第138回MUS研究発表会
【特別企画2】歌声情報処理の過去・現在・未来

ここまで来た＆これから来る
歌声合成・歌声変換技術

高道 慎之介
（東京大学）

MOONSHOT
RESEARCH & DEVELOPMENT PROGRAM
AVATAR SYMBIOTIC SOCIETY

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 (主に音声合成)

時を超えて蘇る50年前の歌声 ～スモールデータを用いたタスク混合深層学習による歌唱再現～ (2022)

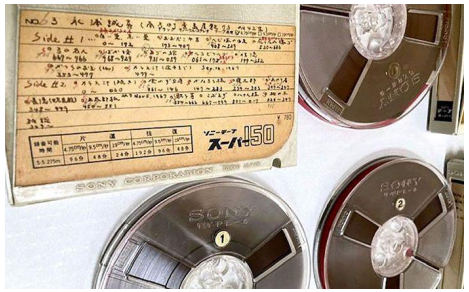
“高道慎之介助教を中心とした研究チームは、(中略)、**歌手の松任谷由実氏が50年前にデビューした当時の歌声を人工再現する技術**を開発しました。(中略)、当時の声色と歌唱表現を忠実に再現することに成功しました。”

https://www.i.u-tokyo.ac.jp/news/files/ist_pressrelease_20221003_takamichi_s.pdf

その他にこんなことをやっています

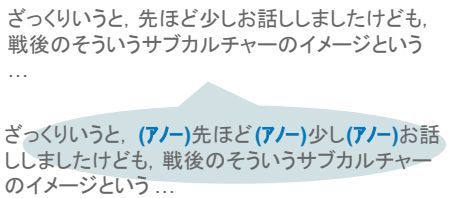
「HUNTER×HUNTER」ボイスチェンジャー (2019)

音声変換技術（人の声を違う声に変える技術）を使って，「誰が喋ってもリアルタイムにそのキャラクターの声になれる」技術を開発．



失われつつある「方言」と「昔話」を復元 (2023)

1960～年代に録音された古い方言昔話音声を現代品質に蘇らせる機械学習技術を開発．また，方言昔話音声データベースを頒布．



人間のように「言い淀む」音声合成 (2023)

考えながら喋ると人間は言い淀む．自然かつ自動で言い淀む音声合成技術を開発．

人間のように「笑う」音声合成 (2023)

