NL研究会 招待講演

基盤モデル時代に言語で音声を処理したい

高道 慎之介（慶應義塾大学／東京大学）

Takamichi Laboratory
慶應義塾大学 高道研究室

自己紹介

@forthshinji

名前

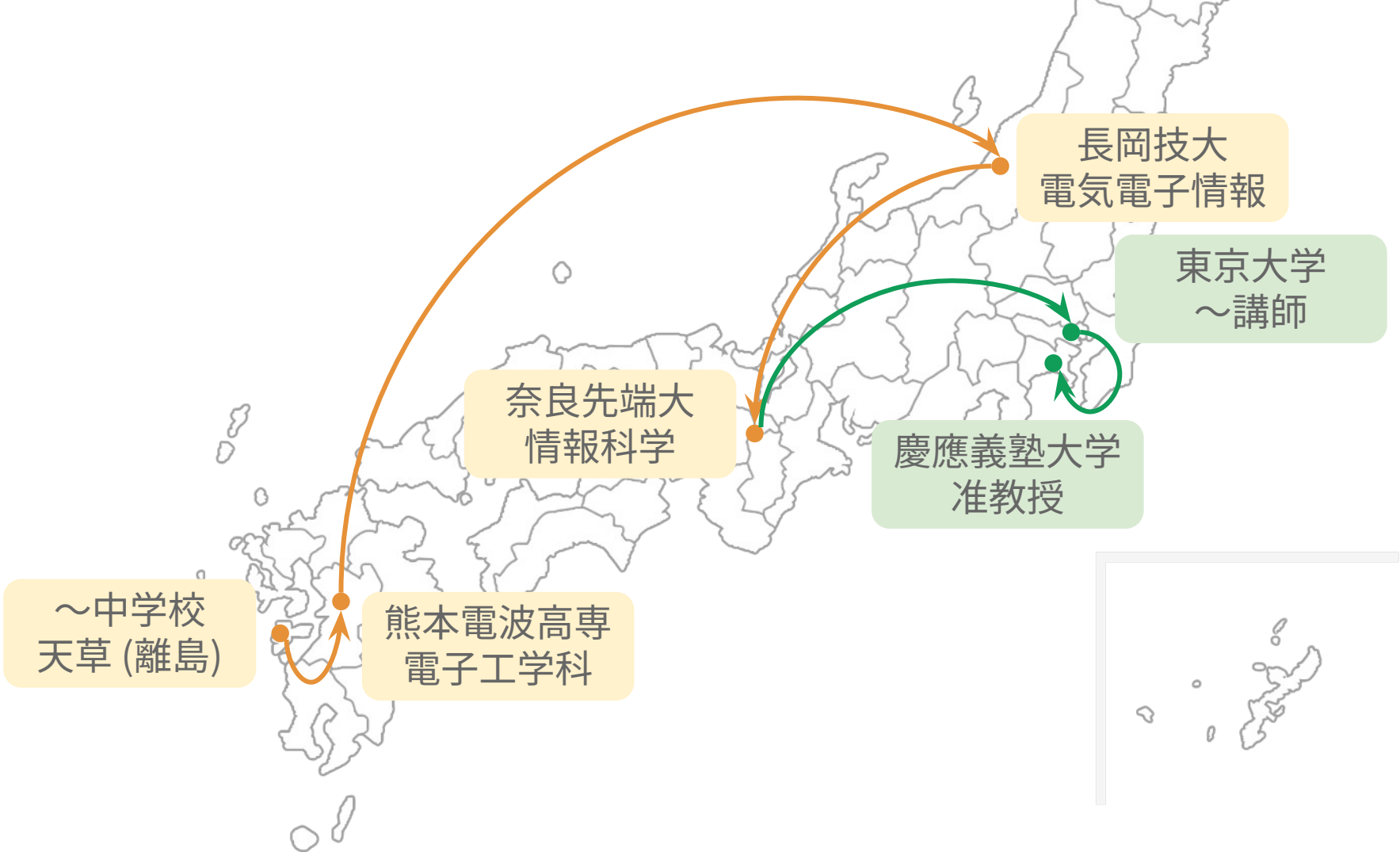
高道 慎之介 (たかみち しんのすけ)

現職

慶應義塾大学 准教授／東京大学 特任准教授

専門

音声工学、機械学習、信号処理


長岡技大
電気電子情報
東京大学
～講師
奈良先端大
情報科学
慶應義塾大学
准教授
～中学校
天草 (離島)
熊本電波高専
電子工学科

慶應義塾大学 高道研究室 (2024~)

音と言葉を研究する

Takamichi (高道) Laboratory, which studies audio (音) and language (言)

高道研究室

Takamichi Lab. / 高道研究室

Official website of Takamichi laboratory / 高道研究室 公式ページ

The Takamichi Laboratory, established in 2024, is dedicated to the science and technology of audio and language. Our laboratory focuses on the theory, machine learning, data resources, and human interaction concerning audio, speech, and music signals, as well as spoken language information. What principles underlie audio and language? What should the future of audio and language be like? We conduct research to answer these questions.

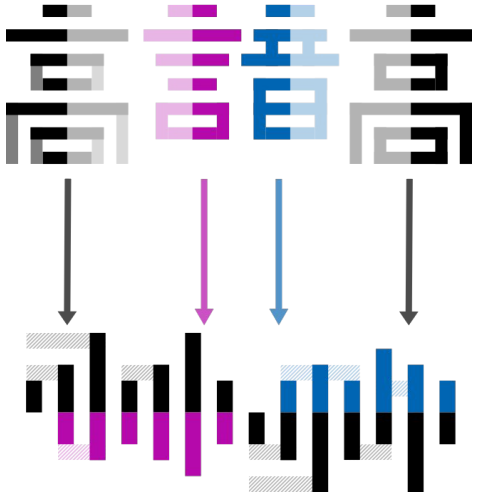
高道研究室は，2024年に設立された，音と言葉の科学技術を扱う研究室です．本研究室では，音声音響音楽信号と音声言語情報を中心として，その理論，機械学習，データ資源，人間との相互作用を扱います．音と言葉はなんの理に基づいているのでしょうか？ 音と言葉について未来はどうあるべきでしょうか？ これらに答えるために研究をしています．

https://takamichi-lab.github.io/

Takamichi Laboratory
慶應義塾大学 高道研究室

About Logo / ロゴについて

Takamichi Laboratory
慶應義塾大学 高道研究室



本講演の内容 (あとで個人HPでスライドを公開します)

NLPっぽく音声を処理できる時代が来そう。どうする？

開催中の interspeech 2024 における text ⇔ speech のサーベイを多く含みます.
世界一早い interspeech サーベイ発表 (だったら良いな)！

その目的1：自然言語処理と音声処理の境界を熱く！

(5) 大規模言語モデルの音声タスクへの応用と分析

水本 智也 (SB Intuitions株式会社), 山崎 天 (SB Intuitions株式会社), 李 凌寒 (SB Intuitions株式会社), 吉川 克正 (SB Intuitions株式会社)

https://sites.google.com/sig-nl.ipsj.or.jp/sig-nl/%E7%A0%94%E7%A9%B6%E7%99%BA%E8%A1%A8%E4%BC%9A/%E7%AC%AC260%E5%9B%9E

特別セッション「音声と言語の分野横断」

本特別セッションでは，音声と言語の両分野にまたがって，単一のモーダルに留まらない情報を扱った研究発表を広く募集します．音声認識，音声合成，音声対話，音声翻訳，音声と言語の双方を含むコーパスの構築や分析等，音声と言語の両分野をカバーする研究を分野に限らず幅広く募集します．また，音声や言語に留まらず，画像や映像等のモーダルを扱った研究も合わせて募集します．

https://sites.google.com/sig-nl.ipsj.or.jp/sig-nl/%E7%A0%94%E7%A9%B6%E7%99%BA%E8%A1%A8%E4%BC%9A/NL258

漸進的な音声分割を用いたストリーミング同時音声翻訳

○福田りょう, 須藤克仁, 中村哲 (NAIST)

タグ付き混合データ学習と自己教師あり学習による同時通訳データを用いたEnd-to-End同時音声翻訳

○胡尤佳, 福田りょう, 西川勇太, 加納保昌, 須藤克仁, 中村哲 (NAIST)

https://www.anlp.jp/proceedings/annual_meeting/2024/

正書法および音韻の複雑さによる音声認識の精度への影響

○田口智大 (ノートルダム大)

ラベル付き系列予測による音声シグナルの Textless 依存構造解析

○神藤駿介, 宮尾祐介 (東大)

SlideAVSR: 視聴覚音声認識のための論文解説動画データセット

○王昊 (早大), 栗田修平 (理研), 清水周一郎 (京大), 河原大輔 (早大)

Creating Heterogenous Transcription of English and Japanese on a Multilingual Audio File

○◊Yuika Sun (Los Altos High School)

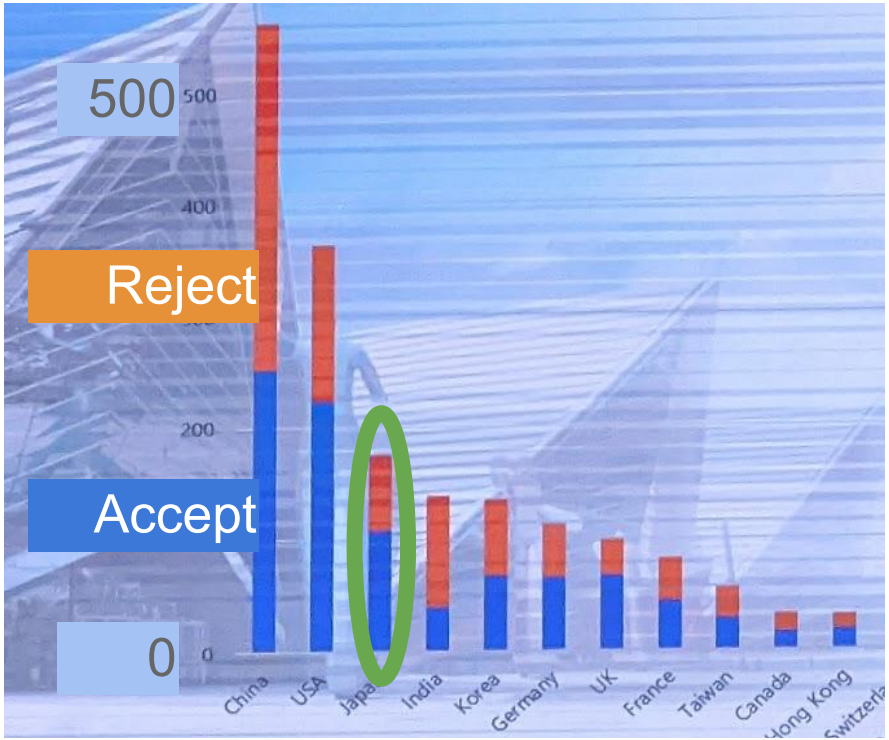
環境音に対する日本語自由記述文コーパスとベンチマーク分析

○岡本悠希 (立命館大), 高道慎之介, 森松亜衣, 渡邊亞椰 (東大), 井本桂右 (同志社大), 山下洋一 (立命館大)

https://www.anlp.jp/proceedings/annual_meeting/2024/

その目的2：音声界隈の日本のプレゼンスを保つ！

国別の投稿件数 (interspeech2022)



