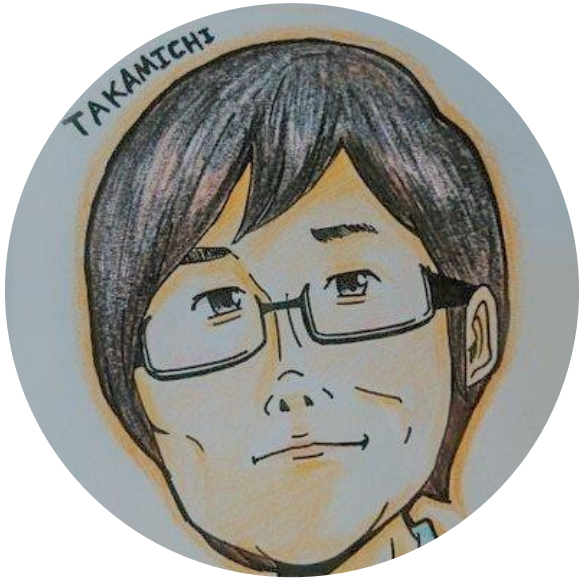
日本大学 文理学部
マルチメディア情報処理 (2024/07/23)

音声合成・歌声合成が拓く未来

高道 慎之介 (慶應義塾大学)

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 → 慶應義塾大学 准教授

経歴

離島→熊本高専→長岡技大→奈良先端大
2016年に博士 (工学).

兼務

東京大学　特任准教授
理研　客員研究員
他

本講演のテーマ

計算機による声合成はどこまで来たのか？
これからどう進んでいくのか？

音声情報処理の基礎知識

発声器官のモデル: ソース・フィルタモデル

声道 (フィルタ)

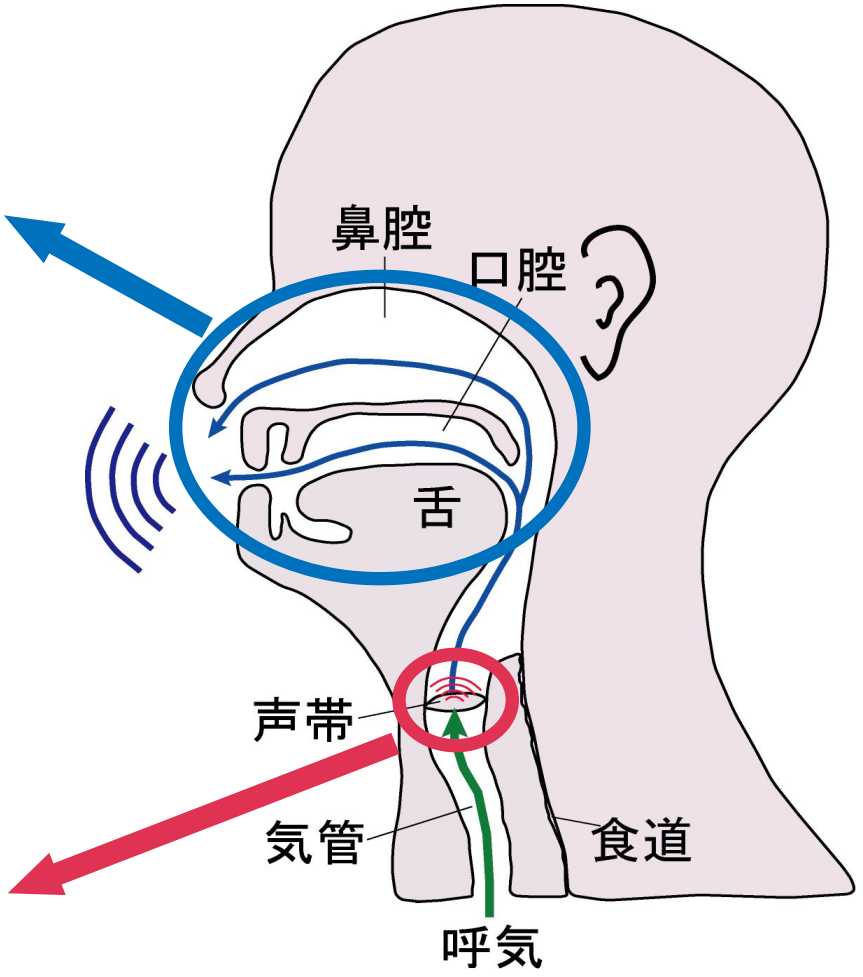
口や喉を動かして
音源信号を音色付け

音声波形

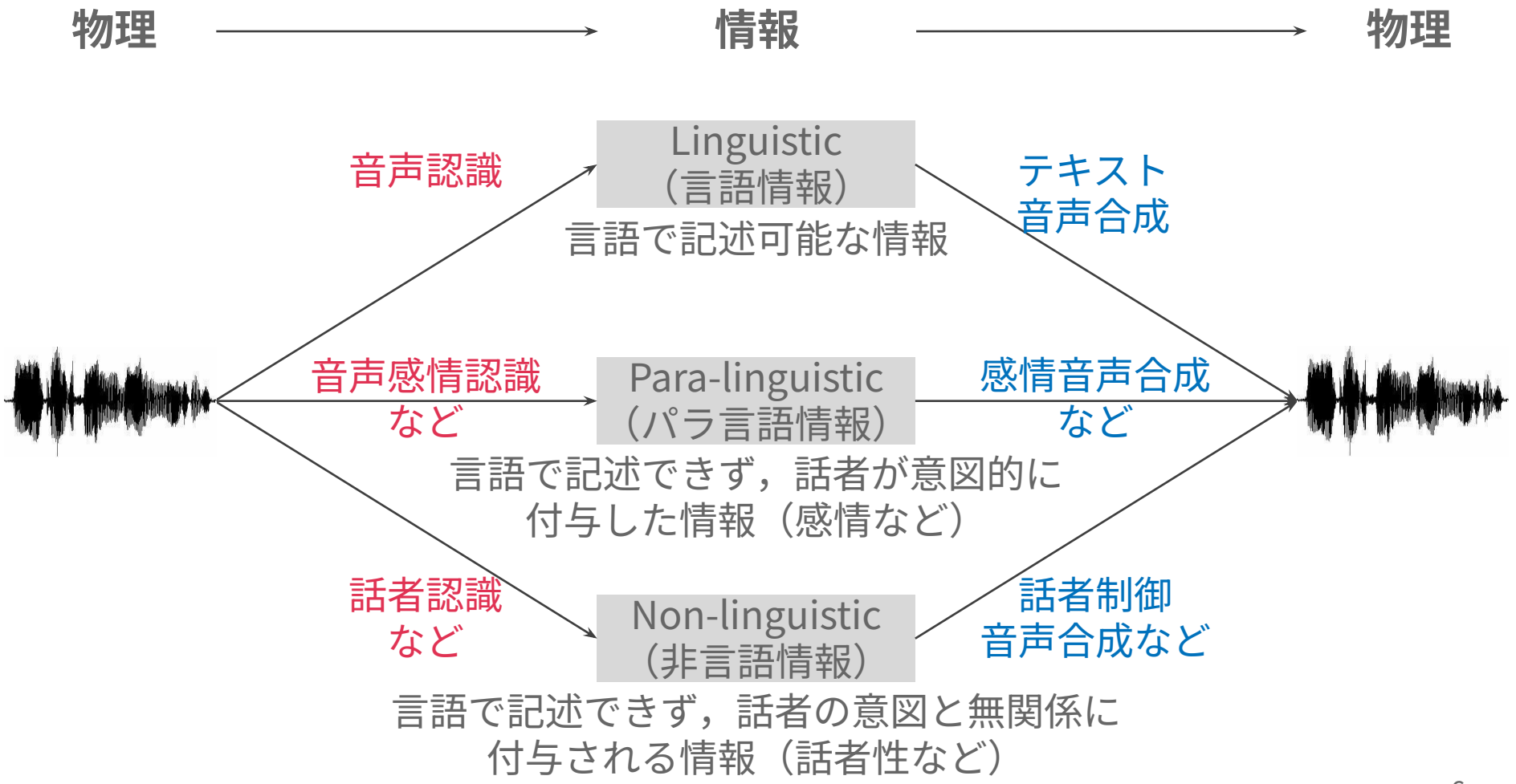
畳込み

音源 (ソース)