「笑い声シンボル言語モデル」を開発し，自然な笑い声を確率的に生成できる生成モデルを開発．

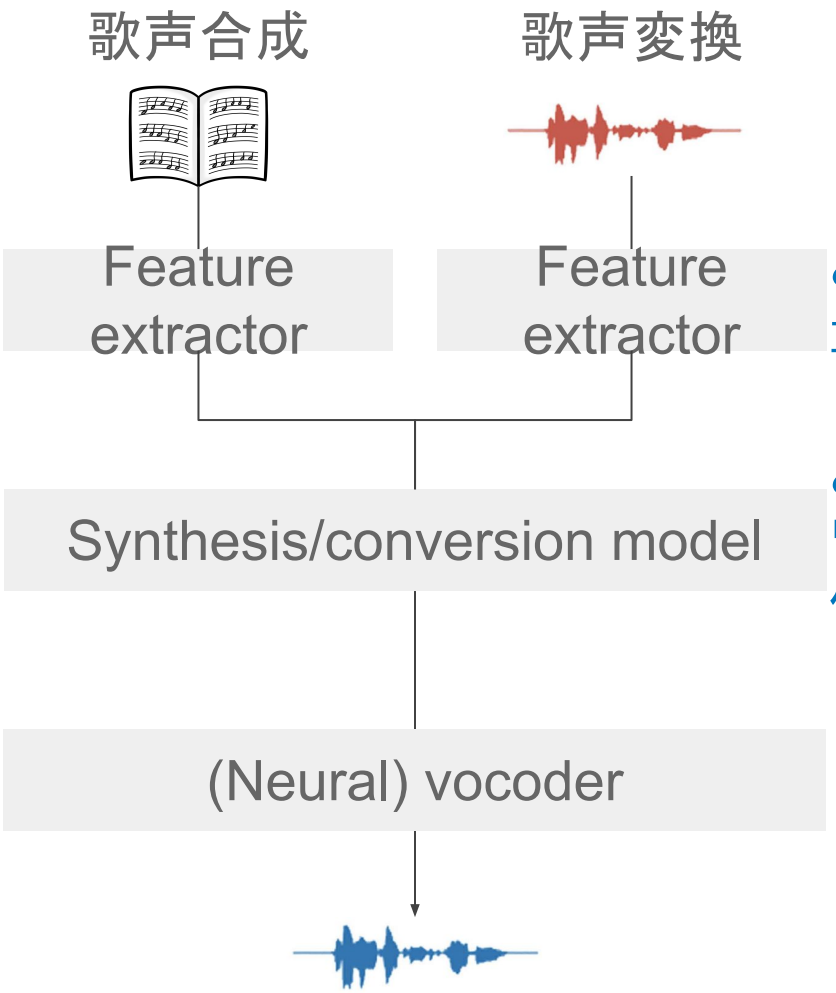
本講義の内容

ここまで来た＆これから来る歌声合成・歌声変換技術．
前半は最近の技術発展，後半は今後の進み方

本スライドは高道のHPで公開予定です．
Google で「shinnosuke takamichi talk」で調べて下さい．

最近の歌声合成の発展

歌声合成・歌声変換のフロー

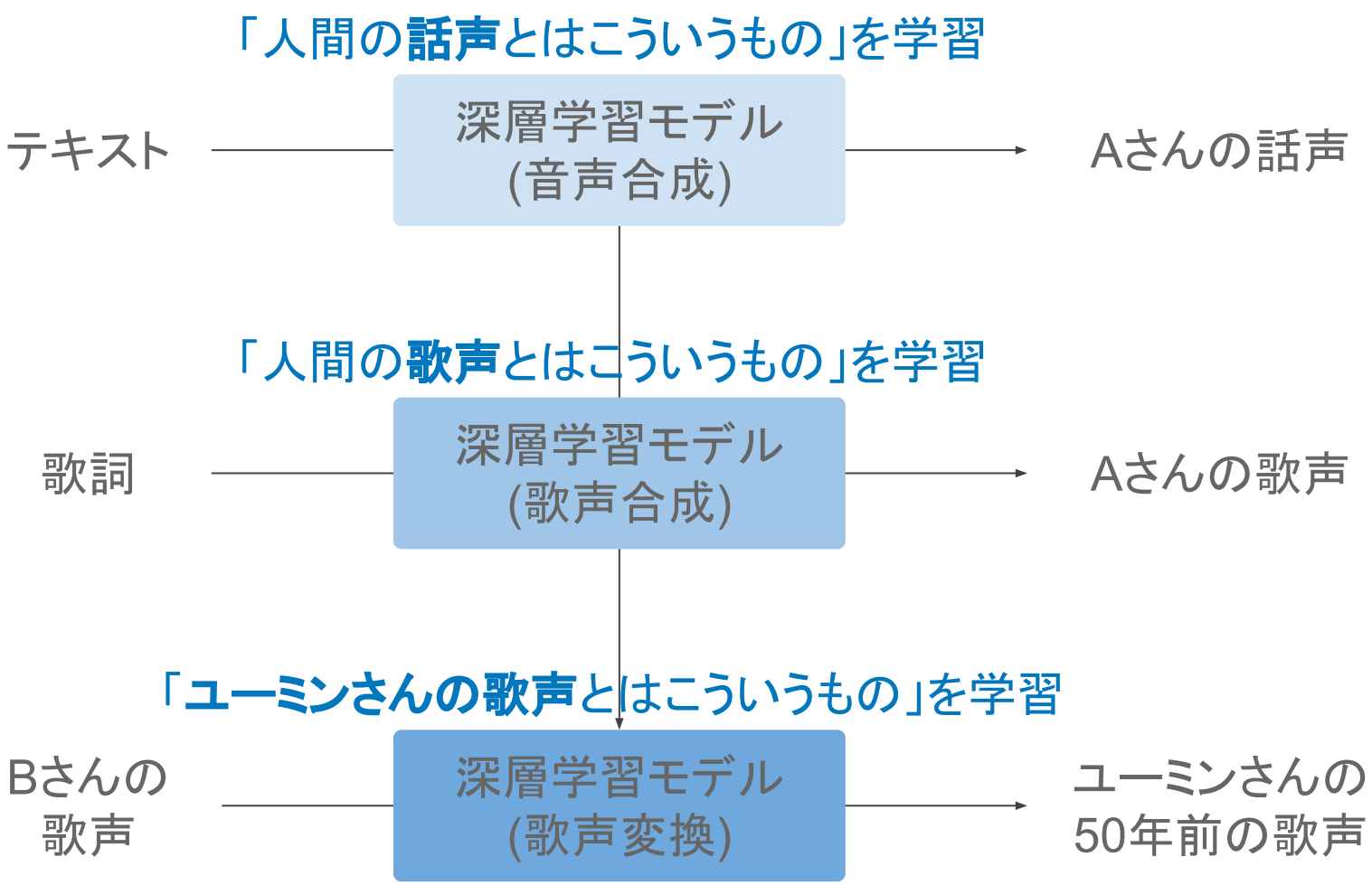


どんな楽譜・音声特徴量が使われるのか？

