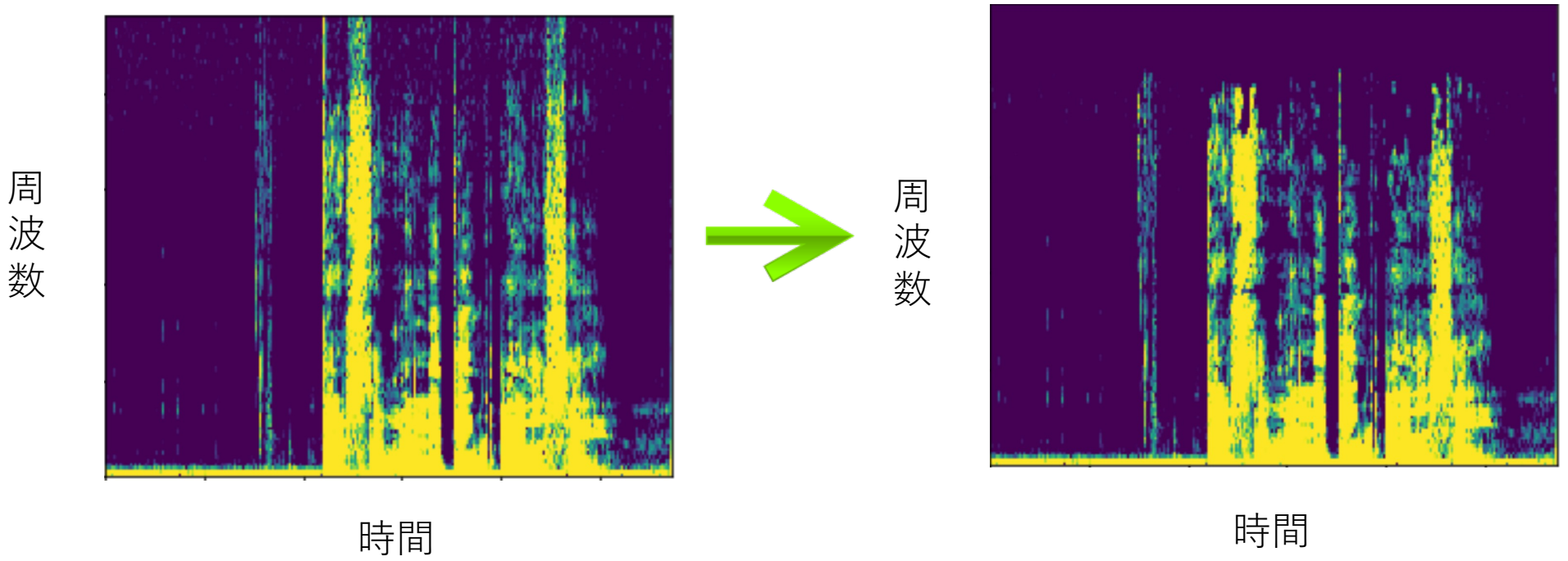
サンプリングレート：48kHz



サンプリングレート：8kHz

背景：MP3

- 人間の聴覚特性を利用し, 聞こえにくい部分の音を
  切り捨てることで, 容量を小さくする圧縮方式.

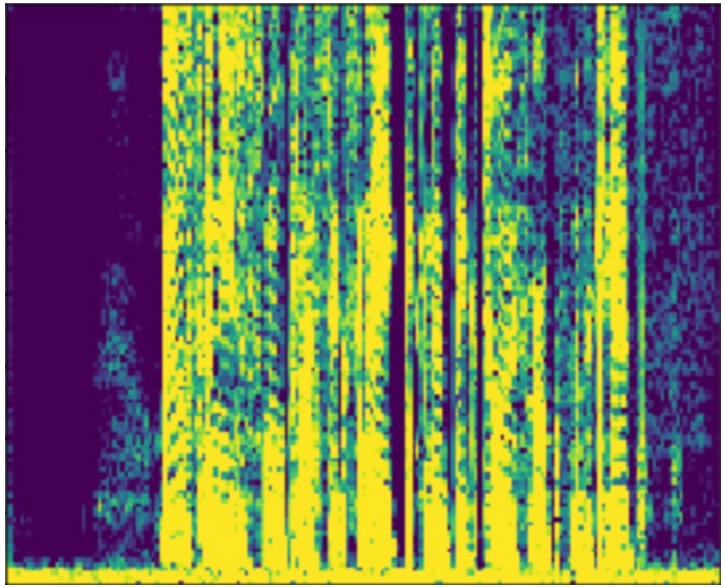


背景：CELP

- 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のみを送信する圧縮方式.