中国 > アメリカ > **日本** > インド > 韓国

著者の発表件数 (interspeech2024)

Watanabe, Shinji	26 ★
Xie, Lei	15
Chang, Joon-Hyuk	13
Meng, Helen	11
Yamagishi, Junichi	9 ★
Busso, Carlos	8
Lee, Chi-Chun	8
Toda, Tomoki	8 ★
Wang, Dong	8
Ginsburg, Boris	7
Noeth, Elmar	7
Qian, Yanmin	7
Takamichi, Shinnosuke	7 ★

トップ10位中4人が日本関係者

音声情報処理

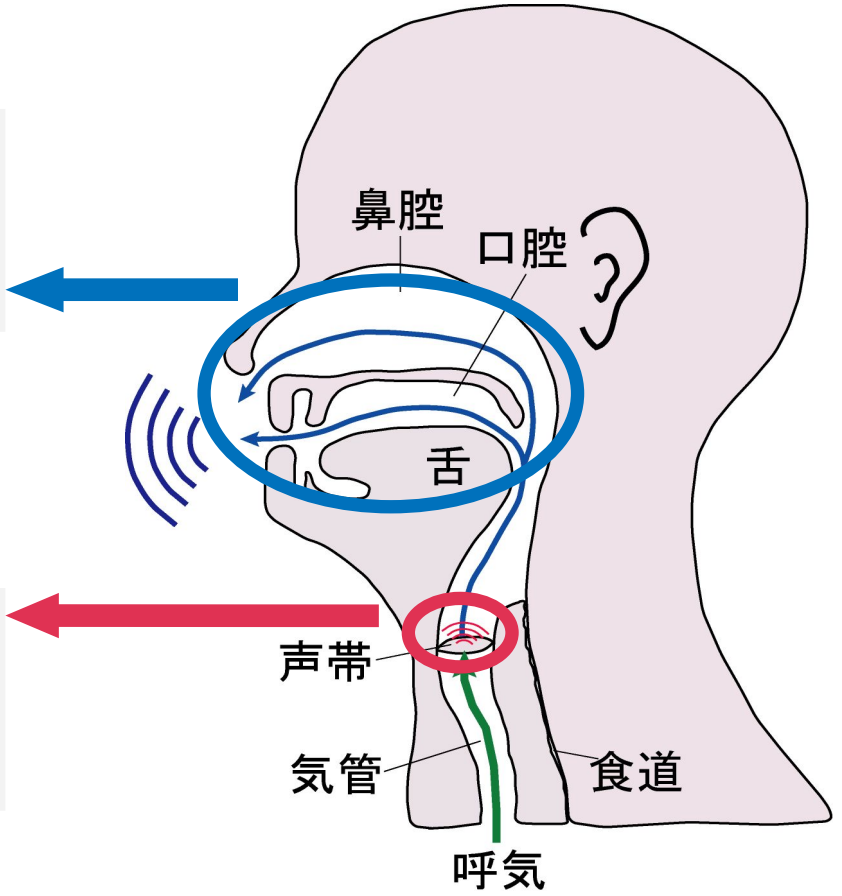
発声器官のモデル: ソース・フィルタモデル

声道 (フィルタ)
口や喉を動かして
音源信号を音色付け

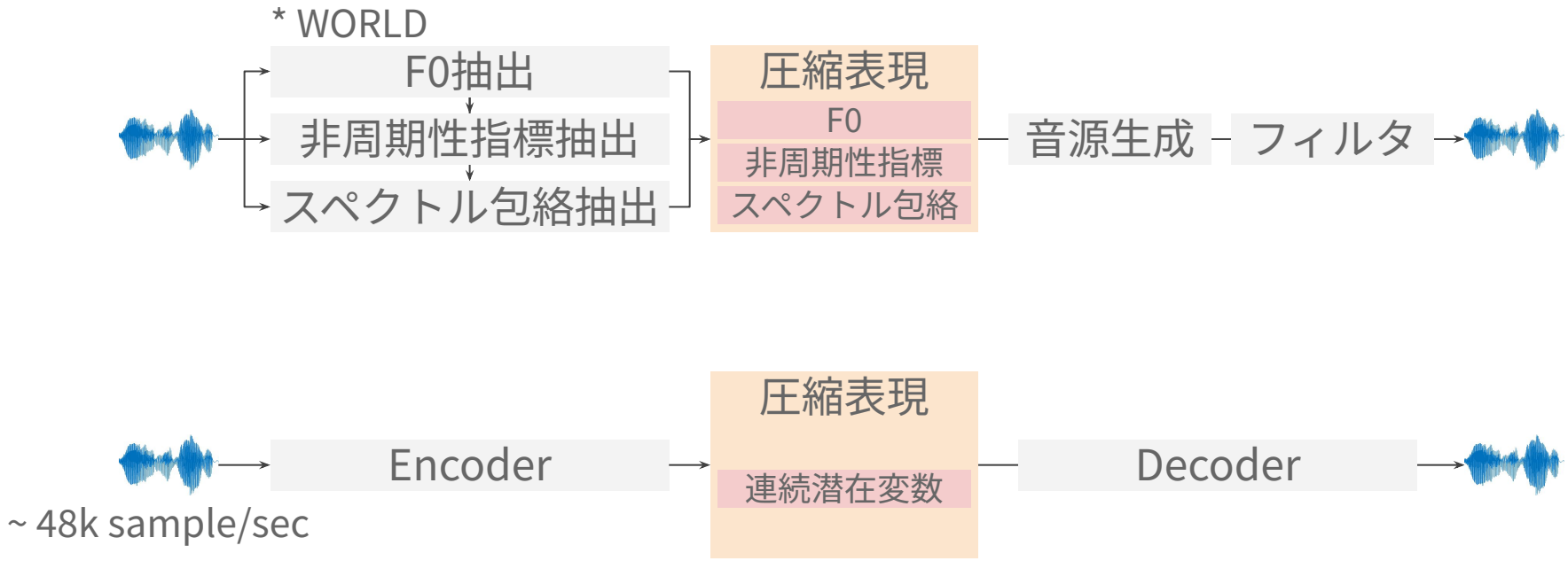