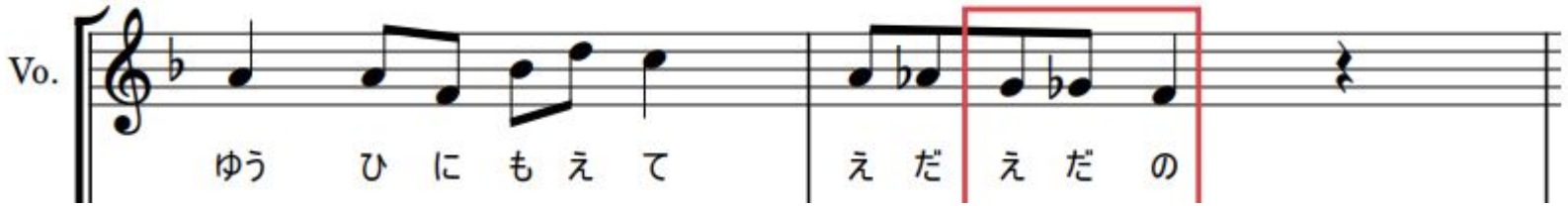
どんなニューラルネットワーク (DNN)が使われるのか？

どんなニューラルネットワーク (DNN)が使われるのか？

余談：先ほどの「Call me back」は多段階学習



どんな楽譜特徴量が使われるのか？



[Nakamura23] より引用

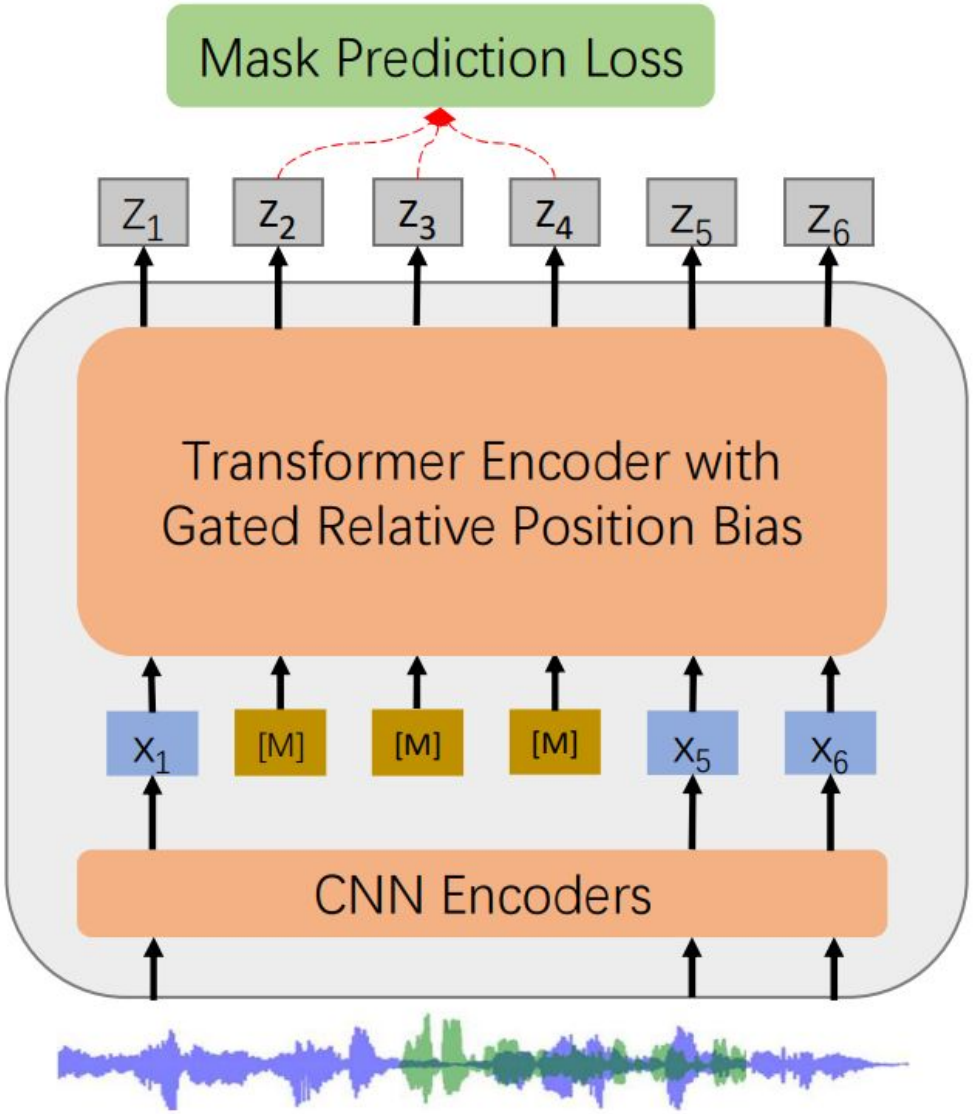
- **言語特徴量**
  - 発音：音素 + 位置 + 言語依存特徴
    - 英語 stress，日本語 mora [Hono21]，韓国語 語頭・核・語尾音素 [Lee19]
  - 意味：BERT [Zhou22]