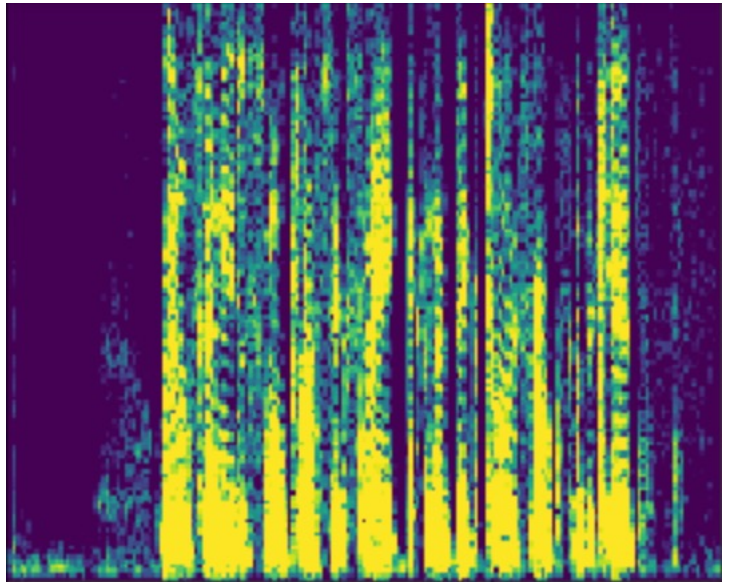
- VoIPなどの用途に広く用いられている.