肺からの呼気を
声帯の開閉で制御



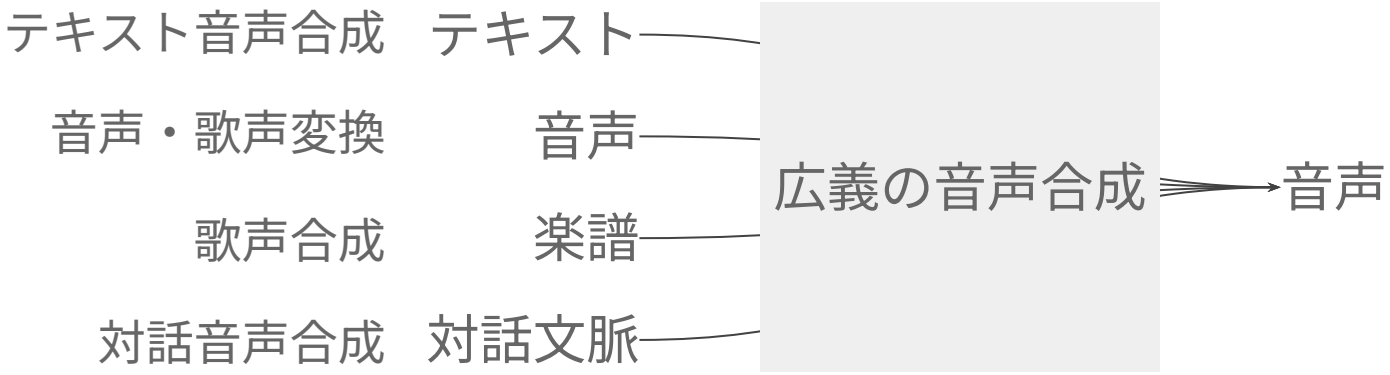
音声のもつ情報とそれを扱う主な音声技術



音声合成とは

- **狭義には，テキストから音声を合成する技術**
  - ＝テキスト音声合成

- **広義には，何らかの情報から音声を合成する技術**
  - {音声, 動画, 話者情報, 概念} to speech



紹介する研究紹介 20連発！(2023年以降を中心に)

研究事例紹介①：人間の音声表現を拡張する

研究事例紹介②：音声の文化を守る

研究事例紹介③：コンピュータが心をもったときに

研究事例紹介④：歌の可能性を広げる

研究事例紹介⑤：音声に関する感性を定量化する

研究事例紹介⑥：環境音を人工的に作り出す

研究事例紹介①
人間の音声表現を拡張する技術

ハンターハンター ボイスチェンジャー

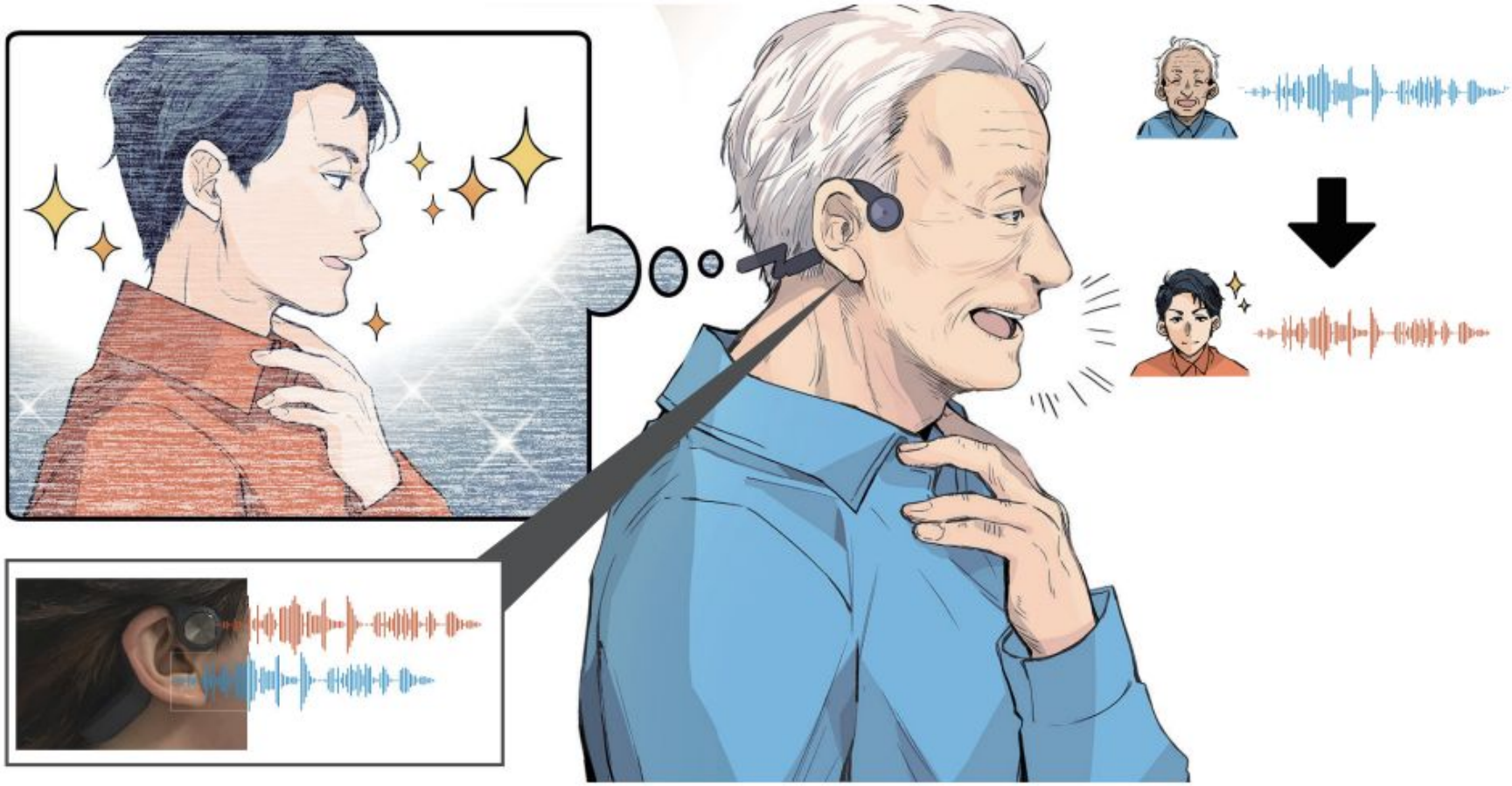
“自分の声が『HUNTER×HUNTER』のキャラクターの声に変換されてスピーカーから流れ、なりきりアフレコ体験ができます”

https://prtimes.jp/main/html/rd/p/000003399.000002581.html

スマホで動くリアルタイムボイスチェンジャー

https://www.youtube.com/watch?v=mc7HXXQ8P7Q

Digital speech makeup：音声の「鏡」を作る



自分で聞く自分の声が変わると，自己の認識が変わる．認識が変わると行動が変わる

Arakawa et al., “Digital Speech Makeup: Voice Conversion Based Altered Auditory Feedback for Transforming Self-Representation” ICMI 2021.

