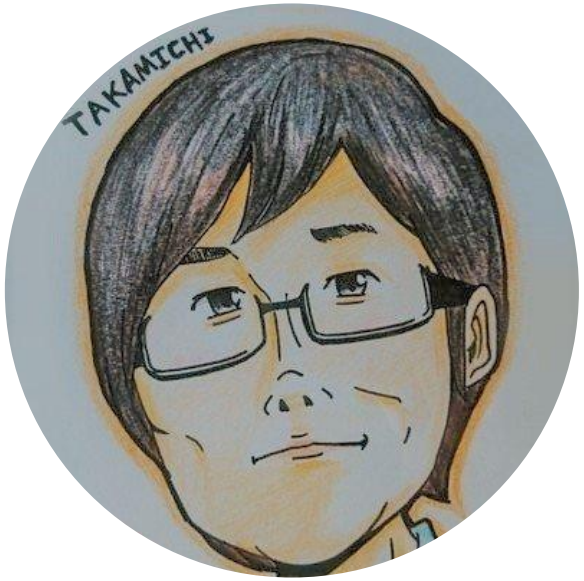
日本地球惑星科学連合大会 招待講演 (2023/05/21)

機械学習を用いた波形モデリング ～人間の音声の場合～

高道 慎之介 (東京大学)

*本資料の一部は以下の協力を得て作成しました．感謝申し上げます．
統数研 矢野恵佑 准教授，　東北大学D2 今井柊平 氏

自己紹介



@forthshinji

名前

高道 慎之介 (たかみち しんのすけ)

現職

東京大学 講師

経歴

熊本高専→長岡技大→奈良先端大.
2016年に博士 (工学).

専門

音声処理

最近の音声合成技術はすごい！

歌声合成技術 [Takamichi22]

松任谷由実氏が50年前にデビューした当時の歌声を人工再現する技術を開発

https://www.youtube.com/watch?v=oWo-TabDt8w
(学術目的での利用の許諾を受けて使用しております)

音声合成技術 [Matsunaga22]

人間のように自然に間違って喋る音声合成技術を開発

“ざっくりいうと，先ほど少しお話ししましたけども，戦後のそういうサブカルチャーのイメージという…”

ざっくりいうと，(ｱﾉｰ)先ほど(ｱﾉｰ)少し(ｱﾉｰ)お話ししましたけども，戦後のそういうサブカルチャーのイメージという…

[Takamichi22] https://www.i.u-tokyo.ac.jp/news/files/ist_pressrelease_20221003_takamichi_s.pdf
[Matsunaga22] https://arxiv.org/abs/2210.09815

本講演のテーマ

機械学習を用いた音声波形モデリングは、なぜ成功したのか？
地震波への応用可能性はあるか？

本スライドは私の個人ウェブサイトで公開予定です.