周波数



時間

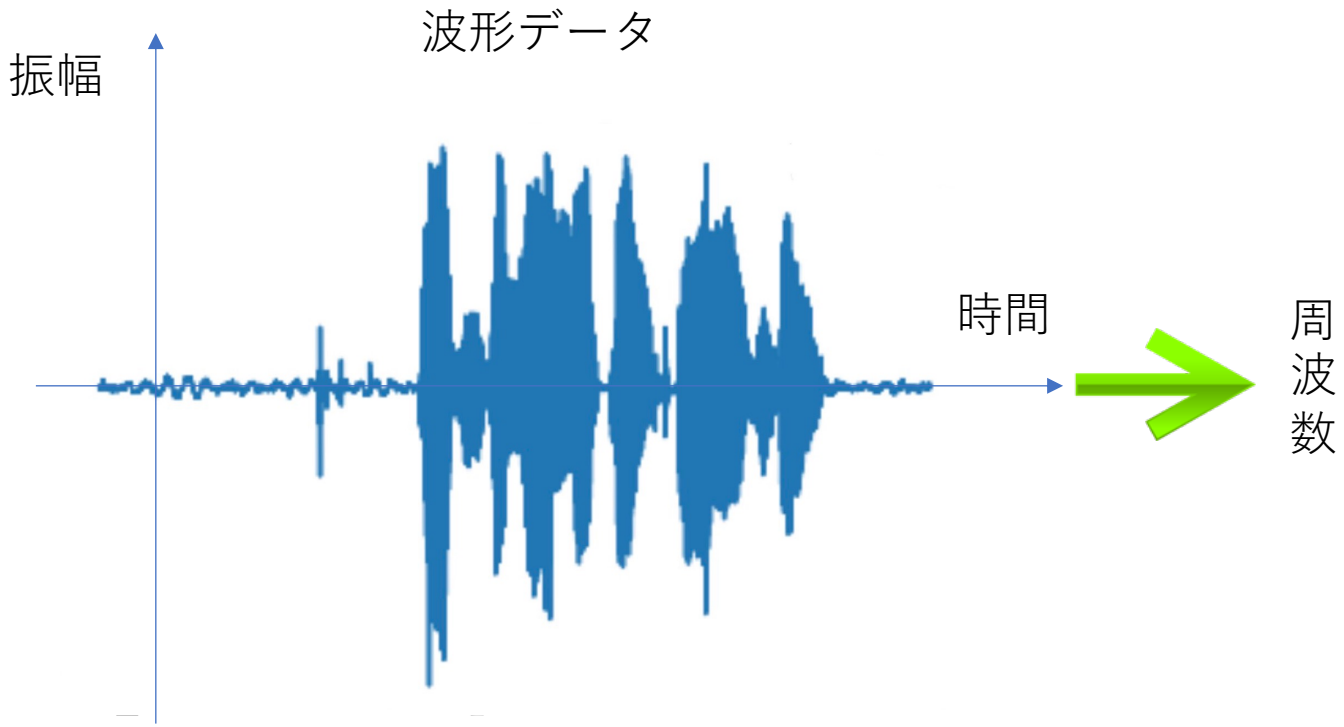
周波数

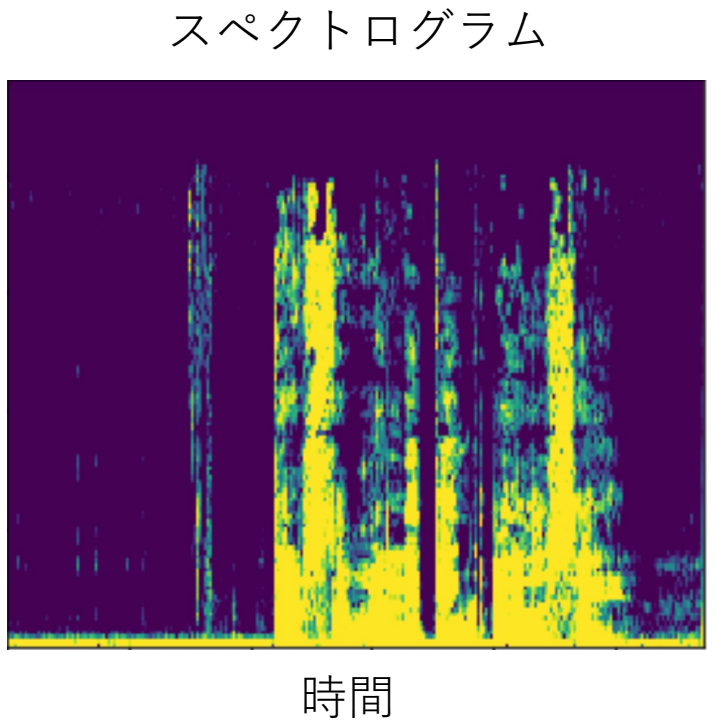


時間

準備：短時間フーリエ変換(STFT)

- 音声波形を周波数ごとの信号(スペクトル)に分解する
  フーリエ変換を短時間ごとに音声を区切り行う.
- 得られたものをスペクトログラムという.

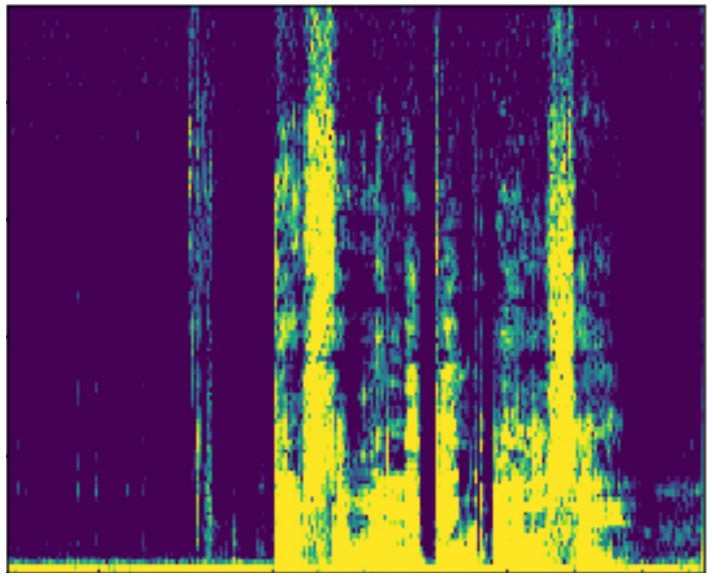




準備：逆短時間フーリエ変換(ISTFT)