エモーションキャンセリング：音声への「盾」を作る

話し手の音声と聞き手の知覚感情を明らかにし，聞き手の心的負担を軽減する

https://www.softbank.jp/corp/news/press/sbkk/2024/20240515_01/

Spatial voice conversion:
空間上の狙った声だけ変える

Two persons are speaking to me (woman from left, man from right.)

Spatial voice conversion

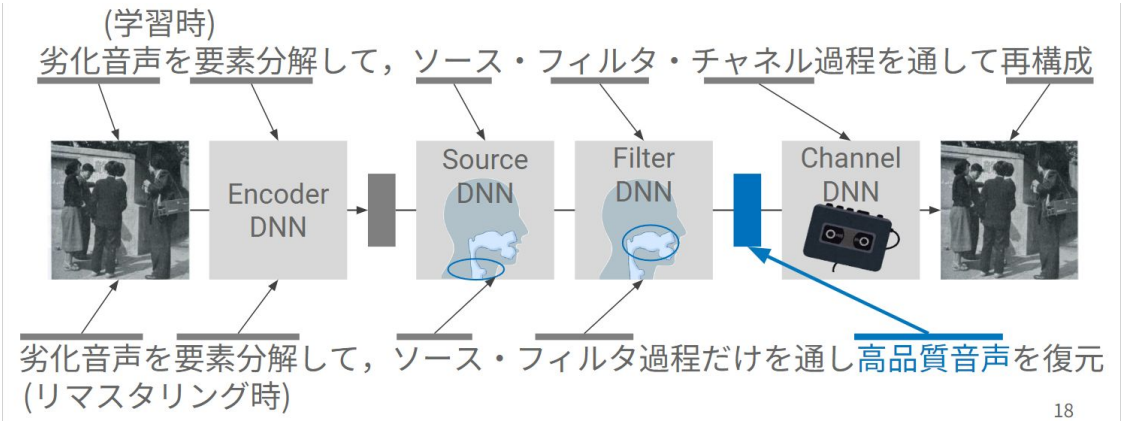
Converts desired speaker's voice (left in demo) without changing other info.

Seki et al., "Spatial Voice Conversion: Voice Conversion Preserving Spatial Information and Non-target Signals," Interspeech, 2024.

研究事例紹介②
音声の文化を守る

Neural speech restoration：音声文化を復元する

ディジタルメディアの前の録音機器は劣化する．廃棄される



音声録音過程を模擬した自己教師あり音声復元．歴史音声だけから現代音質への復元を学習できる

1960~70年代にオープンリールに録音された東北方言昔話をディジタル化 (200時間)

現代品質に近づける

Takaaki Saeki et al., "SelfRemaster: Self-Supervised Speech Restoration for Historical Audio Resources," IEEE Access, 2023.

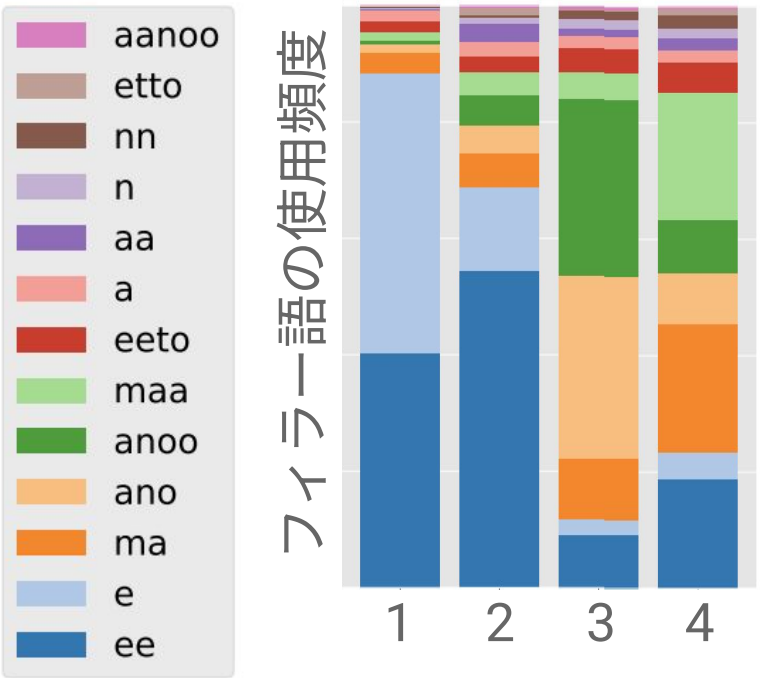
音声文化を復元する→新たな動的コンテンツを作る

音声文化を受動鑑賞する時代を終え，生成 AI と組み合わせた能動鑑賞の時代へ

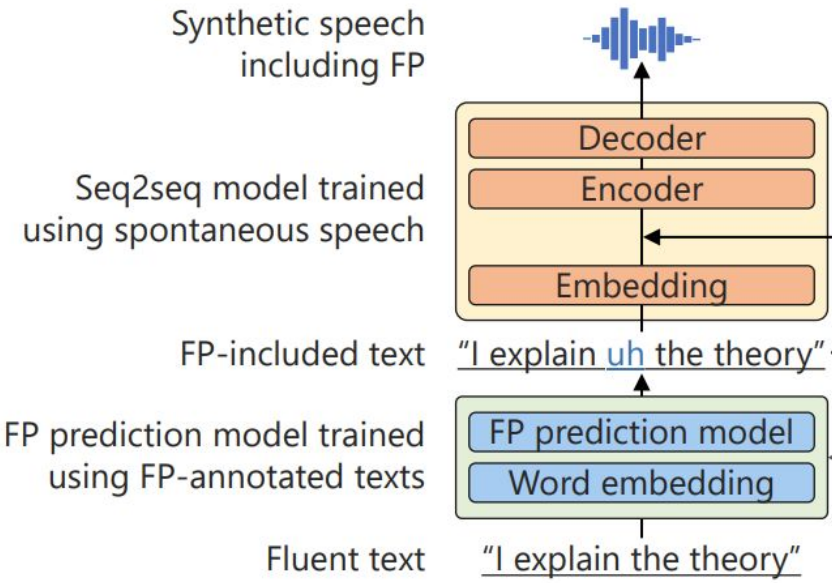
https://sites.google.com/view/ninjal-disciplinary-prj-2023

研究事例紹介③
コンピュータが心をもったときに

自然に間違う音声AI エージェント：「あー」「えー」のフィラーとその個人性



ざっくりいうと，(ァノー)先ほど(ァノー)少し(ァノー)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…

