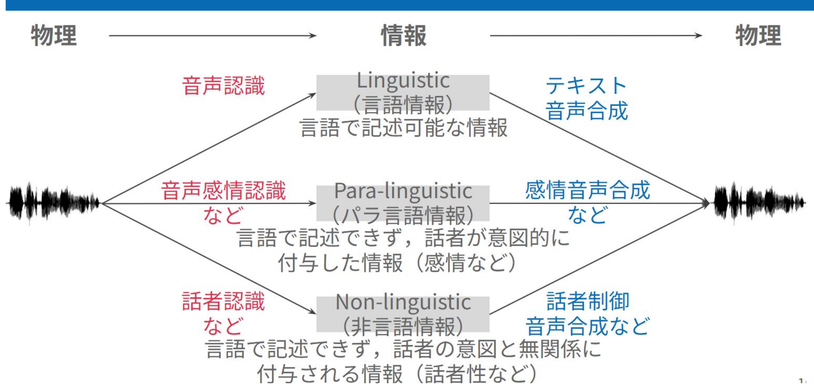
発話内容書き起こしを越えて音声と言語を結びつけたい

高道 慎之介（慶應義塾大学／東京大学）

Takamichi Laboratory
慶應義塾大学 高道研究室

LLM によって，発話内容に限らない音声情報を言語で記述できる？NLPと音声処理で何ができる？



- 音声の情報は言語情報・パラ言語情報・非言語情報に大別される
  - いままで自然言語で扱われていたのは，言語情報 (発話内容文) のみ
  - **でも，人間はパラ言語も非言語も自然言語で表せるよね？**
    - どういう声？ どういう感情？**計算機もそれをできる？ → 関連研究を紹介**

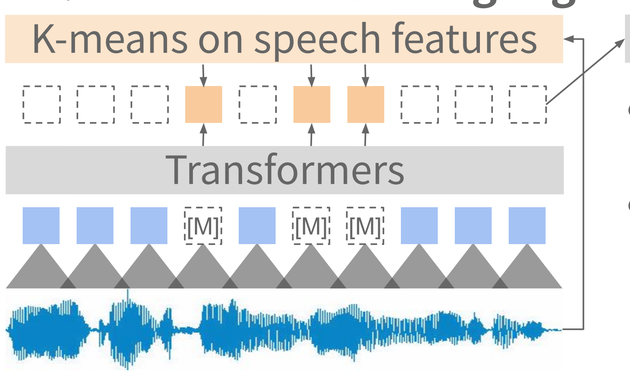
★:発表者 (高道) が関わっているもの

- 音声の表現は連続的で，自然言語(表層)は離散的．ちょっと違った
  - **最近は音声も離散表現するように！じゃあなにができる？ → 関連研究を紹介**

自己教師あり学習の工夫によって，連続的な音声波形を離散シンボル系列として扱えるようになった

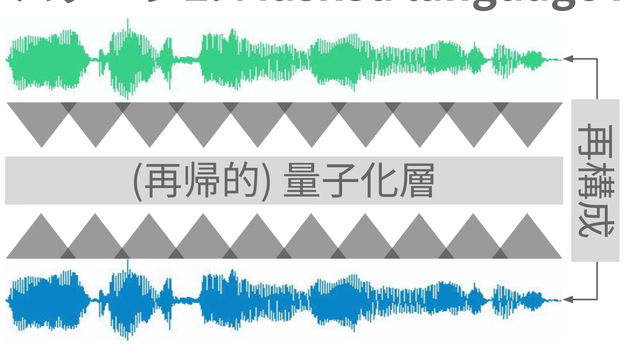
離散表現 (音声シンボル) の獲得

パターン1: Masked language model & k-means [1]



- 自己教師ありモデル (例えば HuBERT) を最初に学習
- 学習後の HuBERT 特徴量を k-means クラスタリング

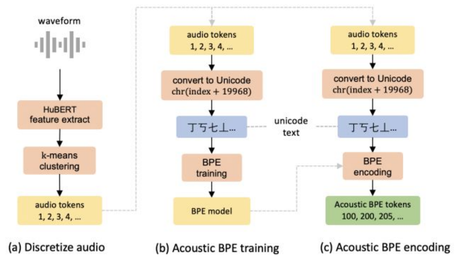
パターン2: Masked language model & k-means [2]



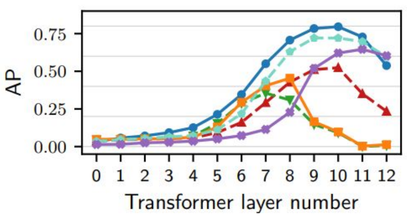
- VQVAE などの量子化層を内包した再構成
- 単一の量子化だと再構成精度が悪いので，残差をさらに量子化する再帰的量子化層を用いる場合が多い

NLP に関連しそうな研究

- 音声シンボルのサブワード化 [3]
  - **1シンボル = 1文字としてBPEを適用**
  - 生成タスクの高品質化・高速化
  - **多言語・サブワードの意味は未調査**