音声波形

畳込み

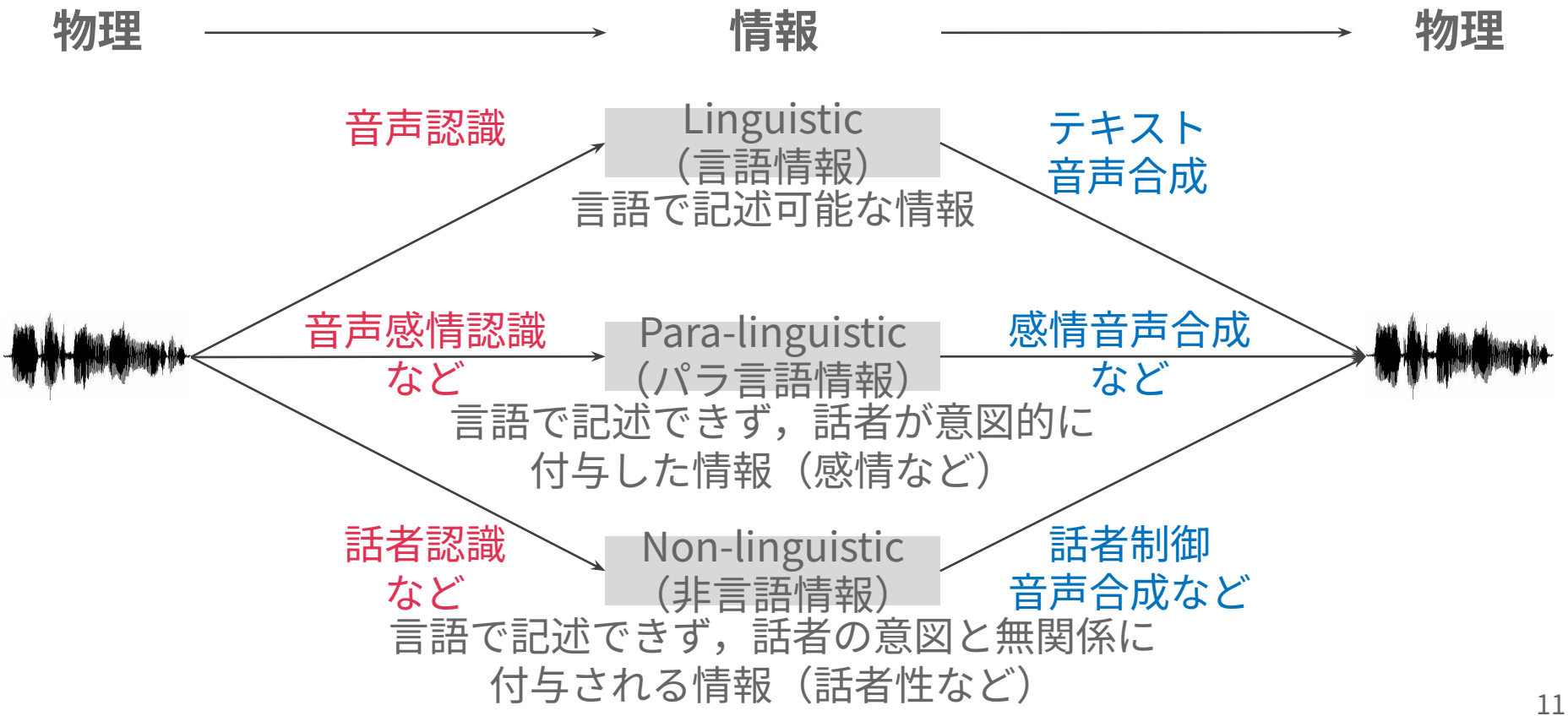
音源 (ソース)
肺からの呼気を
声帯の開閉で制御



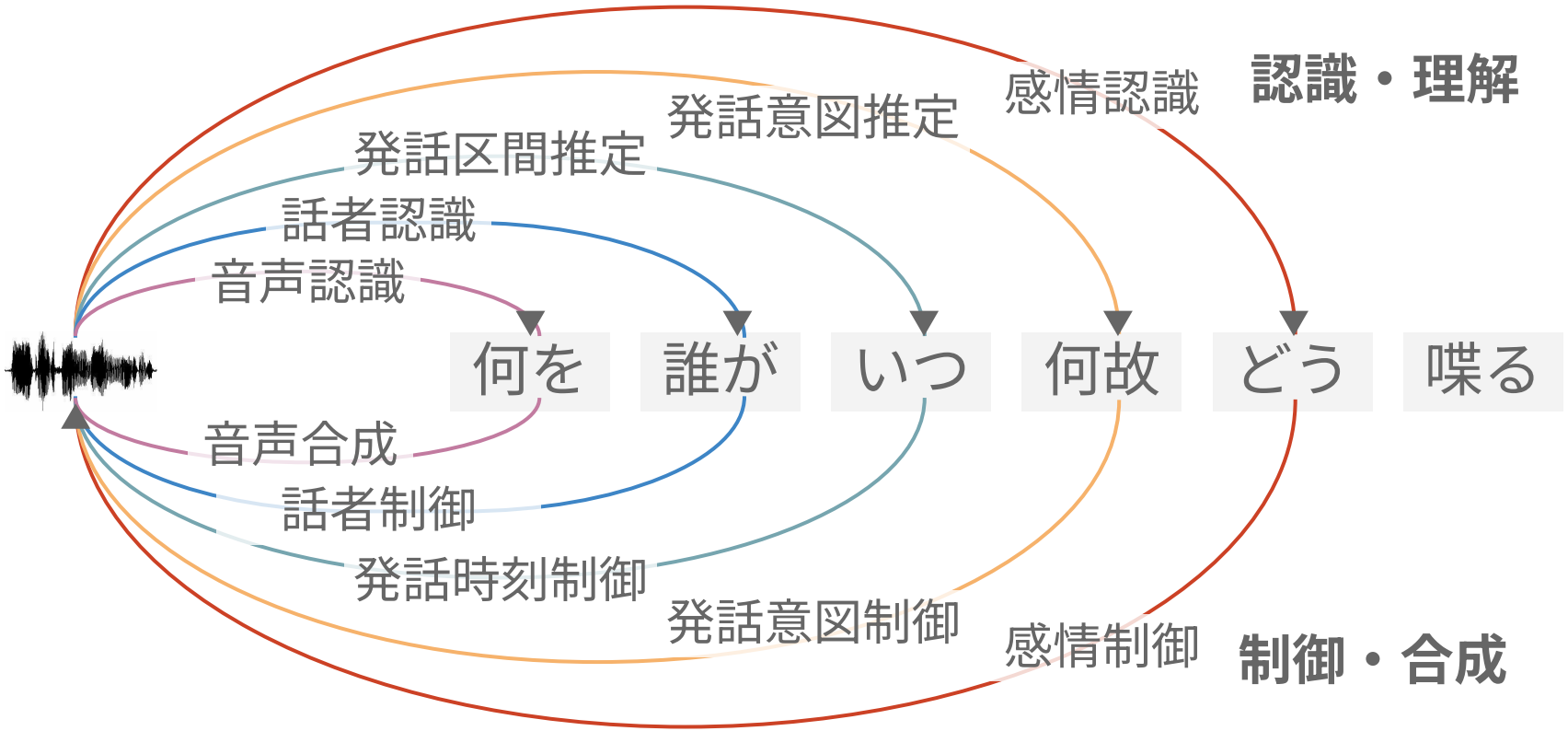
音声のパラメータ表現


* WORLD
F0抽出
非周期性指標抽出
スペクトル包絡抽出
圧縮表現
F0
非周期性指標
スペクトル包絡
音源生成
フィルタ
Encoder
圧縮表現
連続潜在変数
Decoder
~ 48k sample/sec

音声のもつ情報とそれを扱う主な音声技術



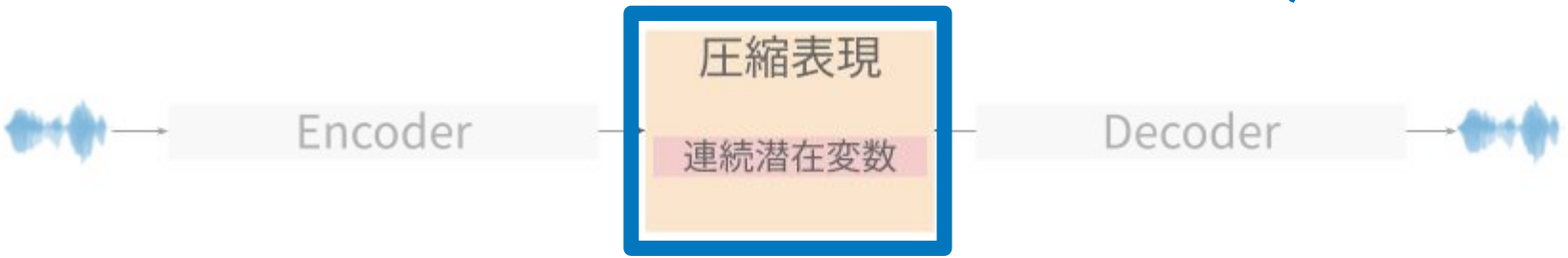
種々の音声処理タスク


認識・理解
感情認識
発話意図推定
発話区間推定
話者認識
音声認識
何を
誰が
いつ
何故
どう
喋る
音声合成
話者制御
発話時刻制御
発話意図制御
感情制御
制御・合成