疎密波である音波のうち，人間の発声器官から
出るものを音声という．この音声はどういう波？

発声器官のモデル: ソース・フィルタモデル

声道（フィルタ）

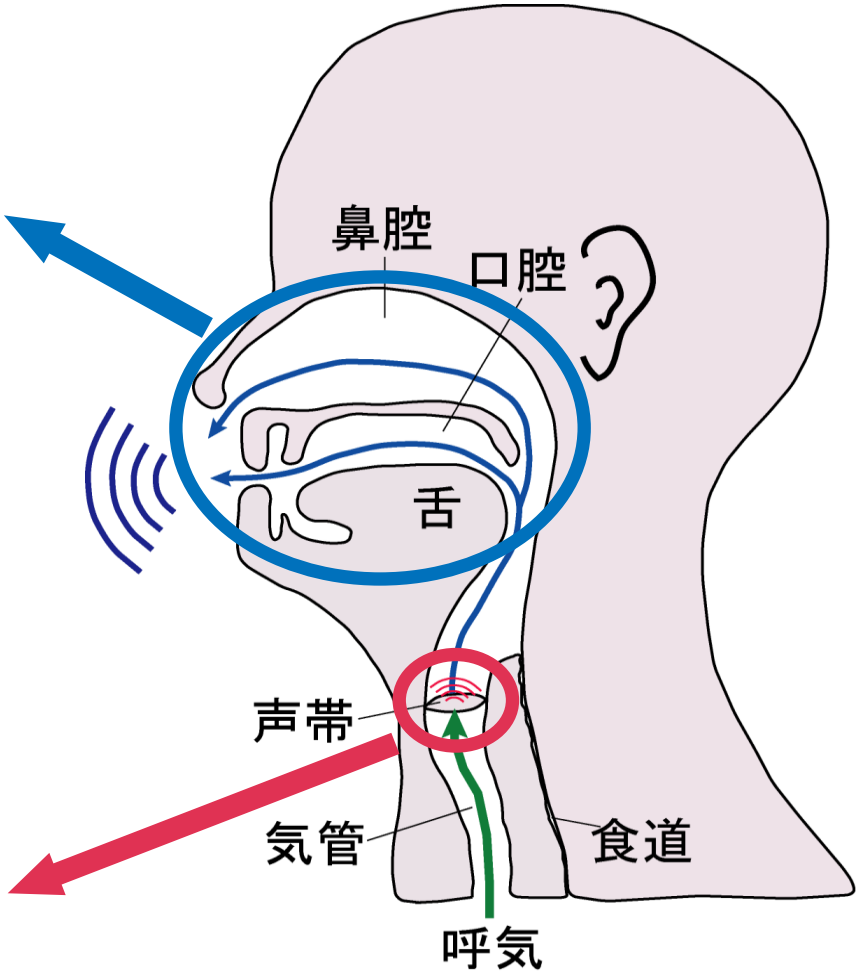
口や喉を動かして
音源信号を音色付け

音声波形

畳込み

音源（ソース）

肺からの呼気を
声帯の開閉で制御



音声波形を数式で表すと

音声生成の式．周波数領域における掛け算（時間領域なら畳み込み演算）

音源 $S(f, t)$ × フィルタ $F(f, t)$ × （チャネル） $C(f, t)$ = 音声 $X(f, t)$

“声の高さ”（時刻で変化）　“声色”（時刻で変化）　空間伝播・マイク特性（本講演では省略）　周波数　時刻

（注意： 次のページで体内を映した映像が出てきます．
苦手な方は音だけを聞いて下さい）

音声波形を数式で表すと

音声生成の式．周波数領域における掛け算（時間領域なら畳み込み演算）

音源 $S(f,\ t)$ × フィルタ $F(f,\ t)$ × （チャネル） $C(f,t)$ = 音声 $X(f,\ t)$

周波数　時刻

“声の高さ”（時刻で変化）

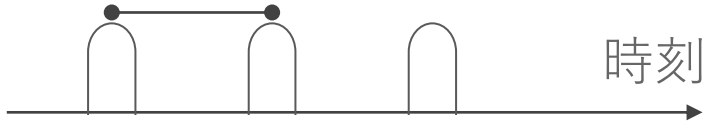
“声色”（時刻で変化）

空間伝播・マイク特性（本講演では省略）

https://www.youtube.com/watch?v=phBjL_BT5Zc

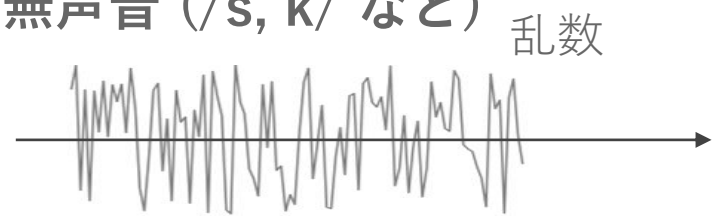
有声音（/a, i/ など）

声門の開閉周期に対応



振幅スペクトルは櫛形（スパース）

無声音（/s, k/ など）



振幅スペクトルは連続的

音声波形を数式で表すと

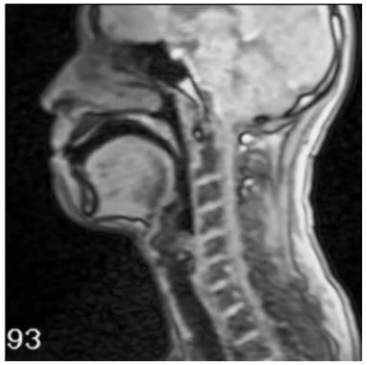
音声生成の式．周波数領域における掛け算（時間領域なら畳み込み演算）

$$\underset{\text{音源}}{S(f,t)} \times \underset{\text{フィルタ}}{F(f,t)} \times \underset{\text{（チャネル）}}{C(f,t)} = \underset{\text{音声}}{X(f,t)}$$

“声の高さ”（時刻で変化）　**“声色”（時刻で変化）**　空間伝播・マイク特性（本講演では省略）　周波数　時刻

リアルタイムMRI動画

- ▶ 現時点でもっとも理想的なデータ
  - 喉頭から唇・鼻孔まで声道の全体像が観察できる
  - 骨が写らない
  - 動画である（毎秒10～80フレーム程度）
  - 被曝の危険性がないので大量のデータ収集が可能
  - 工夫次第でさまざまな精度の情報をとれる
- ▶問題点
  - 撮像できる場所が世界で数箇所しかない
  - それなりにお金がかかる（90分利用するのに18万円）