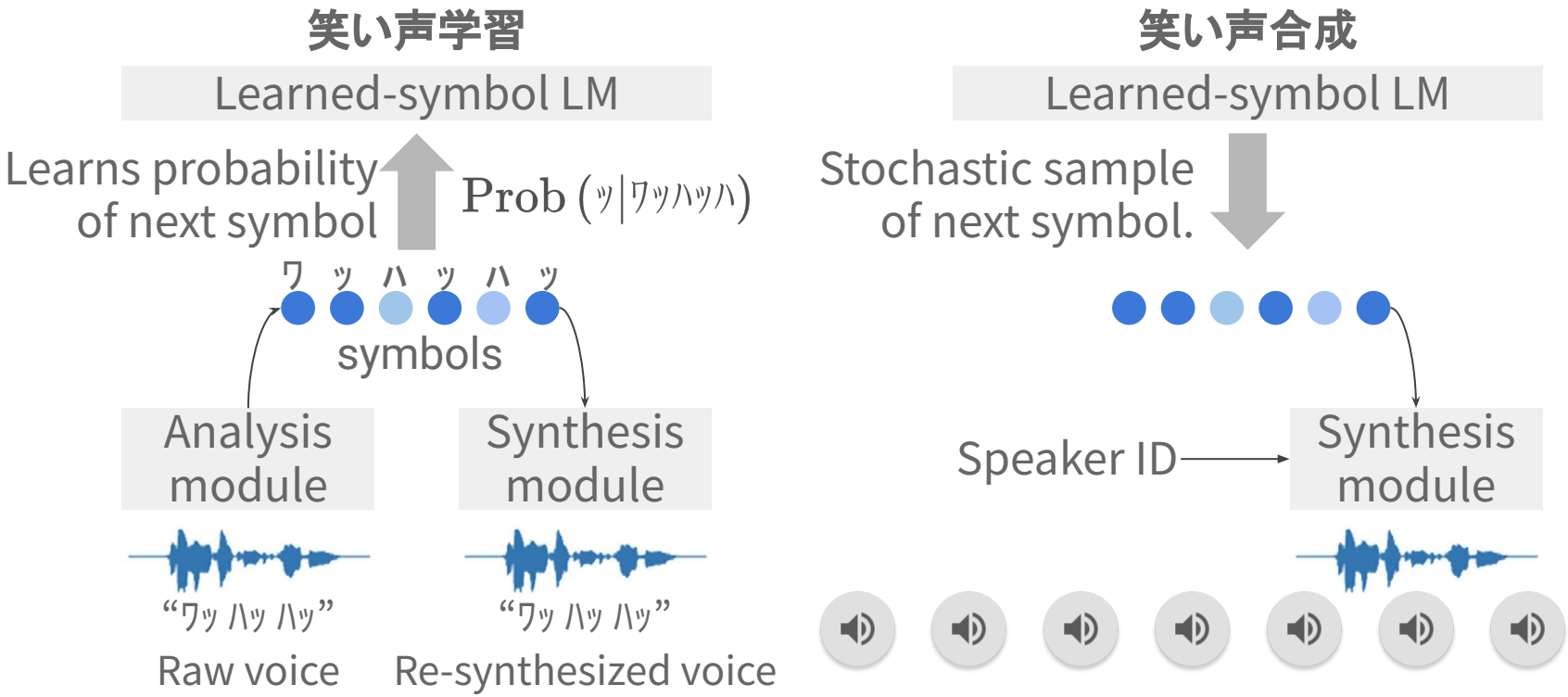


ざっくりいうと，先ほど少しお話ししましたけども，戦後のそういうサブカルチャーのイメージという…

Y. Matsunaga et al., "Improving robustness of spontaneous speech synthesis with linguistic speech regularization and pseudo-filled-pause insertion," Proc. SSW, 2023.

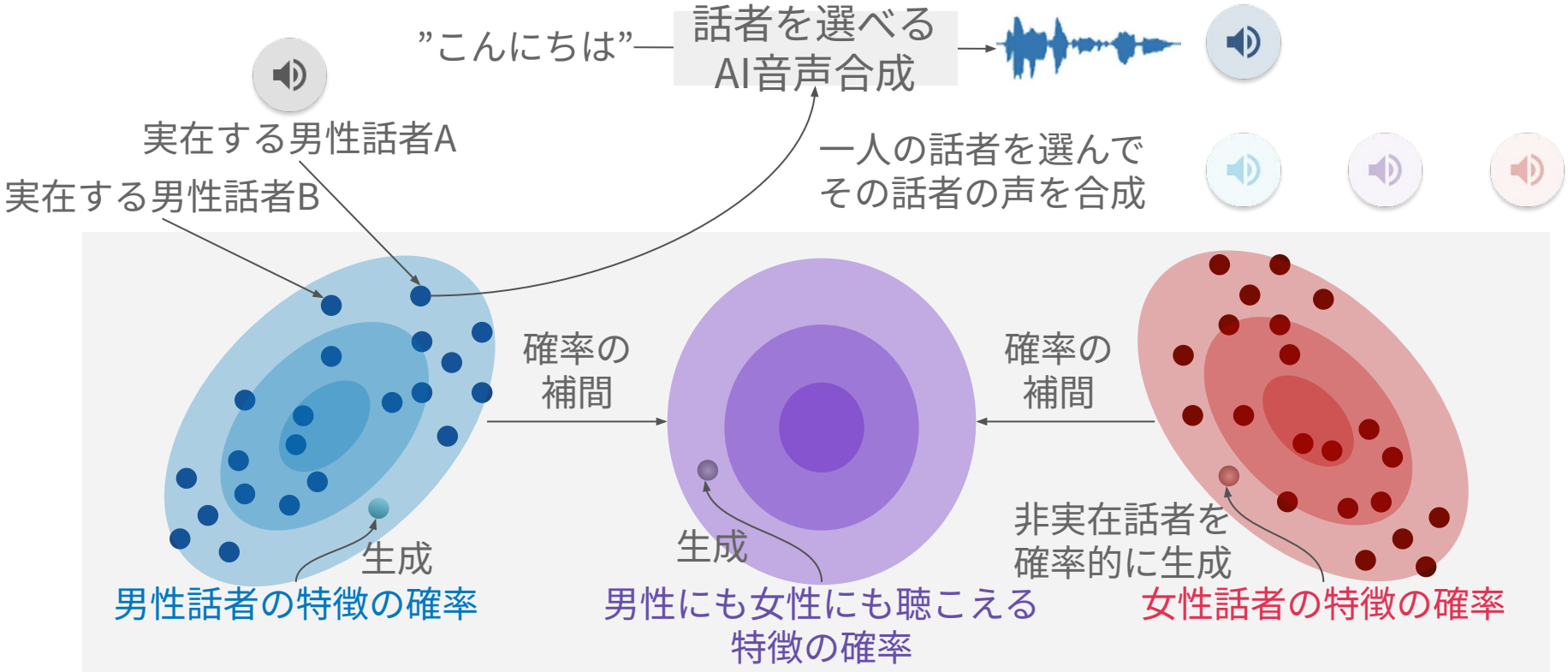
Neural laughter generation：笑い声を合成する

- **コンピュータも笑ってほしい**
  - 笑い声はテキストで書けない．だから，テキストの変わりに「自動で見つけた音の塊」を使う．それを確率モデルで予測．



Detai Xin, et al., "Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus," INTERSPEECH, Dublin, Ireland, Aug. 2023.

Mid-attribute speaker generation：バイナリの属性を超えた非実在話者



Watanabe et al., "Mid-attribute Speaker Generation using Optimal-Transport-based Interpolation of Gaussian Mixture Models," Proc. ICASSP, 2023.

研究事例紹介④
歌の可能性を広げる

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～

“**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”