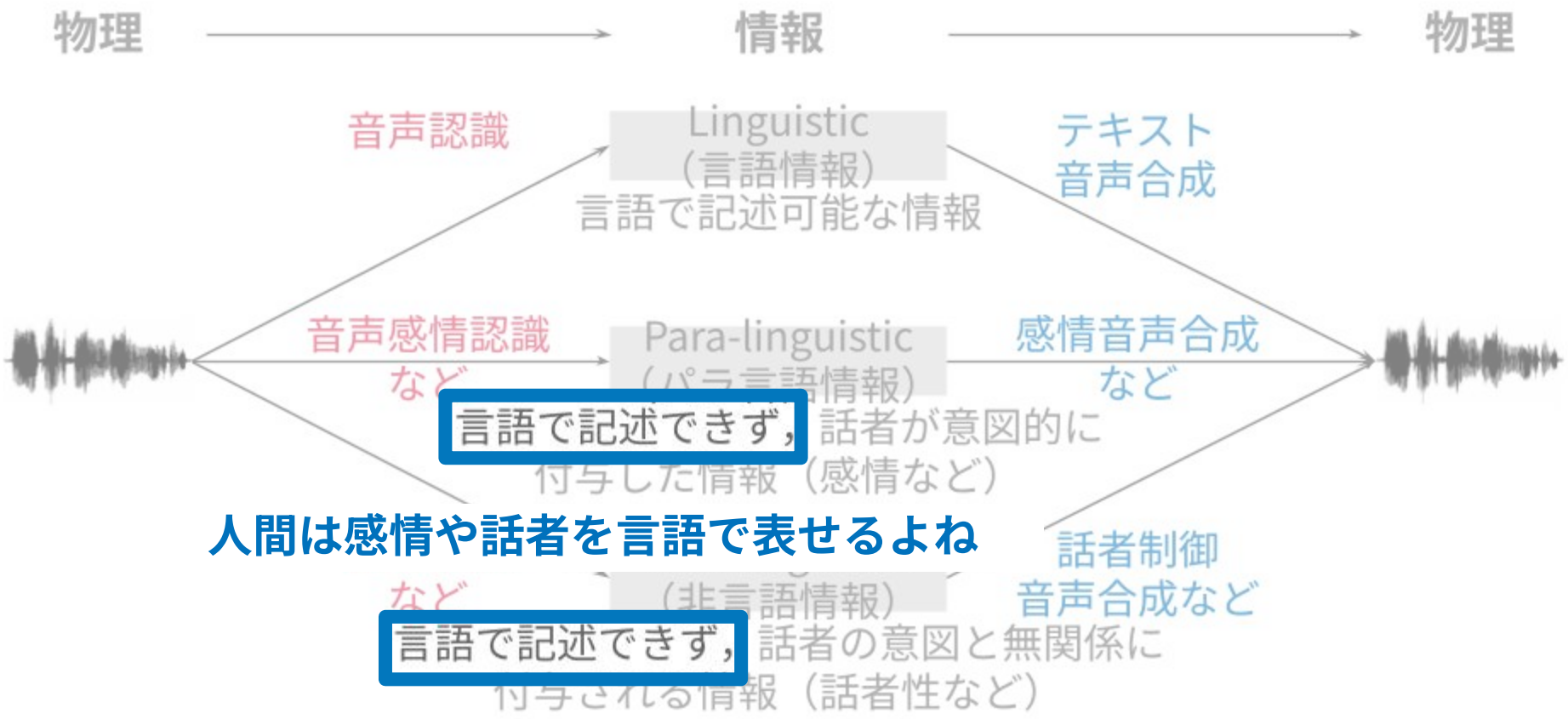
と思っていた．これまでは…

離散表現できるのでは？

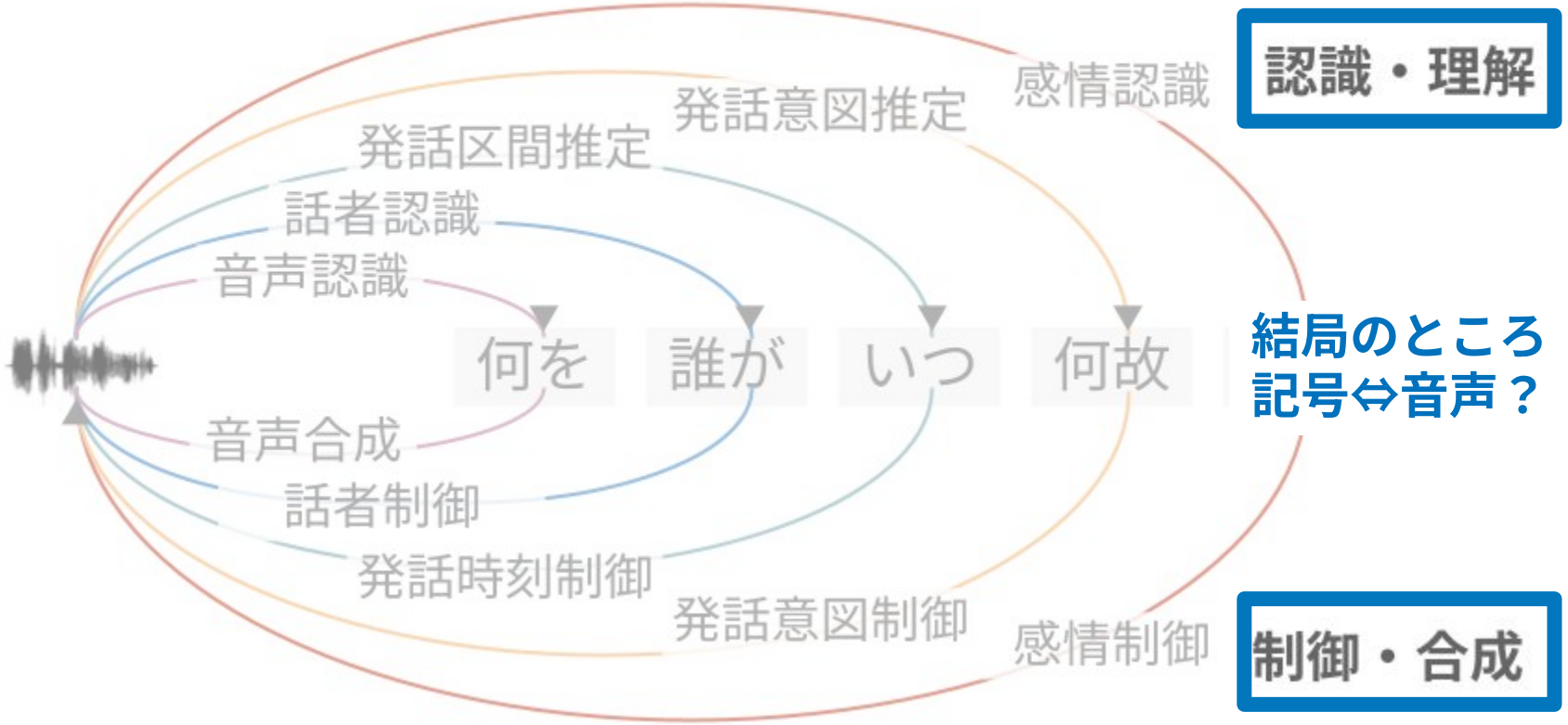
離散表現でもよさそう (音素はそもそも離散的だし
韻律を離散的に扱っても聴感上問題なさそう)



自然言語で表せるのでは？



いろいろLLMでかけるのでは？

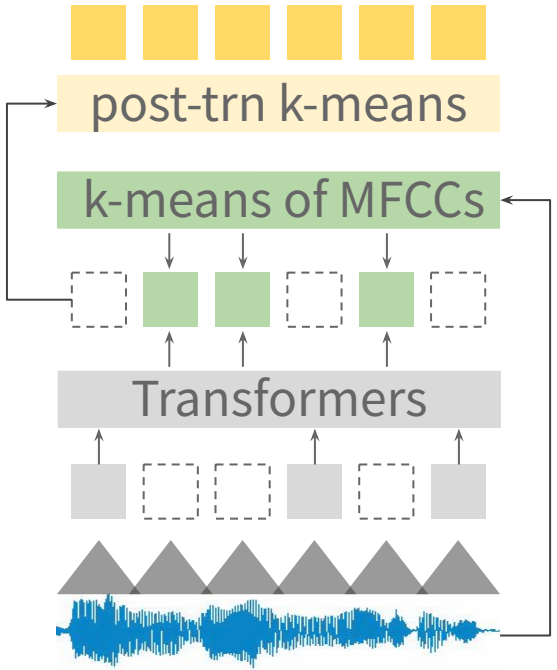


NLPに関連した，昨今の音声処理技術

音声の離散表現の基本的な考え方

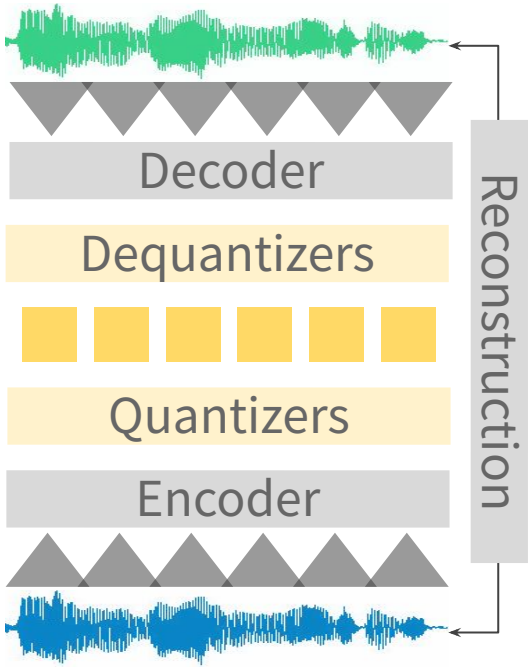
MLM＋k-means

[Lakhotia21]



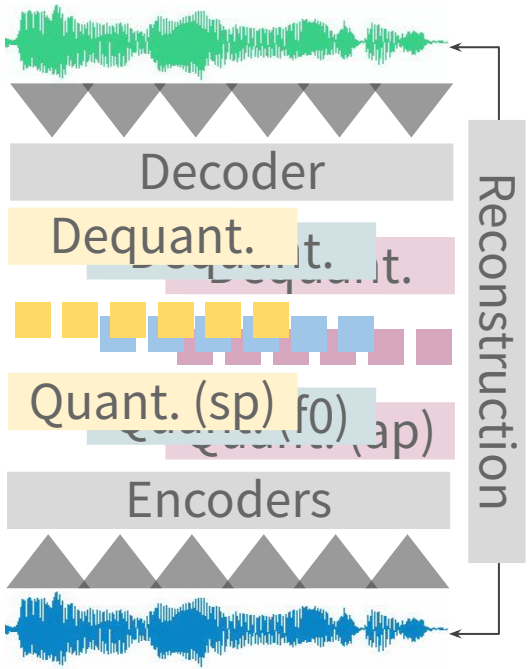
量子化＋再構成

[Zeghidour21][Défossez24][Kumar23]
亜種として[Bai24], MLM併用[Zhang24]



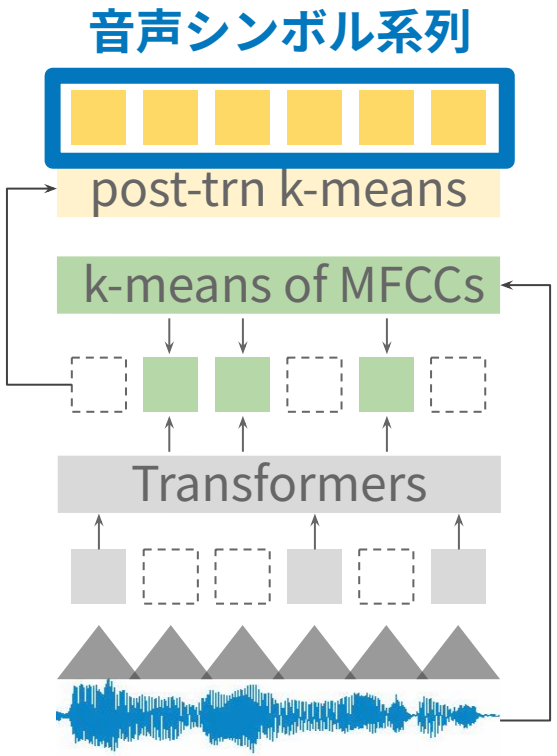
量子化＋分解＋再構成

[Ju24]



[Lakhotia21] https://arxiv.org/abs/2102.01192 (TACL21), [Bai24] https://arxiv.org/abs/2407.15835 (only arxiv), [Ju24] https://arxiv.org/pdf/2403.03100 (only arxiv), [Zhang24] https://arxiv.org/pdf/2308.16692 (ICLR24), [Defossez24] https://arxiv.org/abs/2210.13438 (ICLR24), [Kumar23] https://arxiv.org/pdf/2306.06546 (NeurIPS23), [Zeghidour21] https://arxiv.org/abs/2107.03312 (TASLP21)