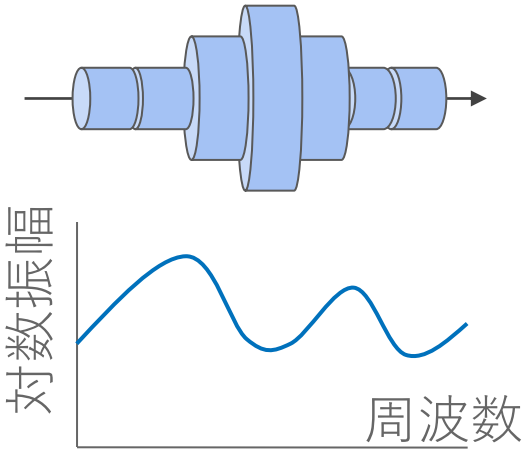


㈱ATR-Promotions 脳活動イメージングセンタにて撮像（14fps）東京語男性話者．デジタル信号処理でMRI装置の稼働音（かなりうるさい）を除去してある

https://www.youtube.com/watch?v=YcIN8CoHmqk&t=491s

音響管の連接でモデル化可能

自己回帰過程を仮定



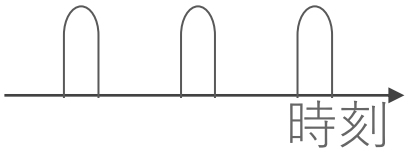
振幅スペクトルは連続的

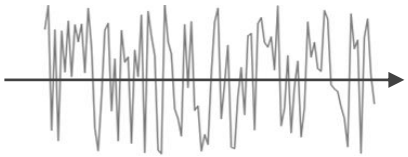
機械学習以前のボコーダ

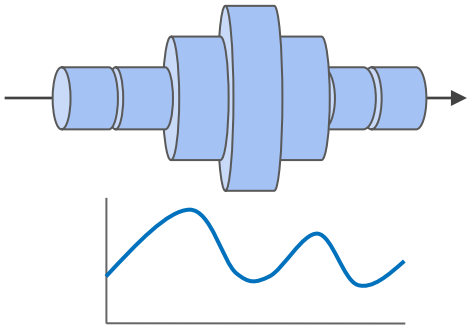
音声生成の式．周波数領域における掛け算（時間領域なら畳み込み演算）

音源 $S(f,\ t)$ × フィルタ $F(f,\ t)$ × （チャネル） $C(f,t)$ ＝ 音声 $X(f,\ t)$

“声の高さ”（時刻で変化）　“声色”（時刻で変化）　空間伝播・マイク特性（本講演では省略）　周波数　時刻







音源とマイクが離れているなら，距離の2乗で減衰

＊省略

暗黙的に仮定していること

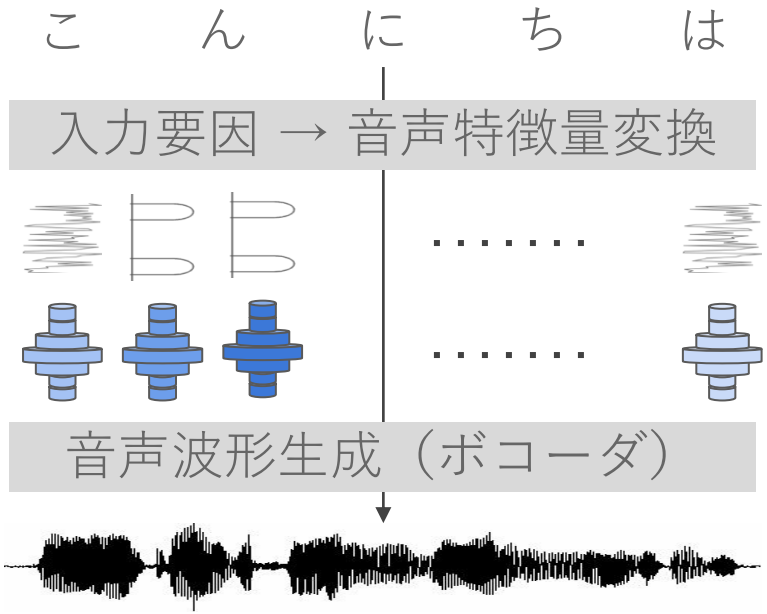
- **全体の仮定**
  - 時間区間内（音声の場合は20~25msec）では信号が定常
    - 時間区間をずらし（例えば5～10msc）ながら分析，合成
  - 各要素の畳み込みで音声が得られる
  - 位相の情報を捨てている（人間の聴覚は音声信号の位相に鈍感）
    - 例えば，乱数で生成 or 区間同士の整合性から推定 or 振幅から推定