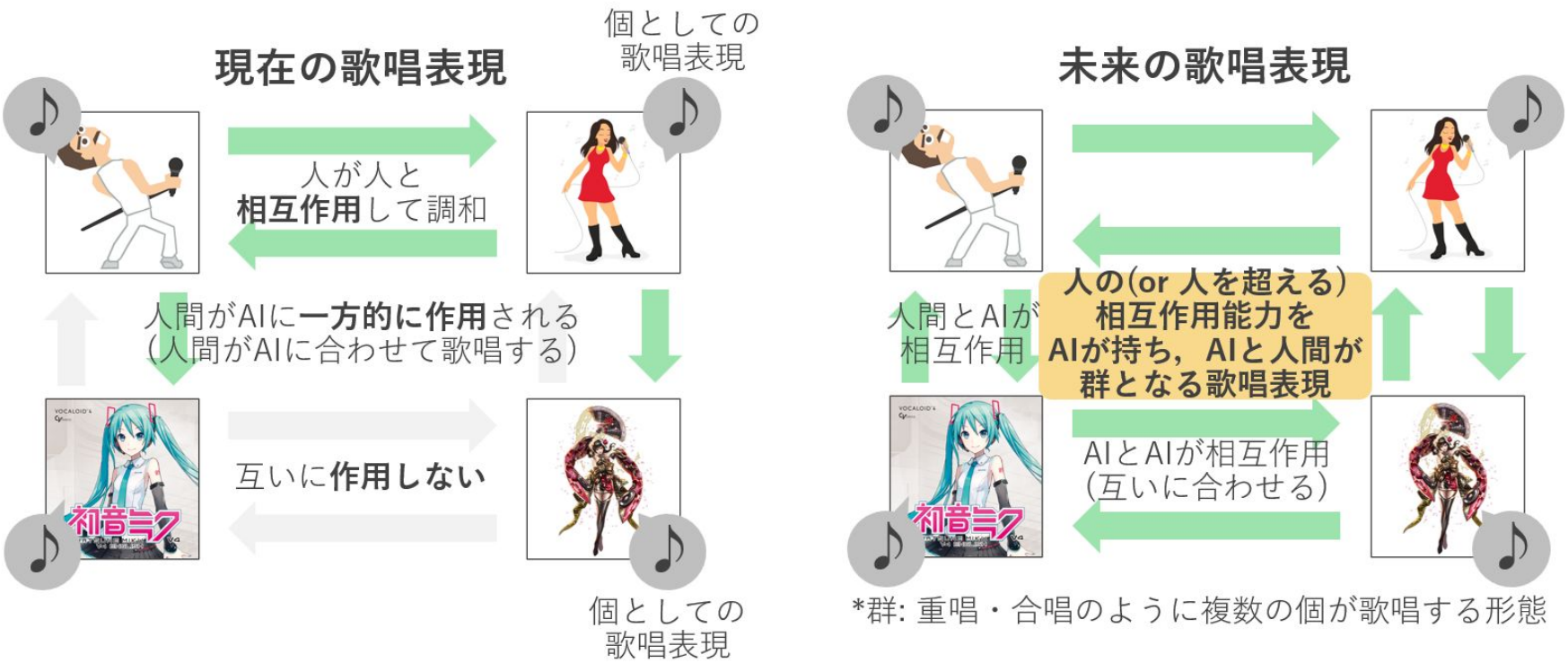
https://www.i.u-tokyo.ac.jp/news/files/ist_pressrelease_20221003_takamichi_s.pdf

歌声合成

https://www.ah-soft.com/synth-v/

jaCappella：
他者と相互作用するAIシンガー& アカペラDB

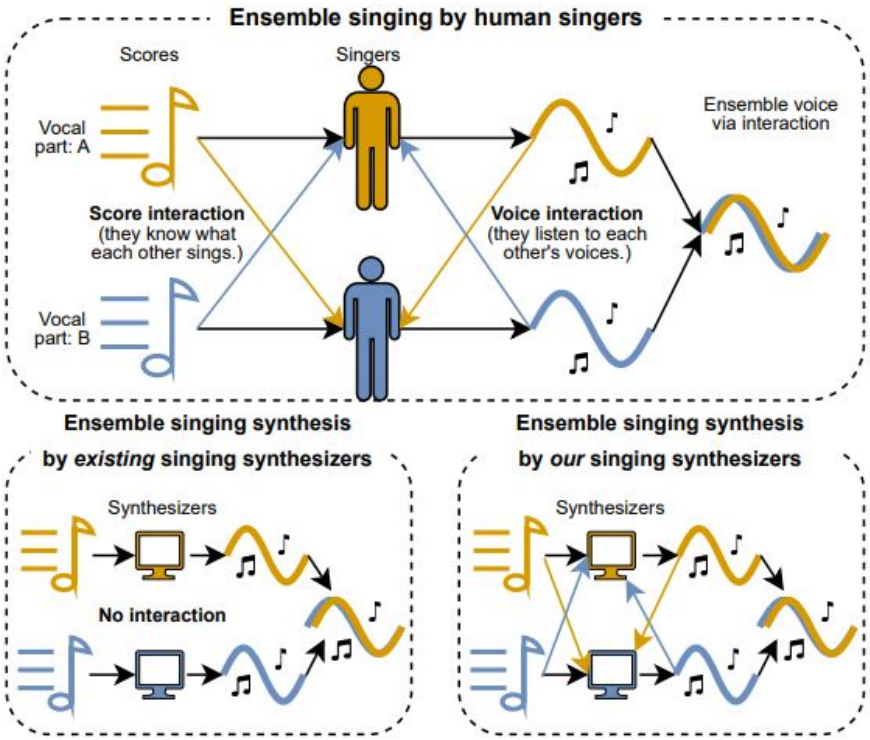
- **人間歌手とAI歌手が相互作用する社会へ**
  - 歌声信号レベルでの相互作用の例がアカペラ
- **相互作用を分析・再現する技術**
  - 相互作用により生まれる一体感とは？
  - 計算機は一体感を生み出せるか？



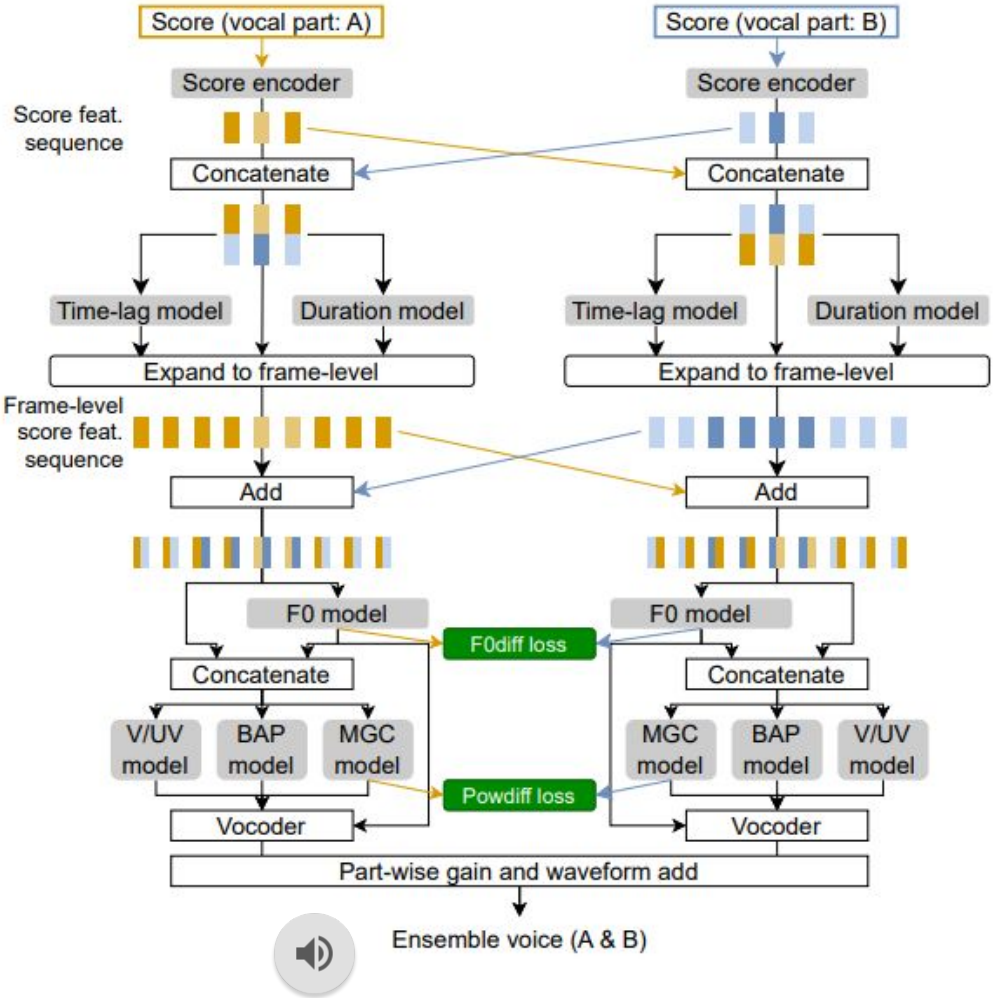
Nakamura: “jaCappella Corpus: A Japanese a Cappella Vocal Ensemble Corpus,” ICASSP, 2023.