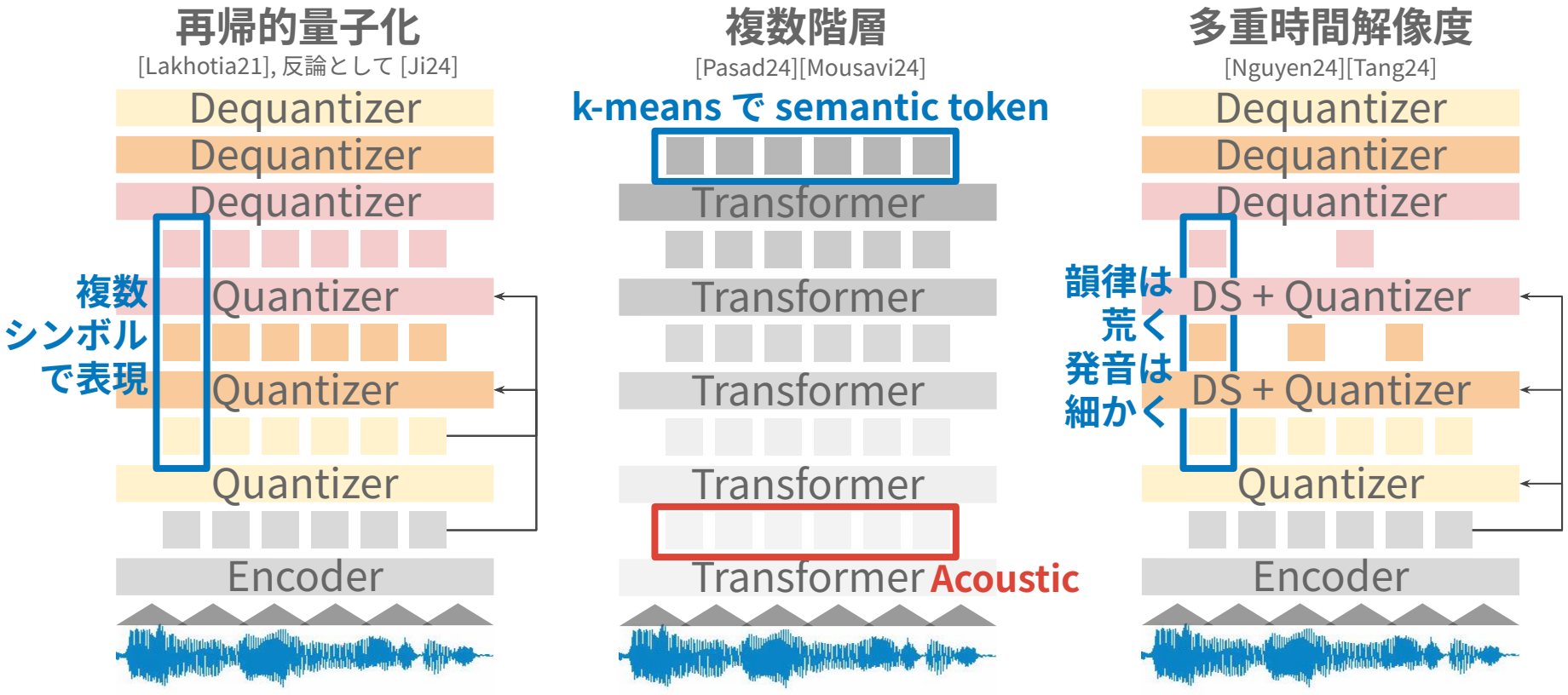
音声シンボルに関する調査



- Subword 化 (byte pair encoding)
  - 音声合成の高品質化＆高速化 [Shen24][Li24]
  - 音声翻訳のモデルサイズ削減 [Lam24]
  - この単位が何を表すのかは未調査？
- Zipf 則
  - Zipf 則ではなくべき乗則に従い，日本語(表意)と英語(表音)で異なる分布 [Takamichi24]
- その他？
  - 多言語化・言語依存性などは未調査？

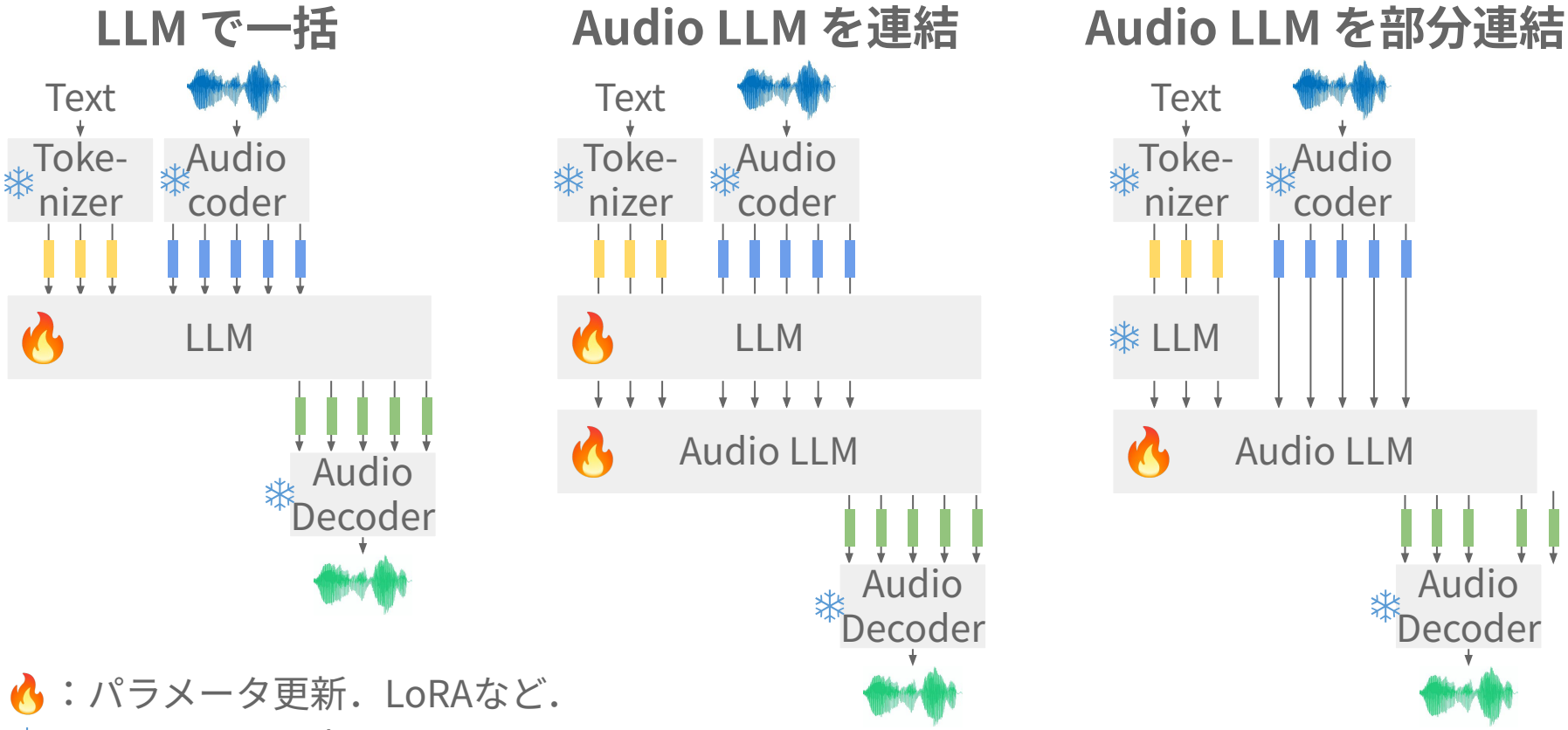
[Shen24] https://arxiv.org/pdf/2310.14580v4 (ICASSP24), [Li24] https://arxiv.org/pdf/2407.03892v1 (ICASSP24), [Lam24] https://aclanthology.org/2024.iwslt-1.16.pdf (IWSLT24)
[Takamichi24] https://arxiv.org/abs/2309.09690 (ICASSP24)

音声特有の表現



[Lakhotia21] https://arxiv.org/abs/2102.01192 (TACL21),[Nguyen24] https://arxiv.org/abs/2402.05755 (arxiv), [Ji24] https://arxiv.org/pdf/2408.16532 (arxiv) [Mousavi24] https://arxiv.org/abs/2406.10735 (interspeech24), [Tang24] https://arxiv.org/abs/2406.08905 (interspeech24), [Pasad24] https://arxiv.org/abs/2307.00162 (TACL24)

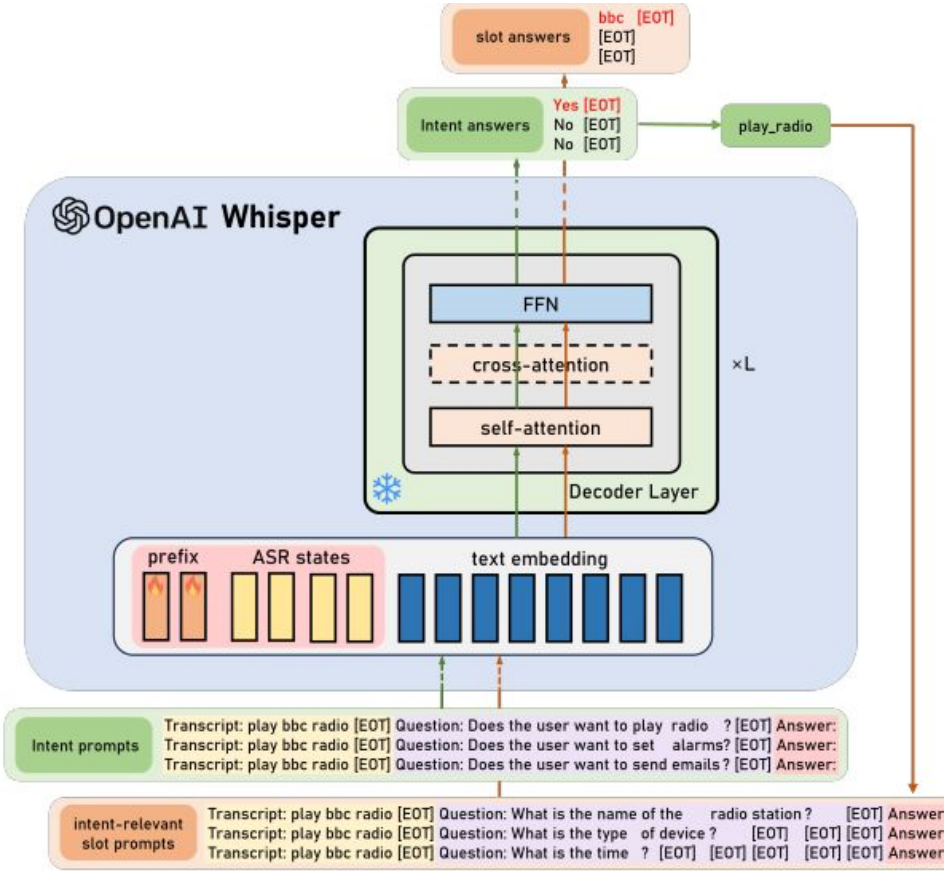
学習済みLLMの利用 (生成タスクに限定)