- **音源の仮定**
  - 有声音と無声音でソフトな切り替えを行う
    - /a, i/ などの有声音の場合：声門の開閉周期に対応する周期波形列．決定論的信号として扱うことが多い．
    - /s, f/ などの無声音の場合：乱数．確率論的信号として扱う．

- **フィルタの仮定**
  - （古典的には）自己回帰過程を仮定． １つ前の時刻の信号に影響．
  - 振幅は，周波数に対して連続的

音声を制御する要因

- **音声を制御する要因は，音声に比べ概して時間解像度が低い**
  - 例えば日本語テキストを読み上げる場合
    - 日本語の話速： ひらがな 5〜8 文字/秒
    - 音声特徴： 100~200 フレーム/秒
      - → ひらがなの時間解像度より10倍以上高い
    - 音声信号： 16000~48000 サンプル/秒
      - → ひらがなの時間解像度より2000倍以上高い



深層学習を用いた音声波形合成技術がこの数年で大成功を納めた．この理由は何だろうか？

深層学習（機械学習）を用いた音声波形モデリング

推論

$$\hat{\boldsymbol{y}} = G_\theta(\boldsymbol{x})$$

合成波形　生成モデル　制御要因

学習

$$\hat{\theta} = \operatorname{argmax} L(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

生成モデルのパラメータ　目的関数　正解波形

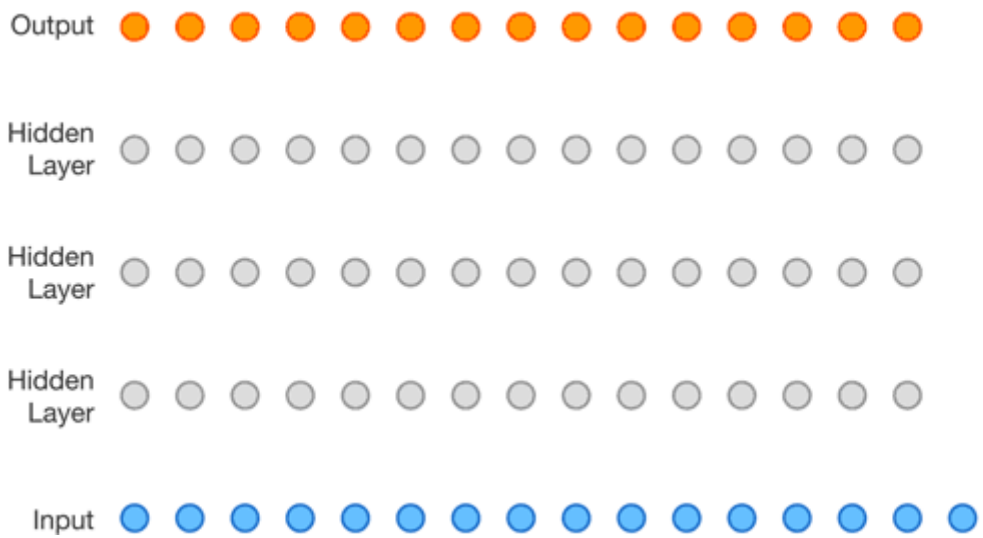
- どんな生成モデルが使われている？
- どんな目的関数が使われている？
- どんな制御要因・確率モデルなら上手く推論できる？
- 合成波形をどう評価すればよいか？