歌唱者間相互作用を模擬するアカペラ合成

人間歌唱者同士は，楽譜レベル・
音声レベルの相互作用を起こす



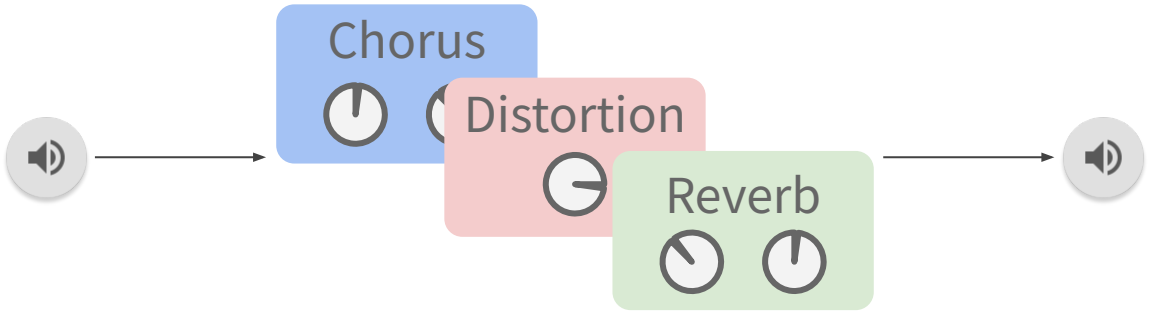
AI歌唱者同士で，楽譜レベル・
音声レベルの相互作用を起こす



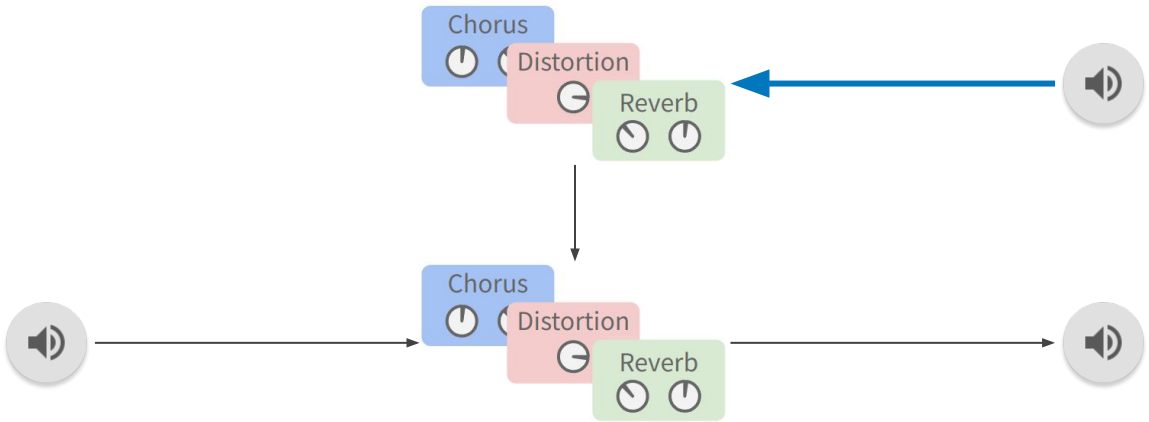
兵藤 他, “歌唱者間相互作用を再現するDNN重唱歌声合成の検討”, MUS研究会, 2024.

Audio effect chain 推定

- Audio effect をかけて音楽制作・ライブの芸術性を高める



- Audio effect chain 推定は、**処理後の信号 (wet signal) から effect を推定**する。その effect を他の音にも適用できる。

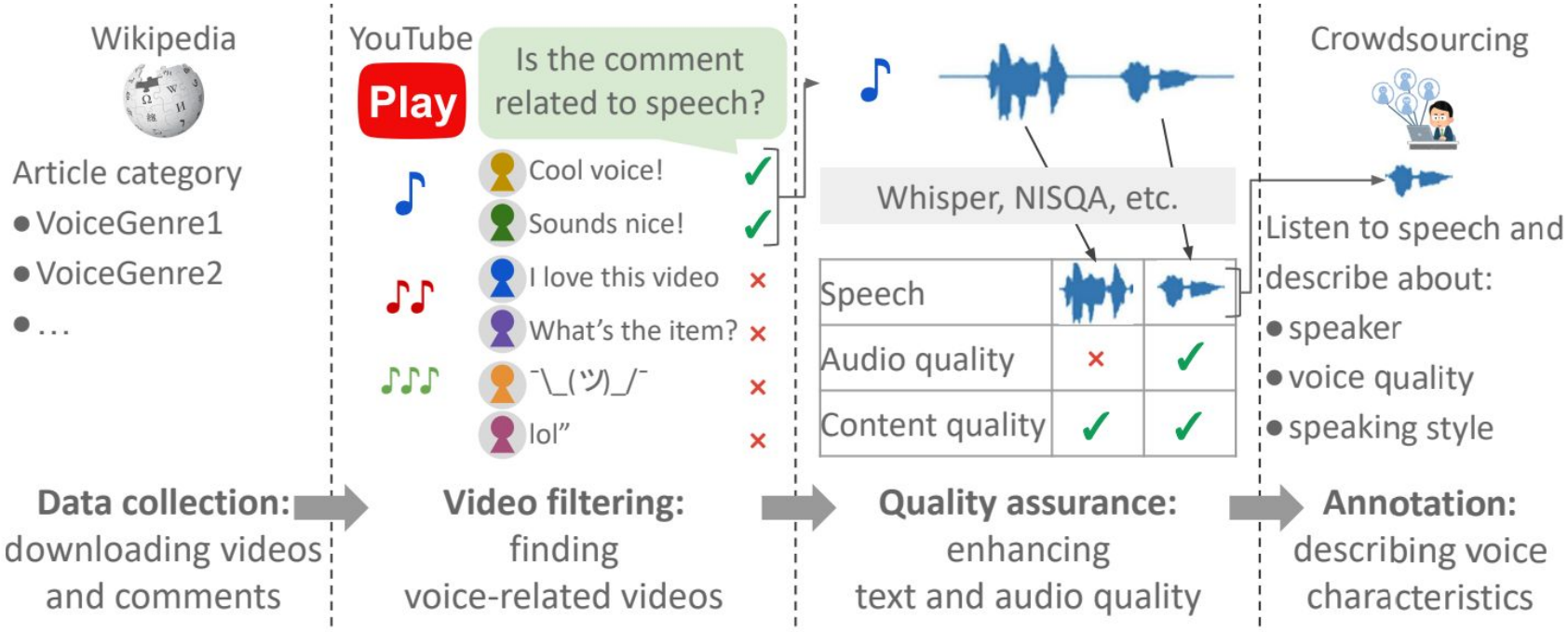


Take et al., “Audio Effect Chain Estimation and Dry Signal Recovery from Multi-Effect-Processed Musical Signals,” DAFx, 2024.