- **音楽特徴量**
  - 音符レベル：音符，ピッチ，音符継続長 [Yamamoto23]
  - 全体レベル：テンポ，ビート [Shi22]

Hand-crafted 特徴量を使うことが大半である印象.
Data-driven 特徴量 (BERT, LLM) の併用は少しずつ出てきている？

[Nakamura23] Nakamura, “jaCappella Corpus: A Japanese a Cappella Vocal Ensemble Corpus,” ICASSP 2023.
[Yamamoto23] Yamamoto, “NNSVS: A Neural Network-Based Singing Voice Synthesis Toolkit”, ICASSP 2023.
[Lee19] Lee, “Adversarially Trained End-to-end Korean Singing Voice Synthesis System” Interspeech 2019.
[Shi22] Shi, ”Muskits: an End-to-End Music Processing Toolkit for Singing Voice Synthesis”, Interspeech 2022.
[Zhou22] Zhou, “Towards Improving the Expressiveness of Singing Voice Synthesis with BERT Derived Semantic Information”, Interspeech, 2022.

どんな歌声特徴量が使われるのか？

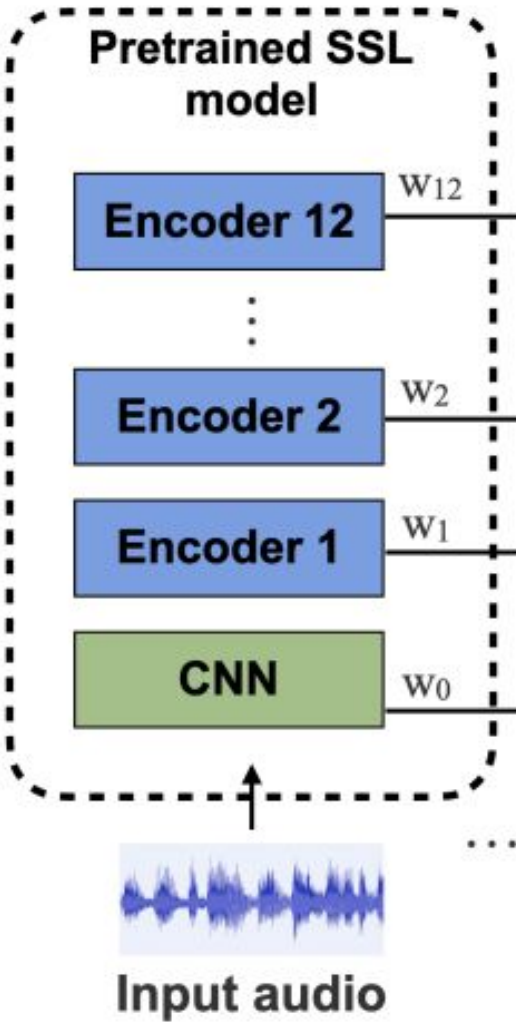


- **既存の特徴量**
  - 音素事後確率，信号処理ベース特徴量（例えばメルスペクトログラム）
  - 基本周波数

- **自己教師あり学習 (SSL)**
  - WavLM (左図)，HuBERT，ContentVec
  - 歌声で学習されたものは（残念ながら）多くないため，話し声で学習されたものを利用することが多い．

[Chen21] “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” 2021.