生成モデル①：
周辺の時刻の波形を利用する方法，およびその派生

- **自己回帰型信号処理の発展として位置づけられる方法**

$$y_t = G_\theta\left(y_{t-T}, \ldots, y_{t-1}, \boldsymbol{x}\right)$$

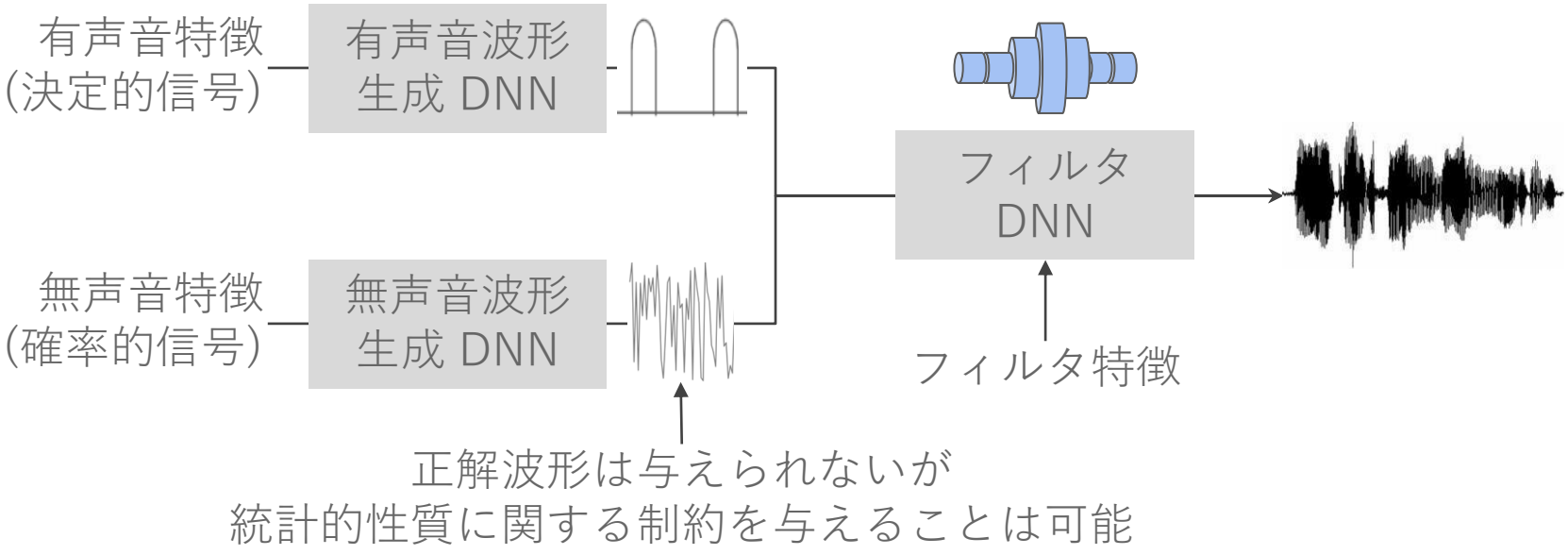
  - WaveNet [Oord16]： 深層学習 x 波形生成の最初の成功
    - Dilated convolution: 参照する過去波形の時間範囲を効率的に広げる
  - 高いモデル化性能から多くの派生が存在（自己回帰とは限らない）
    - ParallelWaveGAN [Yamamoto20]， DiffWave [Kong21]， etc.



[Oord16] https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio
[Yamamoto20] https://arxiv.org/abs/1910.11480 (ICASSP 2020)
[Kong21] https://arxiv.org/abs/2009.09761 (ICLR 2021)

生成モデル②：
音源とフィルタに分ける方法

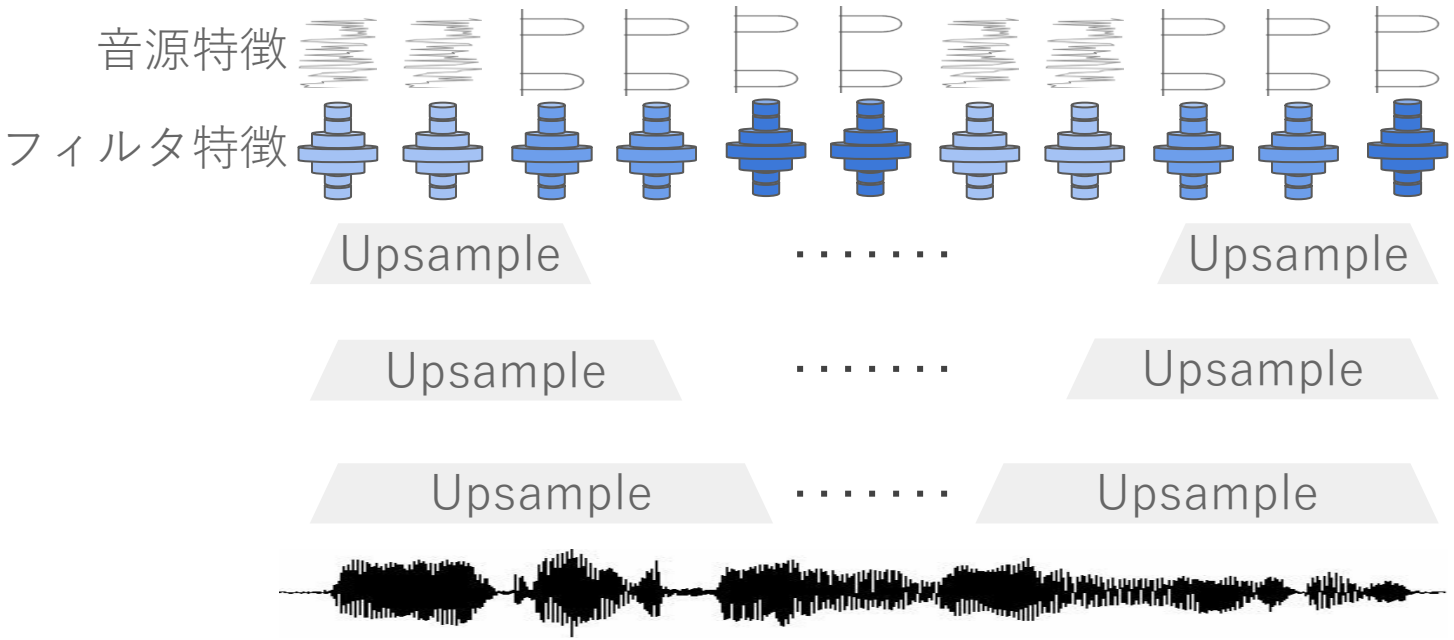
- **音源とフィルタの役割を明示的に分ける方法**
  - Neural source-filter [Wang20]: 音源モデルを周期信号＋乱数で駆動し，フィルタモデルを畳み込みモデルで表現
  - 音源とフィルタの可制御性の高さから，多くの派生が存在
    - DDSP（音源・フィルタ特徴を学習する自己教師あり学習）[Engel20]
    - Source-Filter HiFi-GAN [Yoneyama23]