🔥：パラメータ更新．LoRAなど．

❄：パラメータ固定．

発話内容以外の**言語情報**を自然言語で表す①

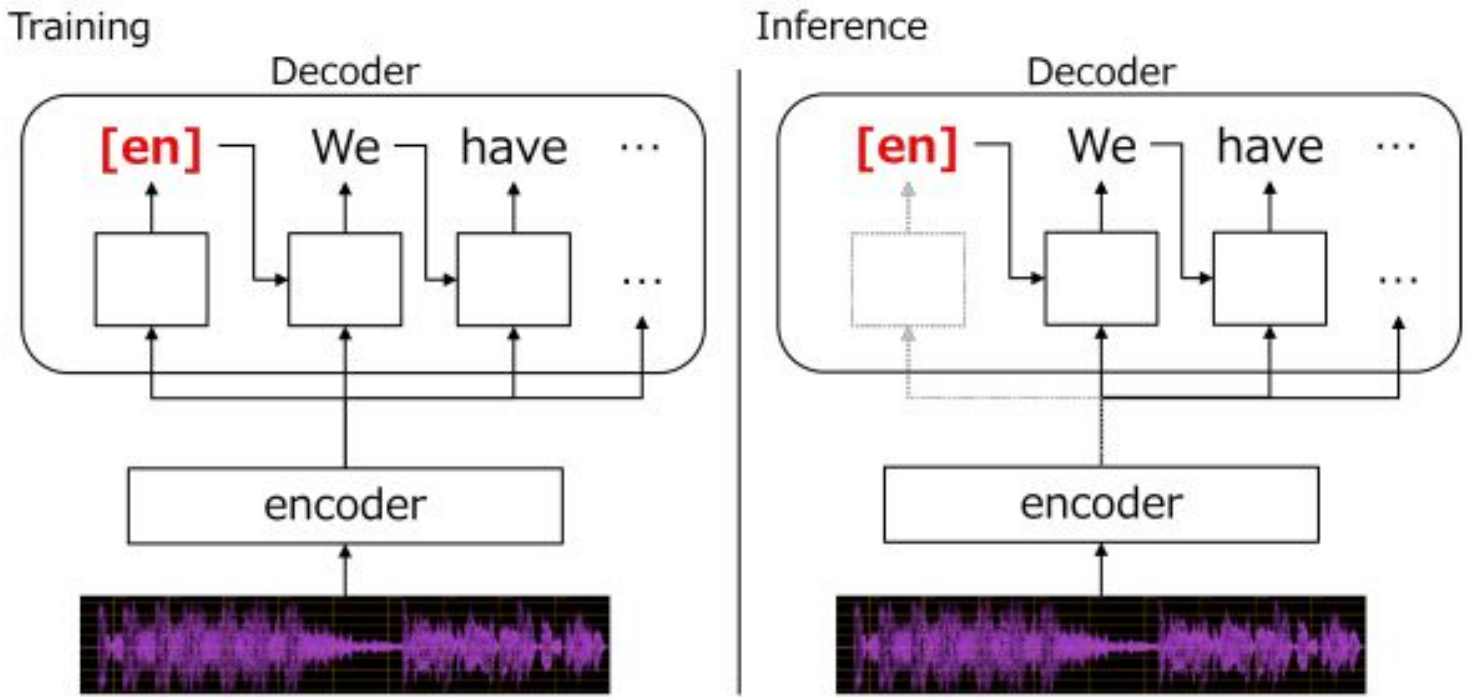


https://arxiv.org/pdf/2406.15209 (interspeech24)

- Intent classification (IC) & slot filling (SF) w/ Whisper decoder

- 音声認識 → IC → SF
  - Chain of thoughts のように、これまでの推論結果をプロンプトとして入れ、意図などを推論

発話内容以外の**言語情報**を自然言語で表す②



言語IDをプロンプトとして入れて、発話内容を推論する音声認識
(テキストのドメインをいれる研究もある)

https://arxiv.org/pdf/2406.12611 (interspeech24)。類似研究としてhttps://arxiv.org/pdf/2309.09443 (icassp24)。[] https://arxiv.org/pdf/2306.16007 (Interspeech24)

パラ言語情報 (感情など)・非言語情報 (話者性など) を自然言語で表す

CoCoCap-beta 日本語声質キャプショニンング with CocoNut Corpus

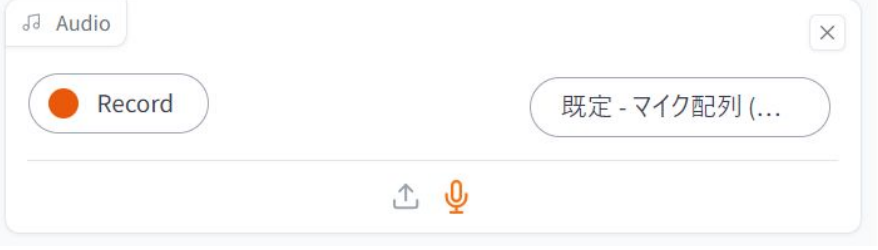
CocoCap-beta

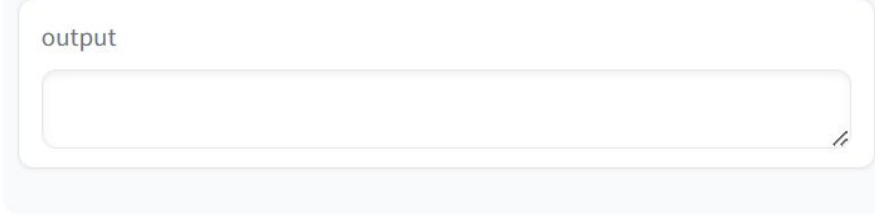
このスペースでは，CocoNutコーパスでfinetuningしたOpenAI whisperによる声質キャプショニング（どんな人がどんなスタイルで喋っているかの文章化）を示します．

Contributors / 貢献者

- 中田 亘
- 渡邊 亞椰
- 高道 慎之介
- 齋藤 佑樹

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



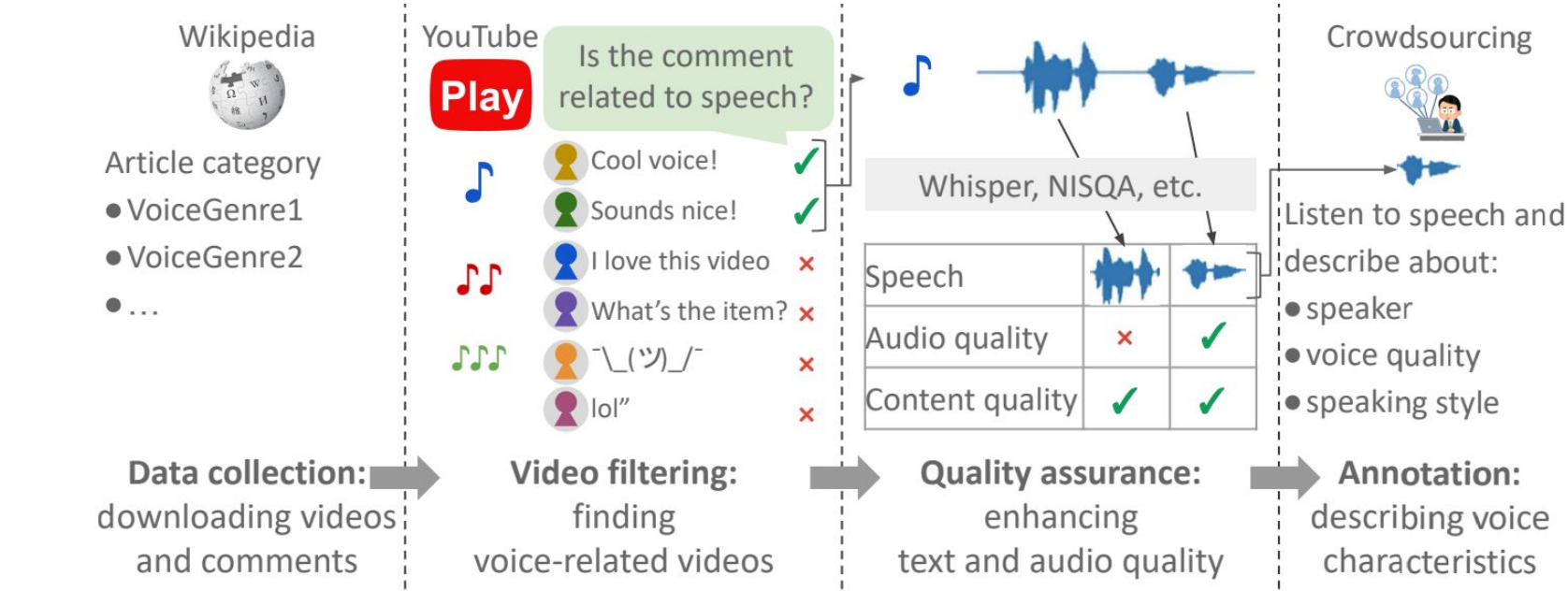


Slot filling 文、客観指標から推測可能な句、多様でない音声データが中心

- 文
  - “a middle-aged man”, “a boy’s voice” [Zhang23]
  - “This woman speaks in a soft voice” [Chen24]
  - “... the volume is normal, but she speaks very slowly” [Guo23][Ando24]
  - “A woman speaks slowly, with a high-pitched voice ….” [Kawamura24]
- 句
  - 高低、年齢、強弱、話速、性別、カテゴリ感情、感情強度[Xu24]、いくつかの声質表現句（固い、明るい、etc.）
- 音声データ
  - 既存の音声合成用データセット (多様性は限定的)

[Zhang23] https://arxiv.org/pdf/2310.05001 (ASRU23), [Chen24] https://arxiv.org/pdf/2406.08812 (Interspeech24), [Guo23] https://arxiv.org/abs/2211.12171 (ICASSP23),
[Kawamura24] https://arxiv.org/pdf/2406.07969 (interspeech24), [Ando24] https://www.isca-archive.org/interspeech_2024/ando24_interspeech.pdf (Interspeech24)
[Xu24] https://arxiv.org/pdf/2312.10381 (AAAI24), 詳細は不明だが https://arxiv.org/pdf/2305.19522 (Interspeech23)も同様