自己教師あり学習は，歌声の何を表しているのか？



- **SSLモデルの各層は何を表している？**
- **各層の役割**
  - 高い層：音韻など
  - 低い層：歌唱テクニック，歌唱者
- **一方で話声の場合は**
  - 高い層：意味
  - 中間層：パラ言語
  - 低い層：音韻

[Yamamoto23] “Toward Leveraging Pre-Trained Self-Supervised Frontends for Automatic Singing Voice Understanding Tasks: Three Case Studies,” APSIPA2023.

合成モデル・変換モデルに
どんなニューラルネットワークが使われるのか？

- **テキスト音声合成や古典的な歌声合成の展開と見做されるもの**
  - Cascade 型：time-lag + duration + acoustic model [Hono21]
  - End-to-end 型：
    - Duration-free 型：source-target attention 派生 [Angelini19]
    - Duration-aware 型：self attention 派生 [Lee21][Zhang22]

- **画像生成の展開と見做されるもの**
  - Diffusion model (拡散モデル):
    - 拡散モデル：事前分布からの反復的逆拡散で目的信号を生成する方法.
    - Latent diffusion 派生：音声波形を直接予測するのではなく，その低次元表現を拡散モデルで予測する [Hwang23]
  - Consistency model (一貫性モデル):
    - 一貫性モデル：拡散モデルの反復的逆拡散を1ステップで達成する蒸留
    - 歌声合成の場合は，楽譜から事前分布を予測し逆拡散 [Ye23]

[Hono21] Hono, “Sinsy: A Deep Neural Network-Based Singing Voice Synthesis System,” IEEE TASLP 2021.
[Abgelini19] Angelini, “Singing synthesis: with a little help from my attention”, Interspeech 2020.
[Lee21] Lee, “N-Singer: A Non-Autoregressive Korean Singing Voice Synthesis System for Pronunciation Enhancement,” Interspeech 2021.
[Zhang23] Zhang, “VISinger 2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer”, Interspeech 2023.
[Hwang23] Hwang, “HiddenSinger: High-Quality Singing Voice Synthesis via Neural Audio Codec and Latent Diffusion Models”, arXiv 2023.
[Ye23] Ye, “CoMoSpeech: One-Step Speech and Singing Voice Synthesis via Consistency Model”, arxiv 2023.

ボコーダに
どんなニューラルネットワークが使われるのか？

- **話声合成と歌声合成は，その処理手順は似ている一方，要請される条件が異なる．**
  - 外れ値への頑健性：話声合成はテキスト側，歌声合成は声側に課される場合が多い（個人的な印象）
    - 文の読み上げ，歌詞の歌い上げの範囲に限る
  - 信号レベルでの制御：ピッチなど

- **外れ値の入力に対して頑健なボコーダ**
  - HiFi-GAN：upsampling 層の反復 [Kong20] (経験上そこそこ頑健)
  - BigVGAN：HiFi-GAN + anti-alias 処理 + 大規模学習 [Lee23]

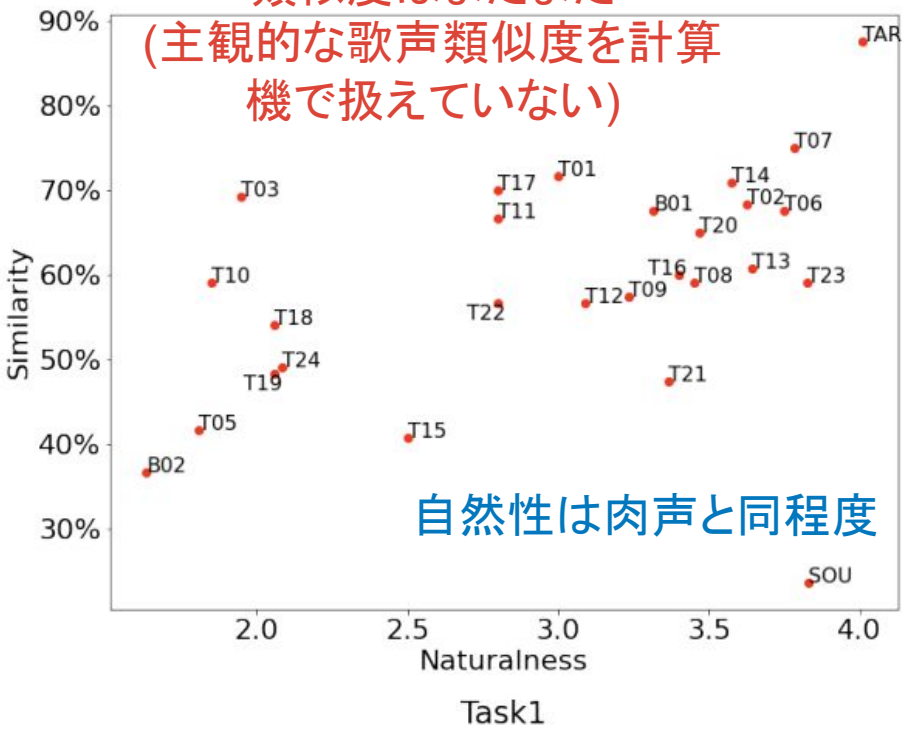
- **ピッチを別途に制御可能なボコーダ**
  - ソースフィルタ系：SiFi-GAN [Yoneyama23], DDSP-subtractive [Wu22]
  - 加算ボコーダ系：Neural-WaveTable [Shan22], DDSP [Engel21]