研究事例紹介⑤
音声に関する感性を定量化する

Coco-Nut:
「魅力的な声」データベース

- **動画サービスにおける「魅力的な声」とは？**
  - いろんなユーザが声質について言及している声！
  - → 「魅力的な声」データベースを作成＆声質説明文を付与



	30代くらいの男性の声。ゆっくりと穏やかな話し方でした。苦悩に満ちた、けだるそうな声でした。
	明るい中年の女性がはきはきとした声で楽しそうに喋っている。

Watanabe et al., "COCO-NUT: Corpus of Japanese Utterance and Voice Characteristics Description for Prompt-Based Control," ASRU, 2023.

Cococap-beta：
声質キャップショニング技術

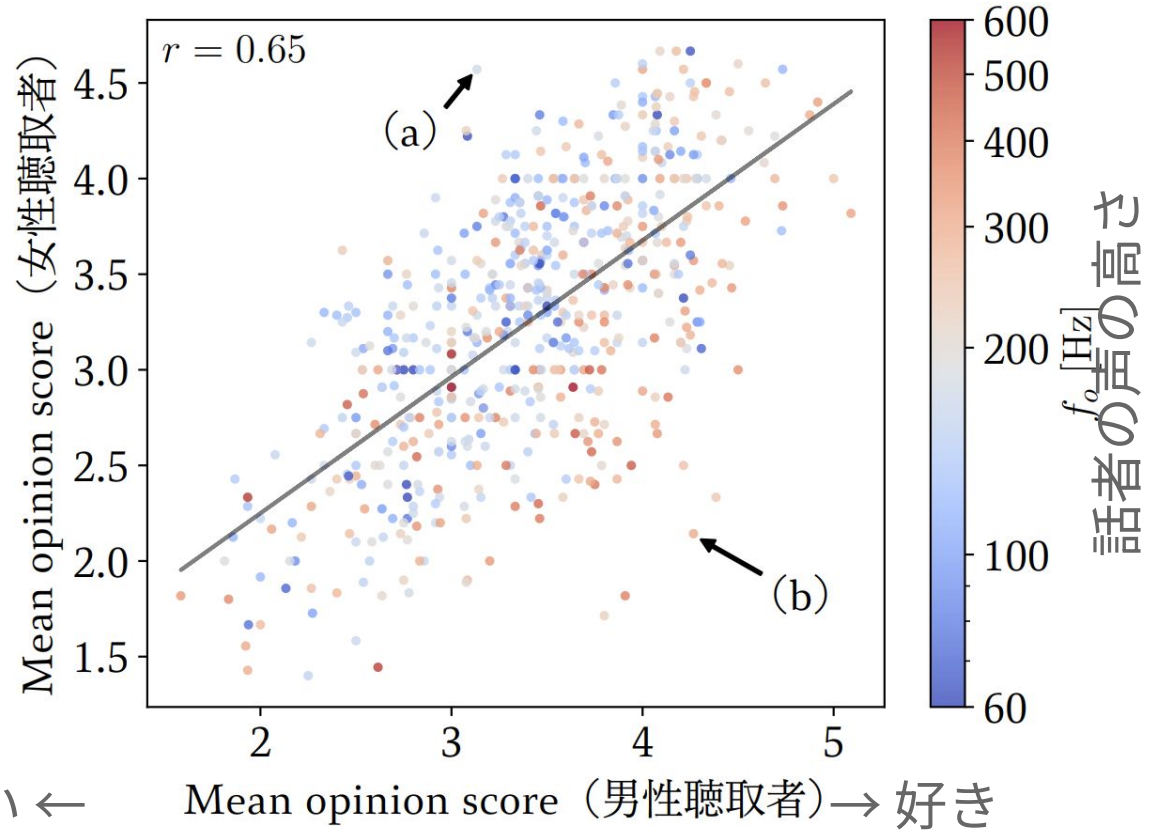
- **Cococap-beta**
  - https://huggingface.co/spaces/sarulab-speech/CoCoCap-beta

- **これ以外にも，声質表現文と音声の相互変換が可能になる！**
  - 声質表現文 → 音声 (自然言語による音声検索)
  - 声質表現文 → TTSシステム (自然言語による音声AIデザイン)
  - それ以外にも：感情キャプショニング

https://huggingface.co/spaces/sarulab-speech/CoCoCap-beta

CocoNut-Humoresque：「魅力的な声」データベースと嗜好分析

各個人にとって「魅力的な声」ってなんだろう？
1800話者の声に対する900聴取者の「この声すきスコア」を集めてみた



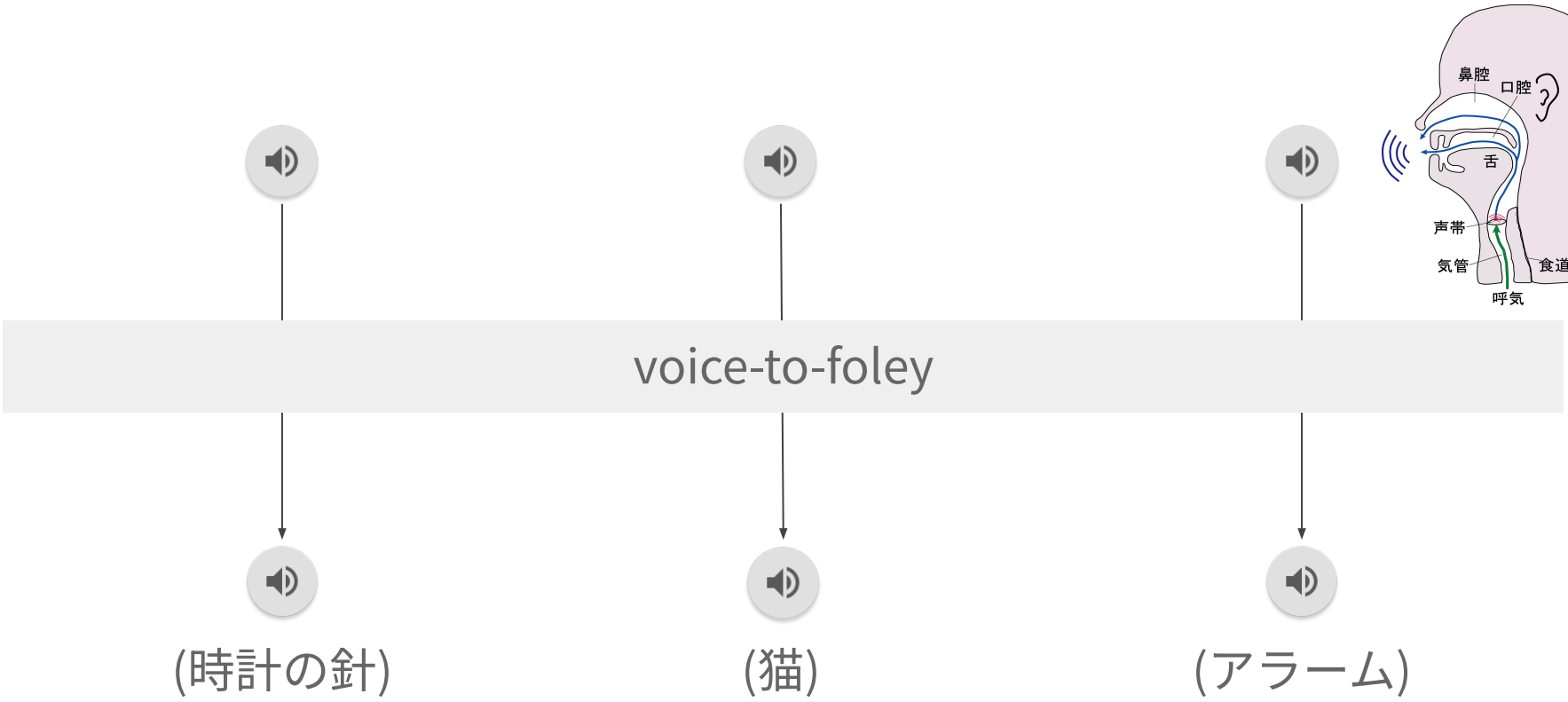
魅力は男女間で相関する傾向．一方で片方の性別のみに好かれる場合も．
(a) 「若い男性が、はきはきした低い声で怒ったように喋っている」
(b) 「10 代の少女が、かわいらしい声でまったりとした口調で喋っている」

Suda et al., “Who Finds This Voice Attractive? A Large-Scale Experiment Using In-the-Wild Data,” Interspeech, 2024.