CoCo-Nut コーパス：Webクロールで作成した多様な音声データ & 多様な声質キャプション

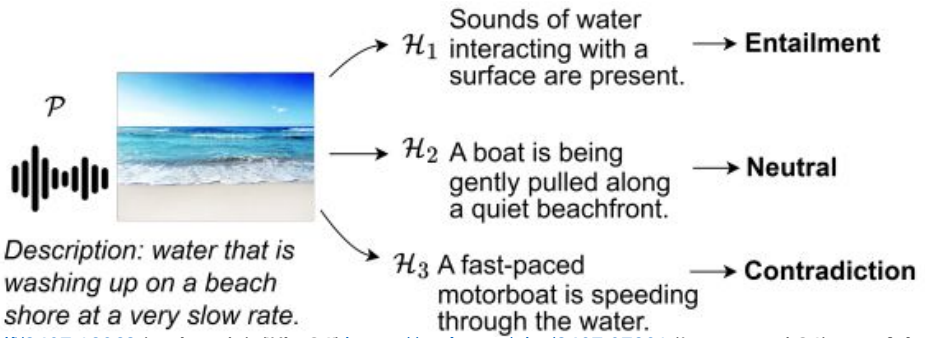


	30代くらいの男性の声。ゆっくりと穏やかな話し方でした。苦悩に満ちた、けだるそうな声でした。
	明るい中年の女性がはきはきとした声で楽しそうに喋っている。

[Watanabe23] https://arxiv.org/abs/2309.13509 (ASRU23), Webベースとして商品レビュー文を利用する研究 [Bott24] https://arxiv.org/pdf/2406.06406 (Interspeech24),

(余談) 音声以外にも使われる、自然言語による表現

- 自然言語 ⇔ 音イベント (無機物の音、自然音、etc)
- 関連研究のみリストアップ
  - 音内容 caption
    - “ロケットが飛び去る音”。日本語の場合はオノマトペあり [Okamoto24]
  - タイムスタンプ付き caption [Goel24]
    - “Sound of Howl … [0.406s-9.237s] … Wind noise … [2.128s-2.584s]
  - Audio Entailment (音の含意関係認識、右下) [Deshmukh24]
  - Audio visual caption [Kim24]
    - 音&画像でわかる説明文
  - Audio editing [Paissan24]
  - 前景背景付き caption [Lagrange24]



[Goel24] https://arxiv.org/pdf/2404.07616 (Interspeech24), [Deshmukh24] https://arxiv.org/pdf/2407.18062 (arxiv only), [Kim24] https://arxiv.org/abs/2407.07801 (Interspeech24)
[Paissan24] https://arxiv.org/abs/2310.12858 (Interspeech24), [Lagrange24] https://dcase.community/challenge2024/task-sound-scene-synthesis (DCASE24)

音声の多元自動評価

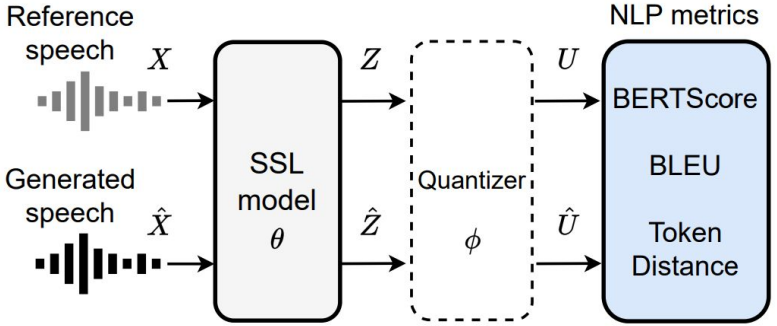
versa

VERSA

VERSA (Versatile Evaluation of Speech and Audio) is a toolkit dedicated to collecting evaluation metrics in speech and audio quality. Our goal is to provide a comprehensive connection to the cutting-edge techniques developed for evaluation. The toolkit is also tightly integrated into ESPnet.

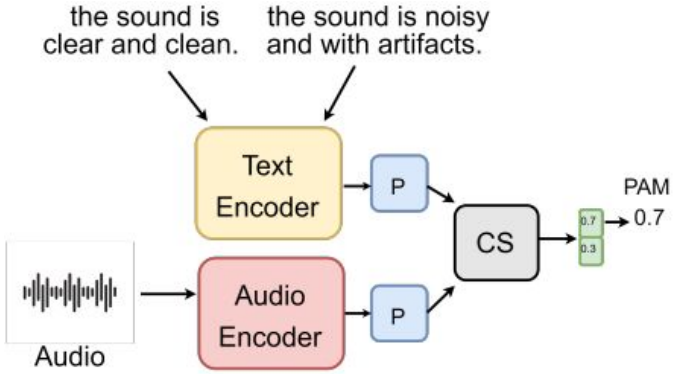
https://github.com/shinjiwlab/versa?tab=readme-ov-file

SpeechBERTScore



https://arxiv.org/abs/2401.16812

PAM (Prompting audio-lang. model)

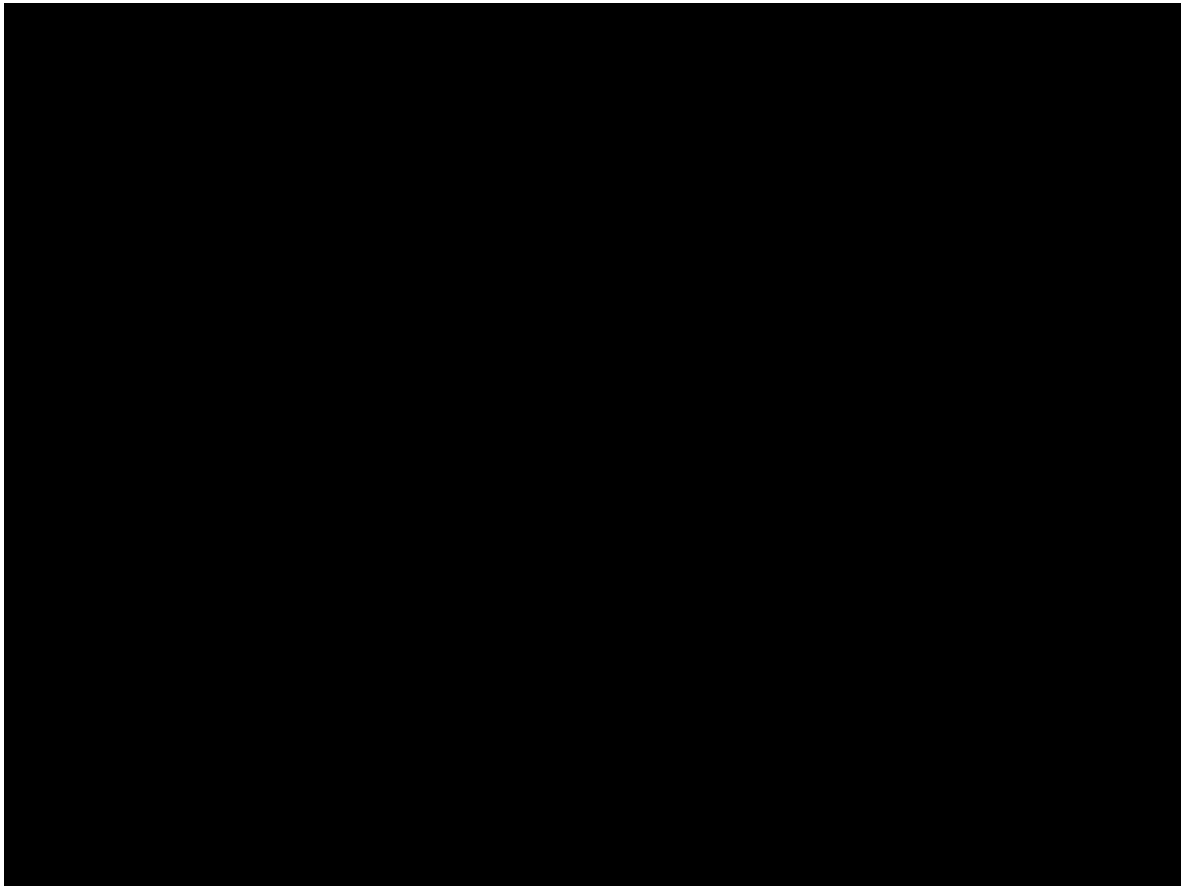


https://arxiv.org/pdf/2402.00282

- NLPほど自動評価は進んでいない
- 本質的に↓を解く
  - Listener, env., and lang. dependencies
  - Caption ambiguity

NLPに関連しそうな，私の研究の宣伝

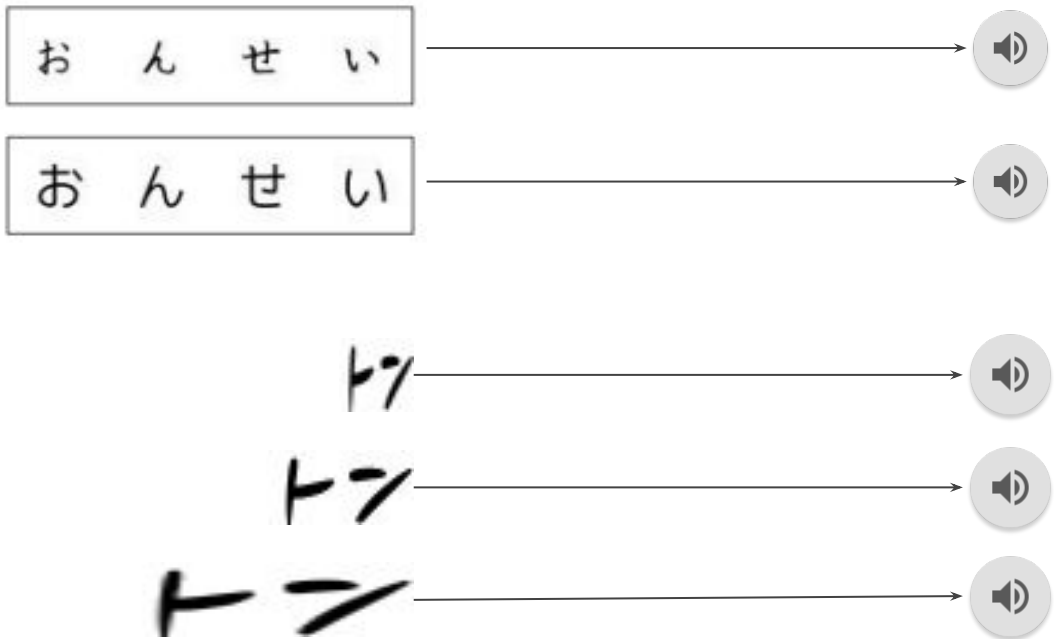
ゲーム実況の自動生成 (産総研 石垣さんプロジェクト. YANS2024 招待ポスターあり)



Saito: "SMASH corpus: a spontaneous speech corpus recording third-person audio commentaries on gameplay," Proc. LREC, 2020.