[Kong20] Kong, HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," NeurLIPS, 2020.
[Lee23] Lee, "BigVGAN: A Universal Neural Vocoder with Large-Scale Training," ICLR 2023
[Yoneyama23] Yoneyama, "Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder," ICASSP 2023.
[Wu22] "DDSP-based Singing Vocoders: A New Subtractive-based Synthesizer and A Comprehensive Evaluation" ISMIR 2022.
[Shan22] "Differentiable Wavetable Synthesis", ICASSP 2022.
[Engel21] "DDSP: Differentiable digital signal processing", ICML 2021.

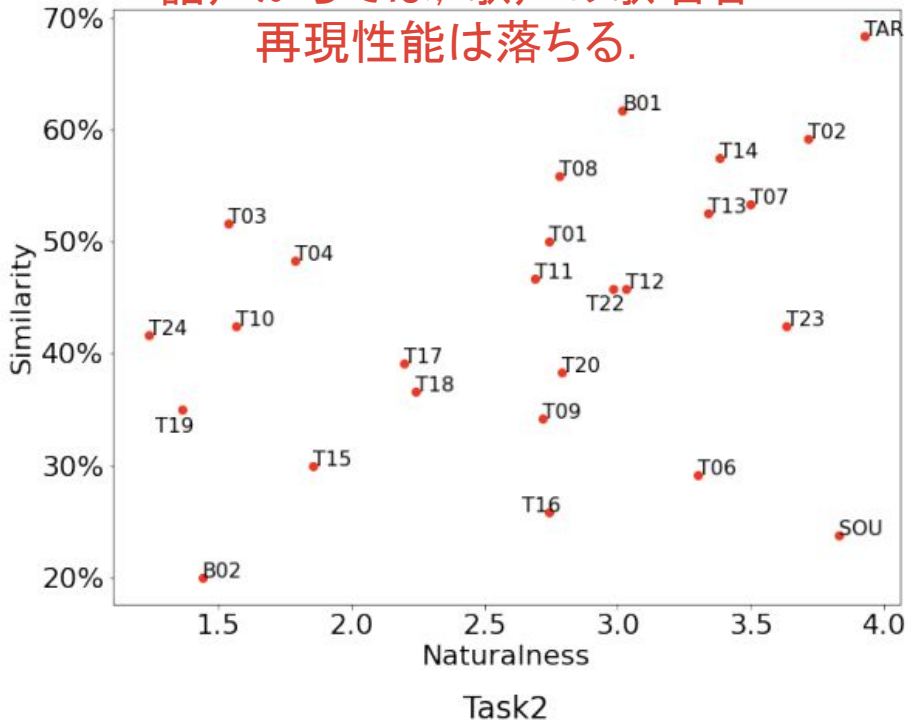
Singing voice conversion challenge 2023

- **同じ学習データの下で歌声変換の性能を評価するチャレンジ**
  - Task 1: 目標歌唱者の歌声が学習データとして与えられ，元歌唱者の声を目標歌唱者の歌声に変換
  - Task 2: 目標歌唱者の話声が学習データとして与えられ，元歌唱者の声を目標歌唱者の歌声に変換

類似度はまだまだ
(主観的な歌声類似度を計算機で扱えていない)