研究事例紹介⑥
環境音を人工的に作り出す

voice-to-foley：
音声模倣からの環境音合成

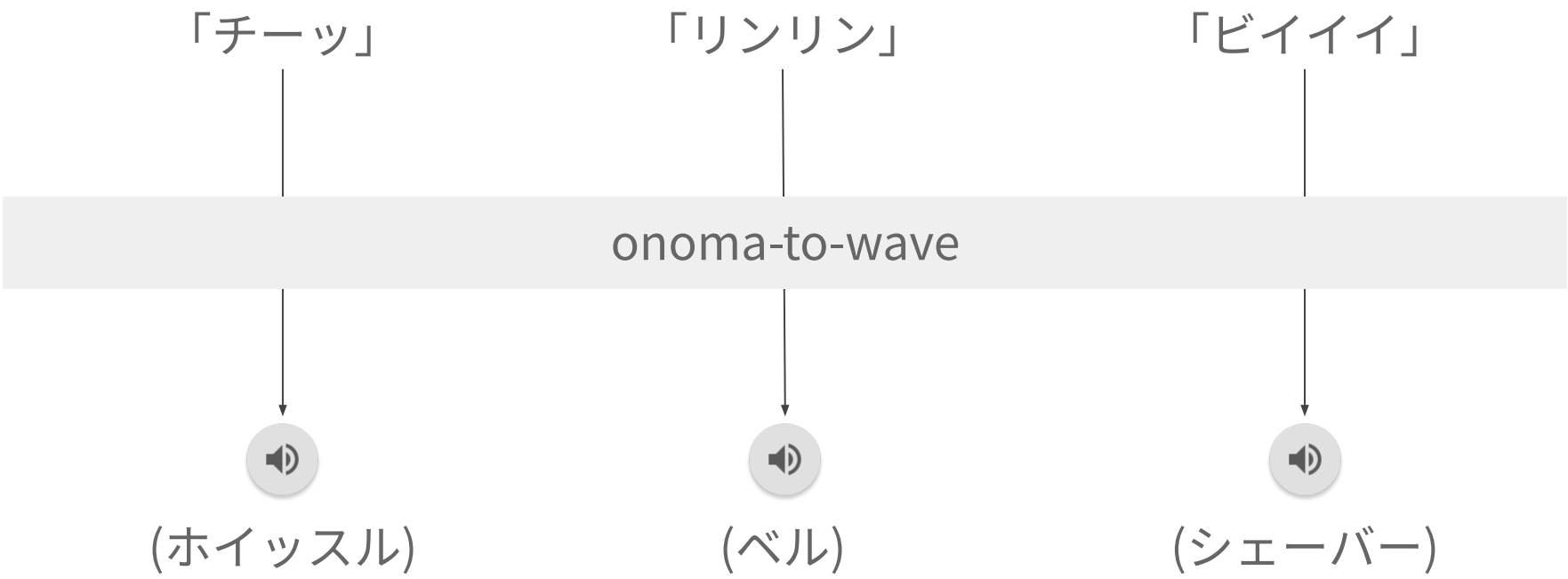
- **音声模倣：非音声を声まねしたもの**
  - 直感的に環境音を合成できる方法
  - → 音声模倣から環境音を合成できるのでは？



Okamoto et al., “Environmental Sound Synthesis from Vocal Imitations and Sound Event Labels,” ICASSP, 2024.

onoma-to-wave：
オノマトペからの環境音合成

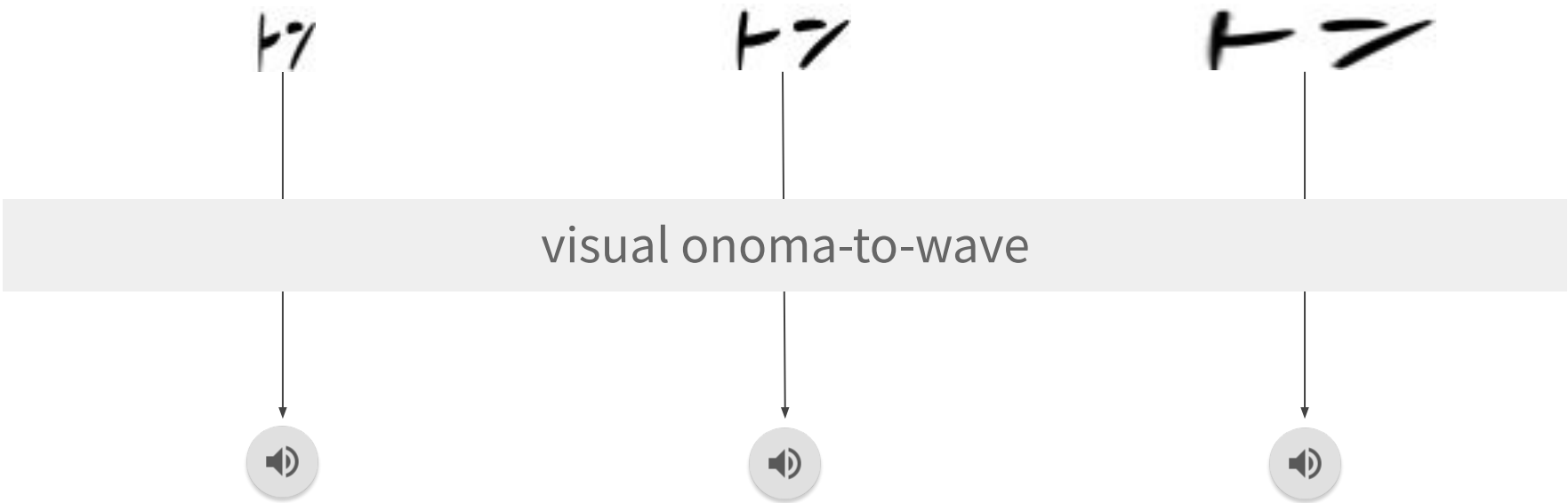
- **オノマトペ：音を文字で模倣したもの**
  - 環境音の時間構造を文字で表してくれる．
  - → オノマトペ文字から環境音を合成できるのでは？



Okamoto et al., "Onoma-to-wave: Environmental sound synthesis from onomatopoeic words," APSIPA transactions, 2022.

visual onoma-to-wave：
オノマトペ画像からの音声合成

- **画像オノマトペ：漫画などで用いられる音表現**
  - オノマトペだけでなく，画像エフェクトでも情感を伝える
  - → オノマトペ画像から環境音を合成できるのでは？



Ohnaka et al., "Visual onoma-to-wave: environmental sound synthesis from visual onomatopoeias and sound-source images", ICASSP, 2023.

まとめ

まとめ

研究事例紹介①：人間の音声表現を拡張する

研究事例紹介②：音声の文化を守る

研究事例紹介③：コンピュータが心をもったときに

研究事例紹介④：歌の可能性を広げる

研究事例紹介⑤：音声に関する感性を定量化する

研究事例紹介⑥：環境音を人工的に作り出す

音声合成・歌声合成 = 音を正確に再現する、だけではありません。
価値観をどう更新できるか、人間をどう幸せにできるかを考えてください。