[Wang20] https://arxiv.org/abs/1904.12088 (IEEE/ACM TASLP 2020)
[Engel20] https://arxiv.org/abs/2001.04643 (ICLR 2020)
[Yoneyama23] https://arxiv.org/abs/2210.15533 (ICASSP 2023)

生成モデル③：
制御要因の系列をアップサンプリングする方法

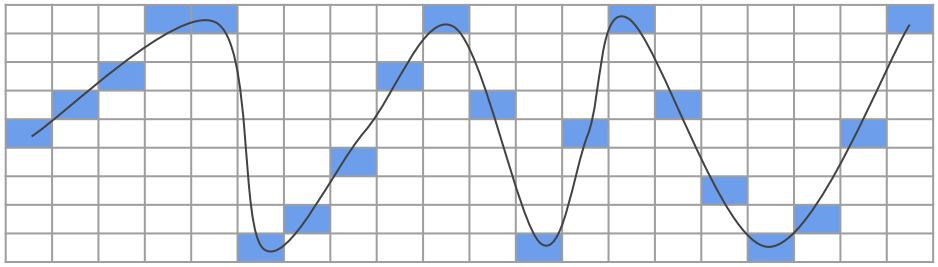
- **制御要因が「時間間隔の粗い系列」であることを利用する方法**
  - MelGAN [Kumar19] … アップサンプリング層の繰り返し
  - シンプルかつ高速な方法であるため，多くの派生が存在
    - HiFi-GAN [Kong20]，BigVGAN [Lee23] (汎化性能が売り), etc.



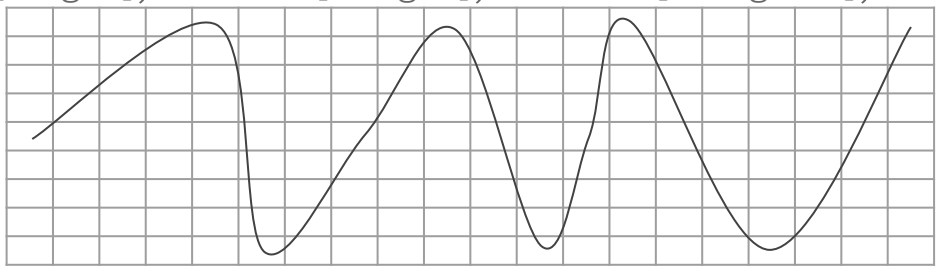
[Kumar19] https://arxiv.org/abs/1910.06711 (NeurIPS 2019)
[Kong20] https://arxiv.org/abs/2010.05646 (NeurIPS 2020)
[Lee23] https://arxiv.org/abs/2206.04658 (ICLR 2023)

学習の目的関数①：
波形レベルで正解に合わせにいく

- **最初期：波形を量子化して離散確率分布で表現**
  - 音声波形は 0 付近に集中するので 0 付近を細かく離散化



- **最近：(複数の時間幅・サンプリング周期で) 連続確率分布で表現**
  - 音声は，局所的には周期信号であり，低周波数が重要な信号
  - Gaussian [Ping19]， GAN [Kong20]， Flow [Prenger19]， Diffusion [Chen20]