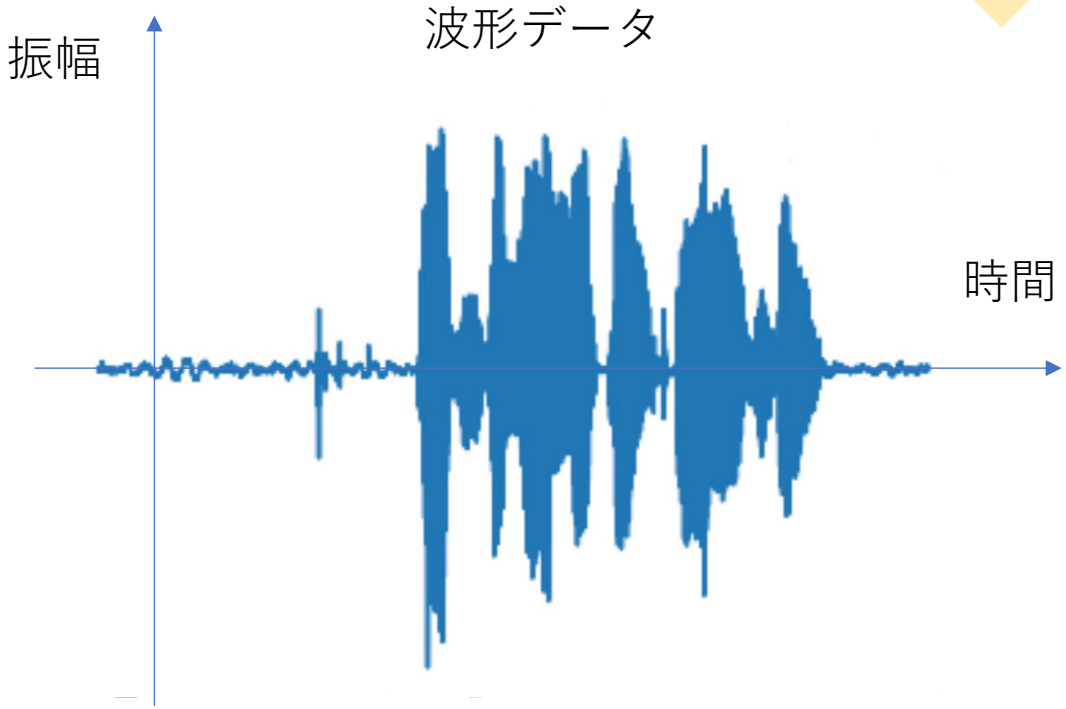
- スペクトルを音声波形に変換する逆フーリエ変換を短時間ごとに音声を区切り行う.