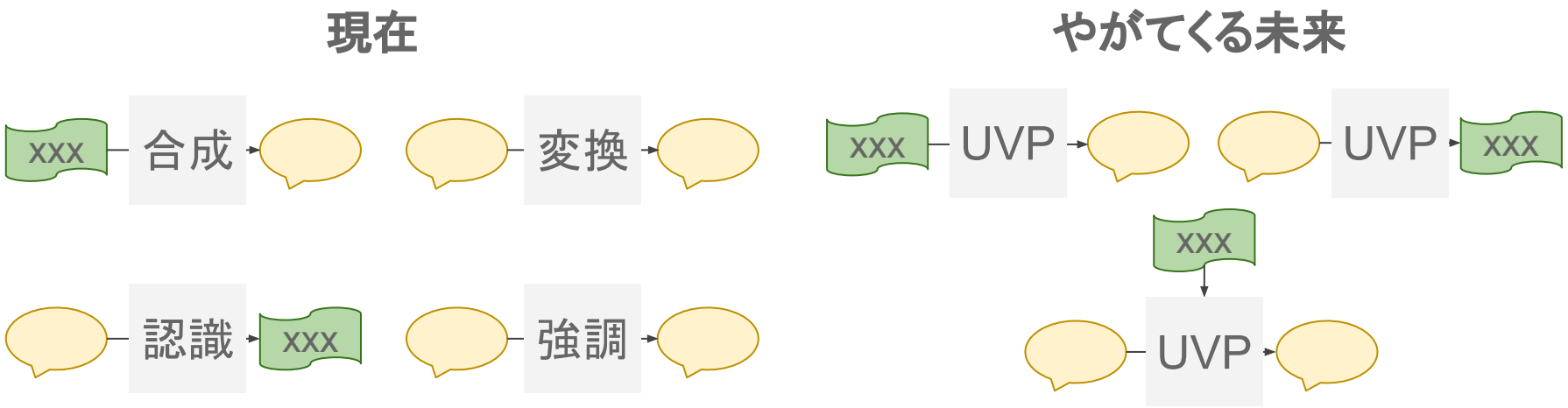
話声からでは，歌声の歌唱者再現性能は落ちる．



[Huang23] Huang “The Singing Voice Conversion Challenge 2023”, arXiv, 2023.

これから何が起こるのか (私見たっぷり)

Universal voice processor に向けて：その評価



- **汎用音声・歌声処理器 (universal voice processor) はもうすぐ？**
  - 言語理解を含めた音声・歌声の認識・加工・合成へ
  - この数年でやがて汎用へ？

- **これをどう評価したら良い？ 言語と音声・歌声の両面から.**
  - どう付き合ったら良い？
  - 言語理解＋音声・歌声の評価法およびデータセットが必要
    - 聴取環境・聴取者の条件付けも含めて

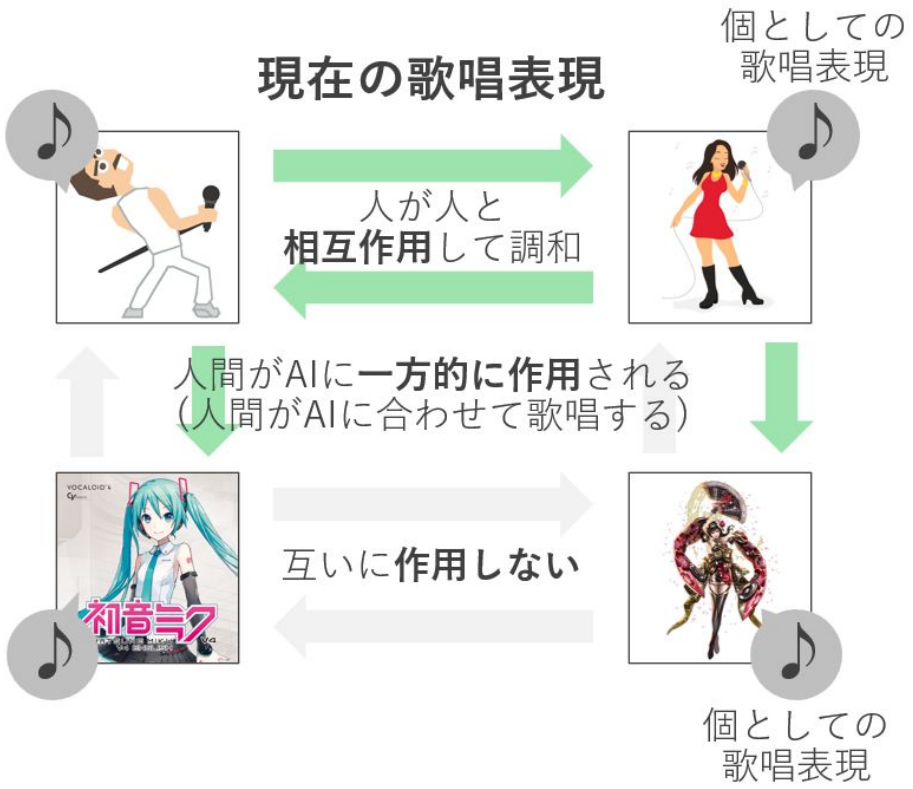
自律する歌声合成：自分の声を聴く歌声合成へ

- **自発性と自律性**
  - 自発性：自己の内部の原因によって行われる性質
  - (行動の)自律性：環境と相互作用して，介入無く動作する性質
  - 共通するのはゴースト (自己のアイデンティティを司る心的機能)？