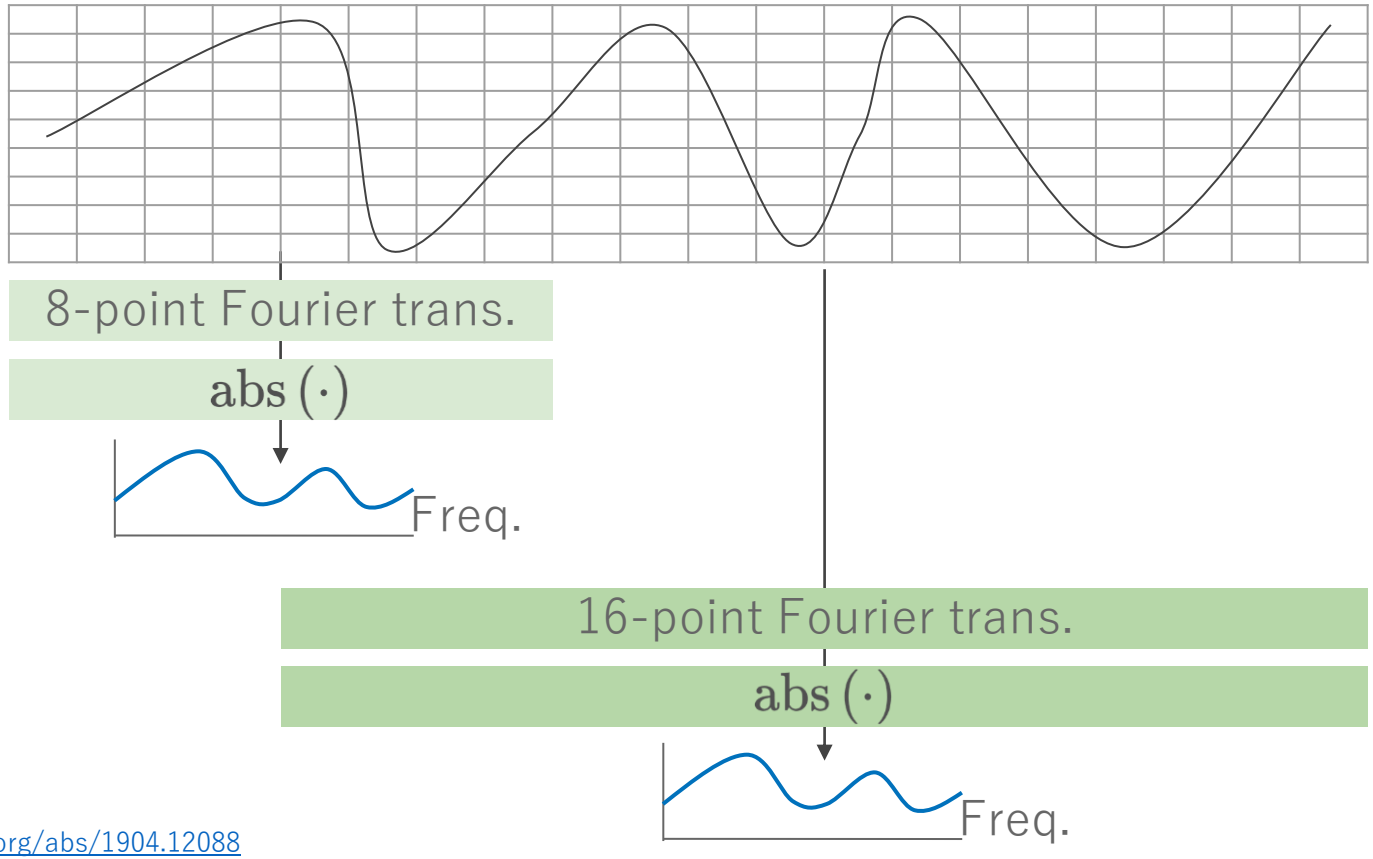


Downsample された音声波形を合わせにいく

[Oord16] https://arxiv.org/abs/1609.03499, [Ping19] https://openreview.net/pdf?id=HklY120cYm (ICLR 2019), [Kong20] https://arxiv.org/abs/2010.05646 (NeurIPS 2020), [Prenger19] https://arxiv.org/abs/1811.00002 (ICASSP 2019) [Chen20] https://arxiv.org/abs/2009.00713 (ICLR 2021)

学習の目的関数②：
振幅のみを正解に合わせにいく

- **音声の知覚には位相よりも振幅が重要であり，また，位相は推定しづらい．そのため，（少なくとも）振幅を正しく推定**
  - 複数の時間幅で振幅を計算し，それを正しく推定しにいく
  - 波形レベルの目的関数の補助としてよく使われる