スペクトログラム

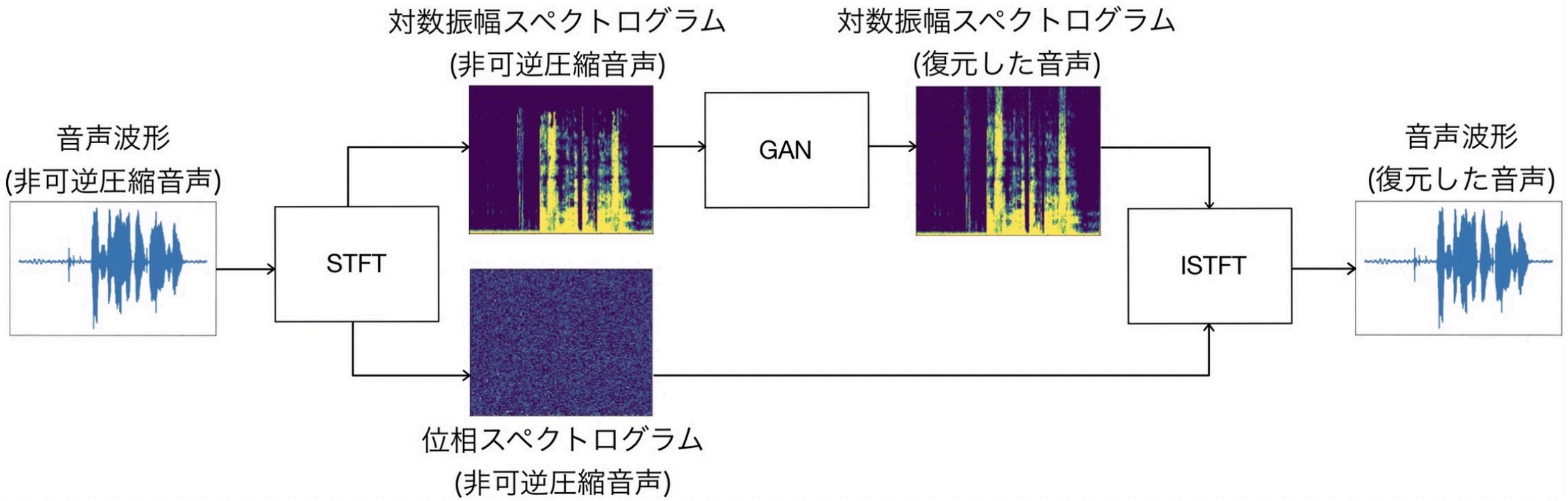


周波数

時間



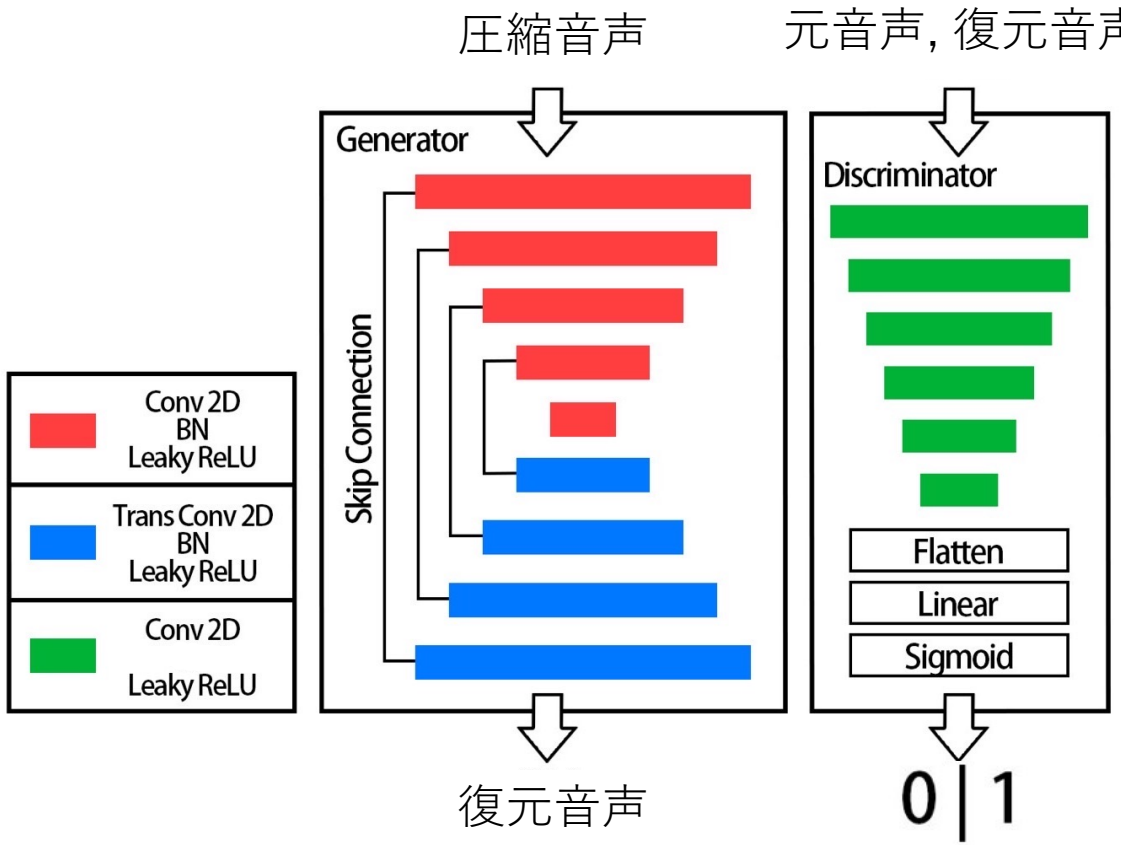
復元の流れ ([1]を参考にした)



[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.