- **現在の方法論の根本的な問題：音声合成AI，歌声合成AIは自分の出力する音声を聞いていない**
  - 人間は，自分の音声をモニタリングする機構を有する（”言葉の鎖”）
  - 同様に，内省による事後の理解も必要なのでは？
  - まずは，この機構を計算機的に実現することが目下の課題
    - モニタリング，フィードバック機構（自己回帰に限らない）
    - 内省の評価関数

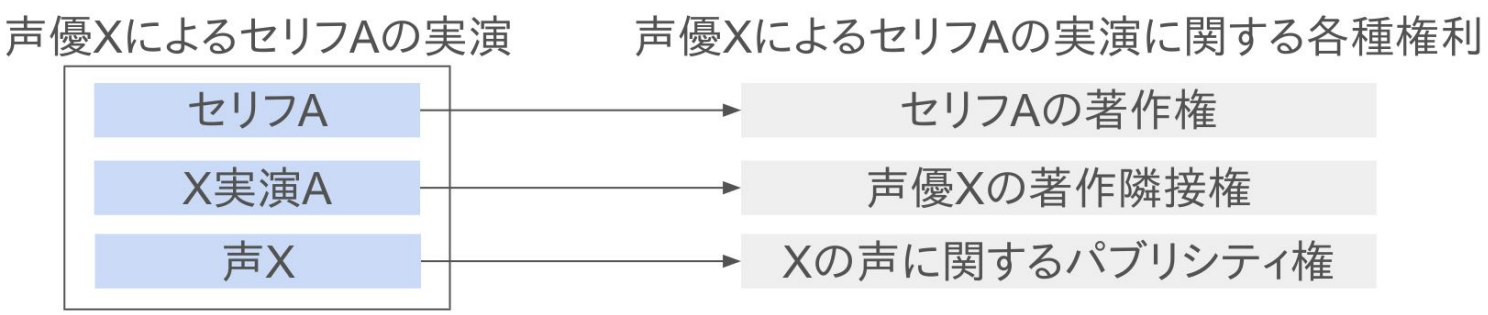
自律する歌声合成：周りと相互作用する歌声合成

- **歌声合成が人間の道具を超えて自律した存在となるとき，何が必要だろうか？**
  - 他の人間，他のAI（下の例），あるいは環境との相互作用？
  - 社会的立場の理解？



*群: 重唱・合唱のように複数の個が歌唱する形態

生成AIと権利 (CEDEC2023での講演資料より)



- ゲーム内声優音声に関する権利関係としては、①セリフの著作権、②声優の実演に関する著作隣接権（セリフという著作物を実演することにより生じる権利）、③声優の声に関するパブリシティ権が存在する。
  - ①セリフの著作権：セリフを作成した著作者（脚本家等）に生じる。
  - ②声優の実演に関する著作隣接権：実演をした声優という実演家に生じる。
  - ③声優の声に関するパブリシティ権：声優に生じる。

1 ピンク・レディー事件（最高裁平成24年2月2日判決）

- 人の氏名，肖像等（以下，併せて「**肖像等**」という。）は，個人の人格の象徴であるから，当該個人は，人格権に由来するものとして，これをみだりに利用されない権利を有すると解される（略）。
- そして，**肖像等は，商品の販売等を促進する顧客吸引力を有する場合があり，このような顧客吸引力を排他的に利用する権利（以下「パブリシティ権」という。）**は，肖像等それ自体の商業的価値に基づくものであるから，上記の人格権に由来する権利の一内容を構成するものということができる。

’音声合成と権利の関係を知る資料になっています．後日公開予定．

[高道23] Huang “AI音声合成の技術動向”, CEDEC, 2023.

まとめ

まとめ

- **最近の歌声合成の発展**
  - 楽譜特徴量・歌声特徴量
  - 自己教師あり学習が表すもの
  - 合成モデル・変換モデル・ボコーダモデル
  - Singing voice conversion challenge 2023

- **これから何が起こるのか**
  - Universal voice processor
  - 自律する歌声合成
  - 音声合成AIと権利