[Wang20] https://arxiv.org/abs/1904.12088

決定的推論 vs. 確率的推論

- **推論時，音声波形は決定的に決めてよいのか？ それとも確率的？**

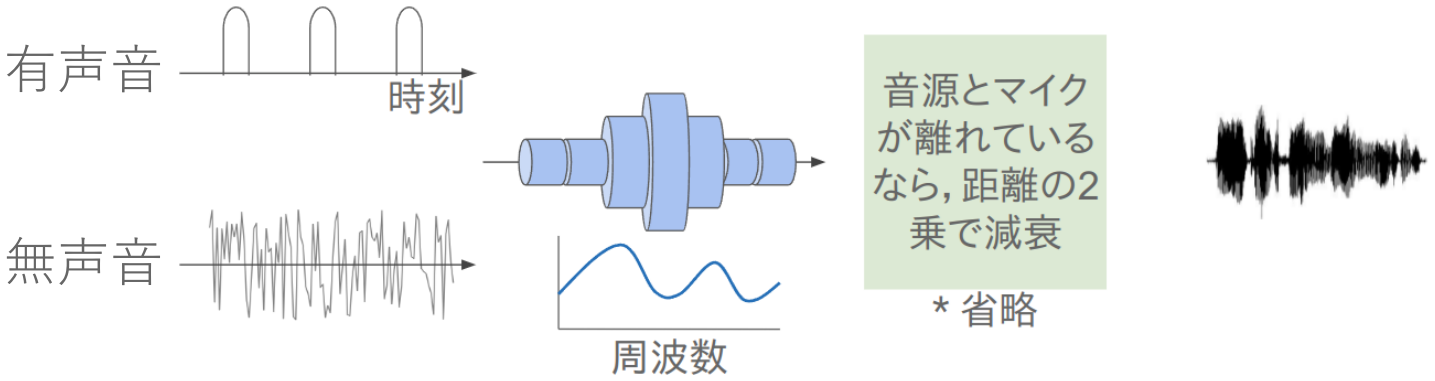
$$y_t = \mathrm{argmax}\ \mathrm{Prob}\left(y_t\right)$$

決定論的（例えば最尤推定）

$$y_t \sim \mathrm{Prob}\left(y_t\right)$$

確率論的

- **結論から言えば，有声音（音源が決定的）生成の場合は決定的でもよいが，無声音（音源が確率的）生成の場合は確率的であるべき**
  - 学習・（推論）を有声音と無声音で分ける場合も



長期依存性をどうモデル化すれば良い？

- **音声波形は時間的な長期依存性を持つ．それをどう学習する？**
  - 波形は解像度が高すぎるため，依存性の学習が困難 [Kong21]
  - 解像度の低い表現を学習し依存性を学習 [Ren22][Lakhotia21]
  - 音声波形自体は高解像度だが，少数クラスで表現可能なことを利用
    - たとえば「あいうえお」のような平仮名

[Kong21] https://arxiv.org/abs/2009.09761 (ICLR 2021)
[Ren22] https://arxiv.org/abs/2110.13900 (IEEE JSTSP)
[Lakhotia21] https://aclanthology.org/2021.tacl-1.79/ (ACL 2021)

合成音声を評価する

- **出てきた合成音声を評価するにはどうしたら良いか？**
  - 基本的には，人間による主観評価
  - 客観指標にはどのようなものがあるか？

- **入力した制御要因をどの程度を反映できているか？**
  - 系列の反映度： 例えば音声認識スコア（合成音声をテキスト認識）
  - ベクトルの反映度： 例えば話者認識スコア
    - 目標とする話者の声色をどの程度再現できているか？
  - disentanglement： 音源・フィルタを個別制御できるか？

- **“信号の質”はどの程度か？**
  - 振幅誤差（音声の知覚に振幅が重要であるため）
  - 主観評価値の予測： 人間の主観評価値を予測する機械学習モデル

地震波モデリングへの応用可能性は有るか？
(地震波の専門家ではありませんが・・・)

地震波モデリングへの応用可能性は有るか？

音源 $S(f,\ t)$ × フィルタ $F(f,\ t)$ × チャネル $C(f,t)$ = 音声 $X(f,\ t)$

断層破壊 $S(f,\ t)$ × 伝播媒質 $F(f,\ t)$ × サイト増幅 $C(f,t)$ = 地震波 $X(f,\ t)$

- 物理的な生成過程を模擬したモデルアーキテクチャ
- 各要素の統計的性質を利用した正則化
- 生成波形の(知覚における)重要度を考慮した学習
- 少数クラス・低時間解像度での表現学習
- 決定論的・確率論的信号を分離したアーキテクチャ・学習

まとめ

まとめ

- 音声合成（機械学習を用いた音声波形生成）はここまできた
- 音声波形の生成過程・統計的性質
- 機械学習を用いた音声波形生成
- 機械学習を用いた地震波形生成に応用できるか？