(公開版につき削除)

画像文字 ⇔ 音声・環境音



画像としての文字と，音声・環境音の相互変換！

Nakano, "vTTS: visual-text to speech," SLT, 2023.
Ohnaka, "Visual onoma-to-wave: environmental sound synthesis from visual onomatopoeias and sound-source images", ICASSP, 2023.

基盤モデルのための音声データベース (YODAS)

3-2-5 YODAS：YouTube 動画から構築される多言語大規模音声データセット

YODAS: YouTube-oriented multi-lingual speech dataset

Li Xinjian, ○高道 慎之介,
佐伯 高明, Chen William, 塩田 さやか, 渡部 晋治

Updates

- 2024/07/09: we also uploaded a new version of YODAS as YODAS2, it provides unsegmented audios and higher sampling rate (24k)

長い文脈の高品質音声

README

This is the YODAS manual/automatic subset from our YODAS dataset, it has 369,510 hours of speech.

世界最大のオープンコーパス (増強予定)

This dataset contains audio utterances and corresponding captions (manual or automatic) from YouTube. Note that manual caption only indicates that it is uploaded by users, but not necessarily transcribed by a human

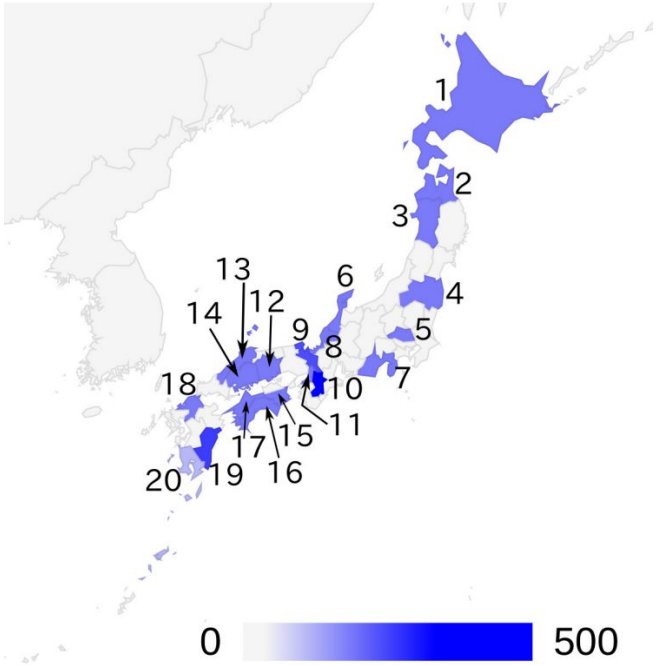
Downloads last month 69,998

Huggingface DLランキングに入っているらしい (?)

Li, "YODAS: Youtube-Oriented Dataset for Audio and Speech," ASRU, 2023.

日本諸方言の音声認識合成

21方言音声コーパス (CPJD)



https://aclanthology.org/L18-1067.pdf

音声認識合成に資する方言コーパス (科研費)

九州・沖縄方言の継承支援に資する音声対話型生成系AIの開発

坂井 美日　鹿児島大学, 総合科学域総合教育学系, 准教授 (00738916)

山田 高明　有明工業高等専門学校, 一般教育科, 助教 (10981285)
横山 晶子　大学共同利用機関法人人間文化研究機構国立国語研究所, 研究系,
宮川 創　筑波大学, 人文社会系, 准教授 (40887345)
中川 奈津子　九州大学, 人文科学研究院, 准教授 (50757870)
重野 裕美　広島大学, 人間社会科学研究科(教), 日本学術振興会特別研究員(P
加藤 幹治　大学共同利用機関法人情報・システム研究機構(機構本部施設等)
久保薗 愛　岡山大学, 社会文化科学学域, 准教授 (80706771)
高道 慎之介　慶應義塾大学, 理工学部(矢上), 准教授 (90784330)
當山 奈那　琉球大学, 人文社会学部, 准教授 (90792854)
高城 隆一　九州大学, 人文科学研究院, 助教 (90991597)

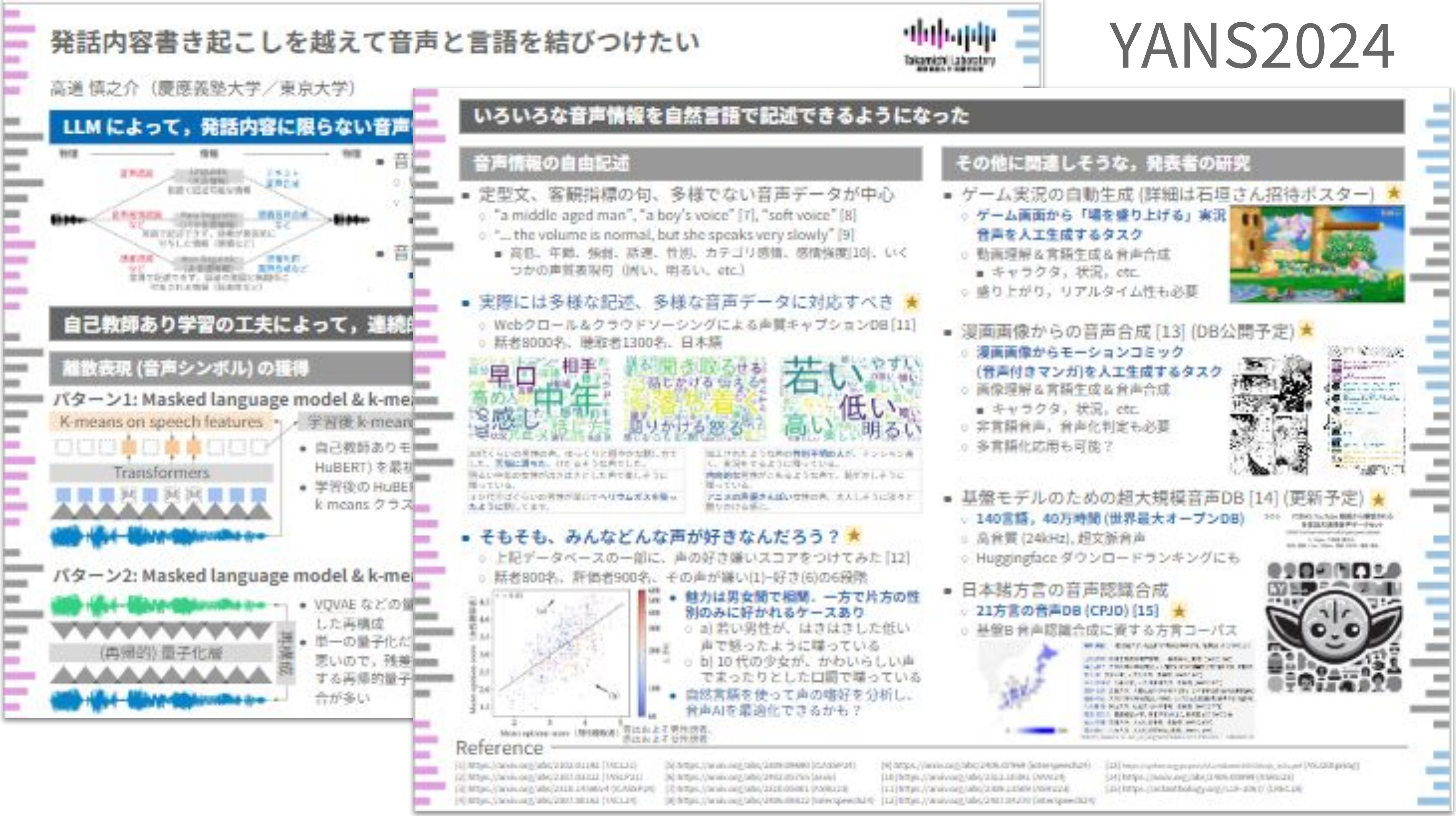
https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-24K00074/

まとめ

まとめ

- 自然言語と音声の境界領域を熱くしたい！
- 音声情報処理の考え方を改めるべきかも？
- 自然言語シンボルっぽく音声シンボルを扱える！
- 発話内容以外も自然言語で表せる！
- 高道の研究紹介

YANS2024


発話内容書き起こしを越えて音声と言語を結びつけたい
高道 慎之介（慶應義塾大学／東京大学）
LLM によって，発話内容に限らない音声
自己教師あり学習の工夫によって，連続
離散表現 (音声シンボル) の獲得
パターン1: Masked language model & k-means
K-means on speech features
学習後 k-means
Transformers
自己教師ありモ
HuBERT) を最初
学習後の HuBER
k means クラス
パターン2: Masked language model & k-means
VQVAE などの量
した再構成
単一の量子化だ
悪いので，残差
する再帰的量子
合が多い
(再帰的) 量子化層
いろいろな音声情報を自然言語で記述できるようになった
音声情報の自由記述
定型文、客観指標の句、多様でない音声データが中心
"a middle aged man", "a boy's voice" [7], "soft voice" [8]
"... the volume is normal, but she speaks very slowly" [9]
実際には多様な記述、多様な音声データに対応すべき
Webクロール＆クラウドソーシングによる声質キャプションDB [11]
そもそも、みんなどんな声が好きなんだろう？
上記データベースの一部に、声の好き嫌いスコアをつけてみた [12]
魅力は男女間で相関、一方で片方の性別のみに好かれるケースあり
自然言語を使って声の嗜好を分析し、音声AIを最適化できるかも？
その他に関連しそうな，発表者の研究
ゲーム実況の自動生成 (詳細は石垣さん招待ポスター)
ゲーム画面から「場を盛り上げる」実況音声を人工生成するタスク
動画理解＆言語生成＆音声合成
キャラクタ，状況，etc.
盛り上がり，リアルタイム性も必要
漫画画像からの音声合成 [13] (DB公開予定)
漫画画像からモーションコミック (音声付きマンガ)を人工生成するタスク
画像理解＆言語生成＆音声合成
キャラクタ，状況，etc.
非言語音声，音声化判定も必要
多言語化応用も可能？
基盤モデルのための超大規模音声DB [14] (更新予定)
140言語，40万時間 (世界最大オープンDB)
高音質 (24kHz)，超文脈音声
Huggingface ダウンロードランキングにも
日本語方言の音声認識合成
21方言の音声DB (CPJD) [15]
基盤B 音声認識合成に資する方言コーパス
Reference