GAN（敵対的生成ネットワーク）

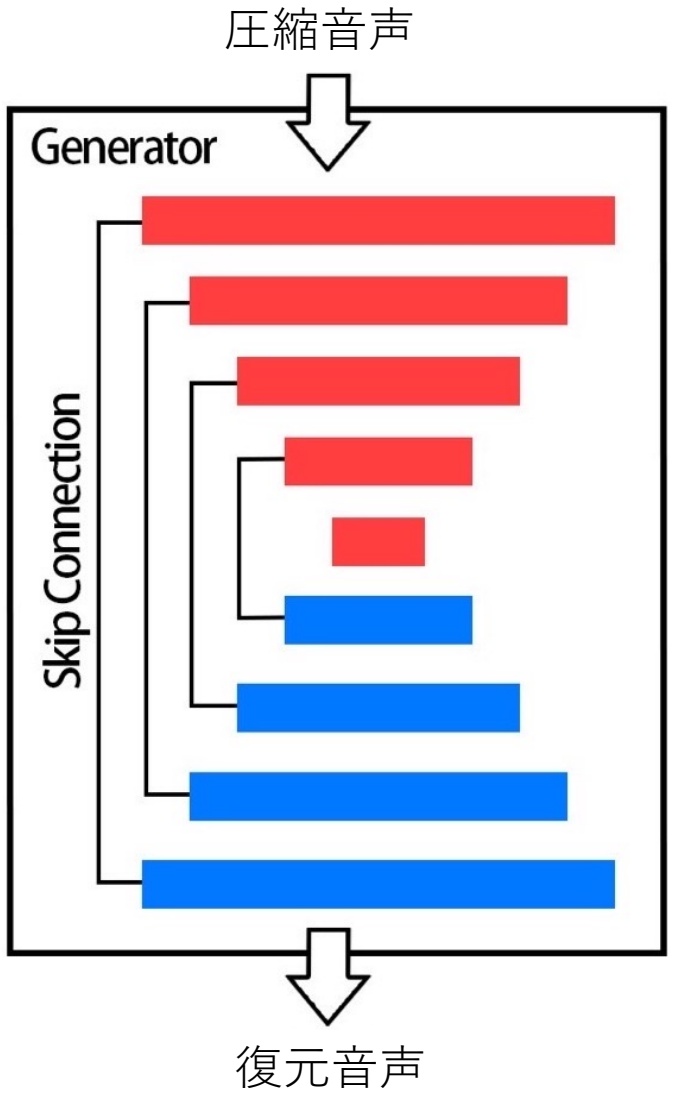
- GeneratorとDiscriminatorの2つのネットワークを互いに競い合わせることで, 復元音声の精度を高める.



Generator :
圧縮音声を入力とし, 復元音声を出力する.

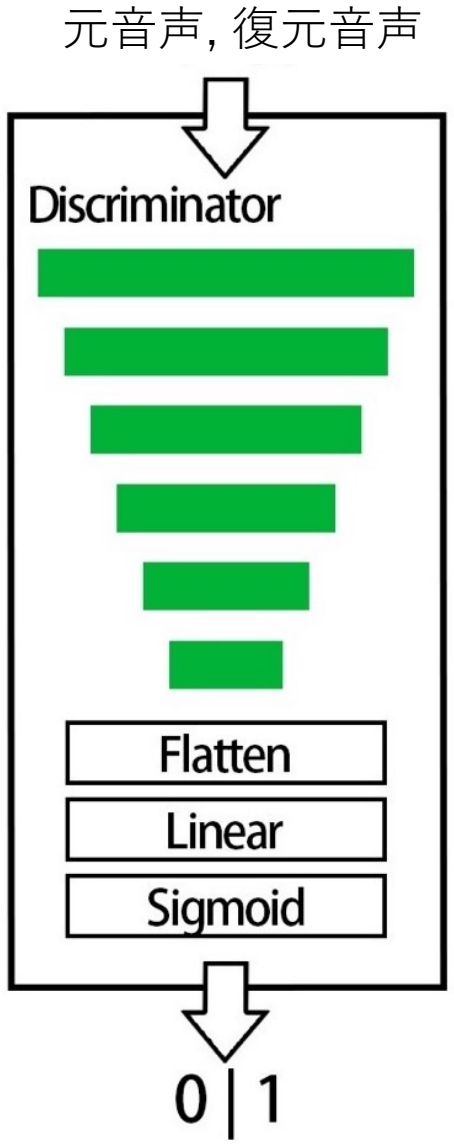
Discriminator :
元音声(本物)または復元音声(偽物)を入力し, 入力された音声が本物か偽物かを判別する.