- Transformer の局所性 [4]
  - Transformer の各層で捉える情報が違う
  - **最終層手前が単語に類似** (→の色はモデル)



- 音声シンボルと Zipf 則 [5] ★
  - **単語のシンボル系列はべき乗則に従う**
  - 日本語(表意)と英語(表音)で異なる分布

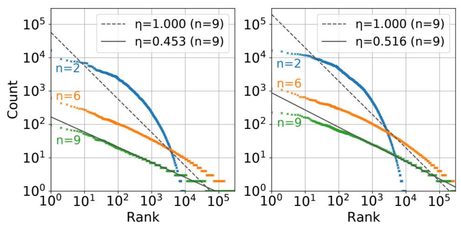


Fig. 4. Speech symbol n-gram rank-frequency distributions. left: Japanese, right: English. $n = 6$ in Japanese and $n = 2$ in English correspond to a character. $n = 9$ corresponds a word.

- 複数時間解像度 [6]
  - 音素に近いところは高い時間解像度
  - 韻律やスタイルは低い時間解像度

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

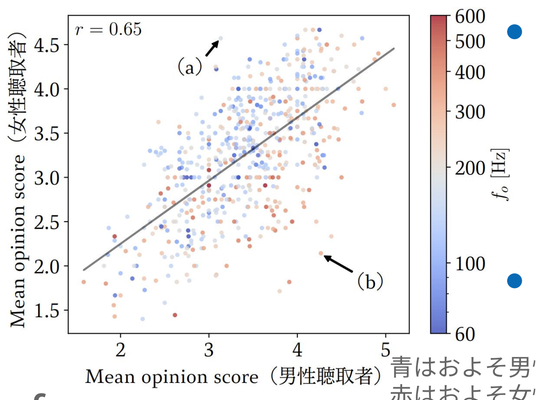
- 定型文、客観指標の句、多様でない音声データが中心
  - “a middle-aged man”, “a boy’s voice” [7], “soft voice” [8]
  - “... the volume is normal, but she speaks very slowly” [9]
    - 高低、年齢、強弱、話速、性別、カテゴリ感情、感情強度[10]、いくつかの声質表現句（固い、明るい、etc.）

- 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]
  - 話者8000名、聴取者1300名、日本語

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。
３０代半ばぐらいの男性が早口で**ヘリウムガスを吸ったように**話してます。

加工されたような声の**性別不明の人**が、テンション高く、実況をするように喋っている。
内向的な男性がこもるような声で、恥ずかしそうに喋っている。
アニメの声優さんぽい女性の声。大人しそうに淡々と語りかける感じ。

- **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]
  - 話者800名、評価者900名、その声が嫌い(1)~好き(6)の6段階

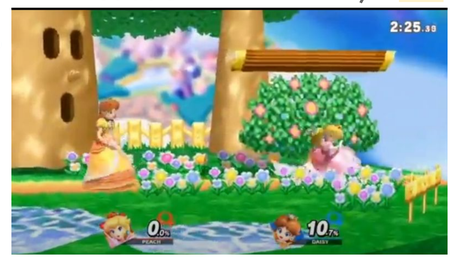


青はおよそ男性話者、
赤はおよそ女性話者

- **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**
  - 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
- 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

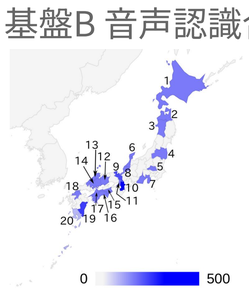
- ゲーム実況の自動生成 (詳細は石垣さん招待ポスター) ★
  - **ゲーム画面から「場を盛り上げる」実況音声を人工生成するタスク**
  - 動画理解＆言語生成＆音声合成
    - キャラクタ，状況，etc.
  - 盛り上がり，リアルタイム性も必要



- 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

- 日本諸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



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(
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

Reference

[1] https://arxiv.org/abs/2102.01192 (TACL21)
[2] https://arxiv.org/abs/2107.03312 (TASLP21)
[3] https://arxiv.org/abs/2310.14580v4 (ICASSP24)
[4] https://arxiv.org/abs/2307.00162 (TACL24)
[5] https://arxiv.org/abs/2309.09690 (ICASSP24)
[6] https://arxiv.org/abs/2402.05755 (arxiv)
[7] https://arxiv.org/abs/2310.05001 (ASRU23)
[8] https://arxiv.org/abs/2406.08812 (Interspeech24)
[9] https://arxiv.org/abs/2406.07969 (interspeech24)
[10] https://arxiv.org/abs/2312.10381 (AAAI24)
[11] https://arxiv.org/abs/2309.13509 (ASRU23)
[12] https://arxiv.org/abs/2407.04270 (Interspeech24)
[13] https://sython.org/papers/ASJ/takamichi2020asjs_m2v.pdf (ASJ20spring)
[14] https://arxiv.org/abs/2406.00899 (ASRU23)
[15] https://aclanthology.org/L18-1067/ (LREC18)