Generator



- 4 層の畳み込み層と4層の脱畳み込み層で構成.
- それぞれに Skip Connection を適用.
- すべての層にBatch Normalization (BN)を適用.
- 活性化関数としてはすべての層に
  Leaky ReLU 関数を使用.

Discriminator



- 5層の畳み込み層で構成.
- 活性化関数としては出力層にはシグモイド関数,
  それ以外の層にはLeaky ReLU 関数を使用.

損失関数（1）

- 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.

- Discriminatorは下記の $\mathcal{L}_{Dis}$, Generatorは $\mathcal{L}_{Gen}$ を損失関数とする.

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

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

$D(\cdot)$: Discriminator
$G(\cdot)$: Generator

$\mathbb{P}$: 正解データの分布
$\mathbb{Q}$: 入力データの分布

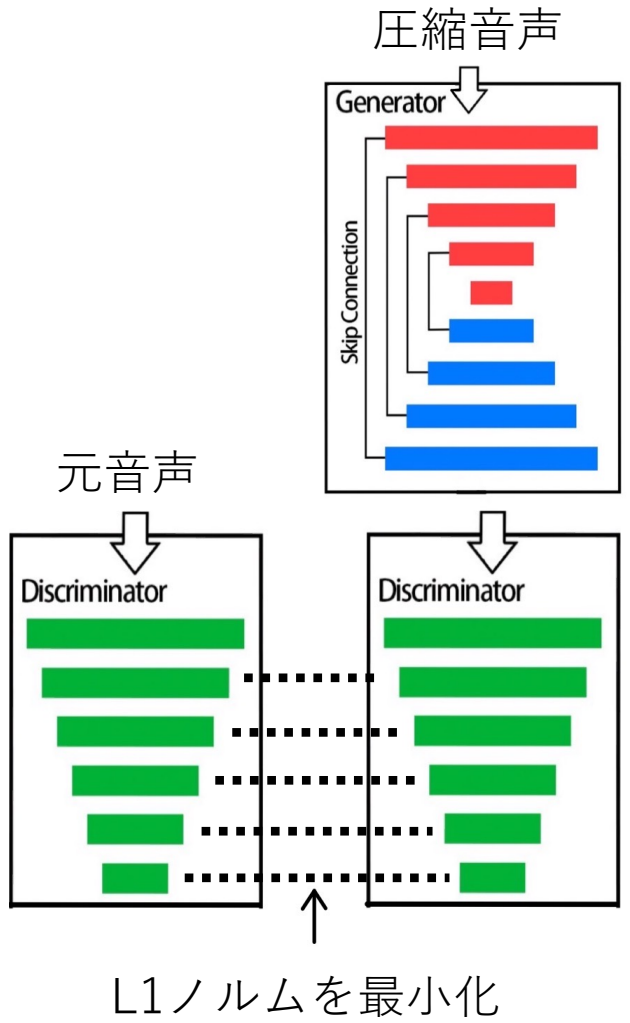
α, β, γ : 重み付け係数
$\mathcal{L}_{FM}$: 次のページで説明

損失関数（2）

- $\mathcal{L}_{FM}$は以下で定義されるFeature Matching Loss [2]：

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

$D^{(i)}$：Discriminatorのi層目の状態
N_i：$D^{(i)}$のユニット数



[2] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300, 2015.

実験設定

- 使用データ：VCTK Corpus [3]
  - 109人の話者 ・ 合計約44時間の音声.
  - 48kHz, 16ビットリニアPCM として収録.
  - 以下, これを x Hzにダウンサンプリングしたものを「元音声(x Hz)」と表記.

- 短時間フーリエ変換をする際のパラメータ
  - フレームサイズ : 256, フレームシフト : 192
  - 窓関数 : ハン窓

MP3による圧縮と元音声への復元

- MP3による圧縮にはffmpegを利用.
- ビットレートは48k, 96k, 128k, 196kbps の4種類.
- サンプリングレートは全て48kHzとして圧縮.
- 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).

CELPによる圧縮と復元

- 圧縮は次の手順で行う.

- 正解データは元音声(16kHz)とする.

- 復元は次の手順で行う.

ABテストによる主観的評価

- 2つの音声A, Bを聞き, 音声が良い方を選ぶ.
- 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
  - 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
  - 元音声(48kHz)
  - 元音声(24kHz)

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%

- 結果より, 元音声(48kHz) > GAN(48kbps MP3) > 48kbps MP3 という選択の傾向が認められる.

客観的評価

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.

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

- X：圧縮前の音声から得た対数パワースペクトログラム
- $\hat{X}$：比較したい音声から得た対数パワースペクトログラム
- L：$X,\hat{X}$の時間方向のサイズ
- K：$X,\hat{X}$の周波数方向のサイズ

MP3に関する実験結果：元音声(48kHz)に対するLSD

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

- LSD(LF)は圧縮の際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.
- どのビットレートでもモデルによる復元でLSD, LSD(LF)が改善されたことがわかる.

CELPに関する実験結果(客観的評価)

元音声(8kHz)に対するLSD

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

元音声(16kHz)に対するLSD

	LSD
8kHz CELP → Cubic補間	3.61
8kHz CELP → Cubic補間 → GAN(16kHz)	3.59
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

- 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．

フレームシフトによるLSDへの影響 (MP3)

元音声(48kHz)に対するLSD

	LSD	LSD(LF)
96kbps MP3 shift 64	3.28	1.91
GAN(96kbps MP3 shift 64)	1.32	1.15
96kbps MP3 shift128	3.28	1.91
GAN(96kbps MP3 shift 128)	1.32	1.17
96kbps MP3 shift 192	3.19	1.84
GAN(96kbps MP3 shift 192)	**1.27**	**1.11**
96kbps MP3 shift 224	3.27	1.90
GAN(96kbps MP3 shift 224)	1.30	1.15

- フレームシフト192のときのLSDが最も良い結果となっているが, 一般に逆短時間フーリエ変換の結果はフレームシフトの値が小さいほど元の波形に近くなるので一概にLSDの結果のみでは判断できない.

フレームシフトによるLSDへの影響 (CELP)

元音声(8kHz)に対するLSD

	LSD
CELP shift 32	2.40
GAN(CELP shift 32)	**2.31**
CELP shift 64	2.40
GAN(CELP shift 64)	**2.31**
CELP shift 128	2.42
GAN(CELP shift 128)	2.34
CELP shift 192	2.44
GAN(CELP shift 192)	2.37

元音声(16kHz)に対するLSD

	LSD
16kHz 補間音声 shift 64	3.38
GAN(16kHz 補間音声 shift 64)	3.03
16kHz 補間音声 shift 128	3.38
GAN(16kHz 補間音声 shift 128)	3.00
16kHz 補間音声 shift 192	3.39
GAN(16kHz 補間音声 shift192)	**2.96**

- 8kHz, 16kHzともにフレームシフトの値によってLSDに大きな差は見られなかった.
- 計算時間と音声品質を考慮しながらフレームシフトの値を決定する必要がある.

うまくいった例

圧縮した音声(CELP 8kHz)

復元した音声(16kHz)

高周波成分がうまく復元され, CELPによる圧縮を
施した音声特有のこもった音も改善されている.

うまくいかなかった例

圧縮した音声(CELP 8kHz)

復元した音声(16kHz)

復元した音声にノイズが入っている.

まとめと今後の課題

- 非可逆圧縮を施した音声の復元をGANを用いて行った.
- 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.
- 今後はさまざまなデータに対して学習することで, 音楽データなどに対してもより効率の良い圧縮を実現することが課題.

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] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300, 2015. [https://arxiv.org/abs/1512.09300]

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

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