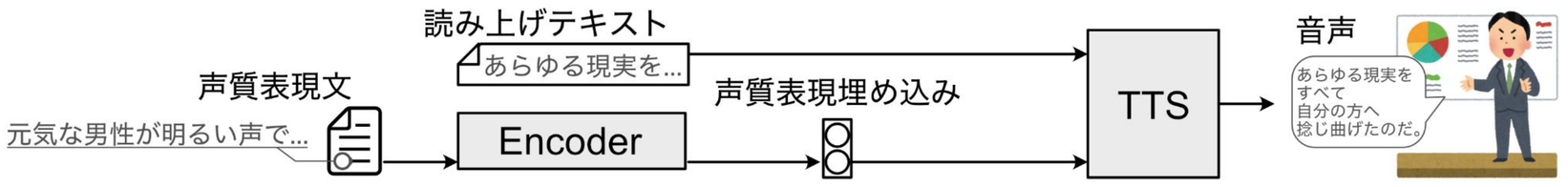
2-P-11 対照学習モデルによる音声-声質表現文の埋め込み表現獲得

☆渡邊 亞椰, 高道 慎之介, 齋藤 佑樹,中田 亘, 辛 徳泰, 猿渡 洋 (東大院・情報理工)

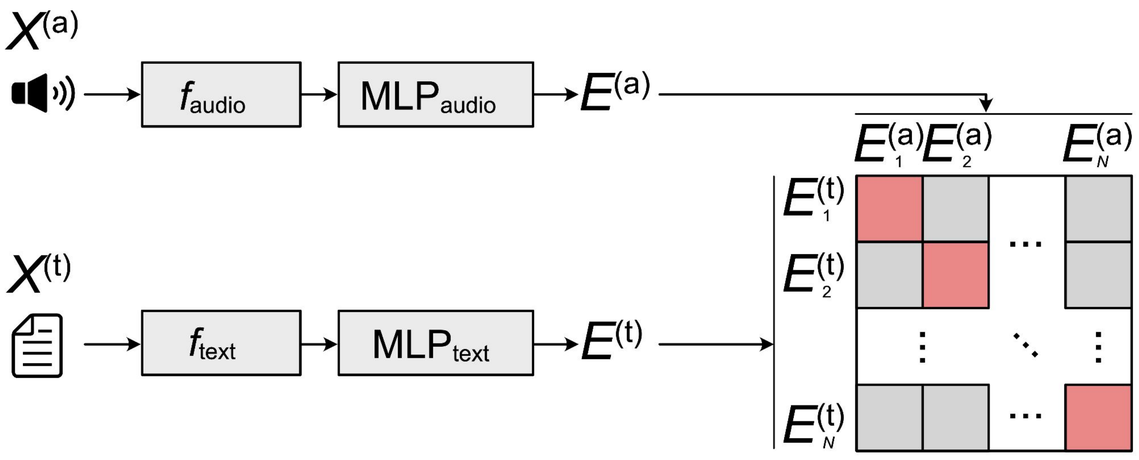
概要: テキスト音声合成(Text-to-speech: TTS)の声質操作に向けた自由記述-声質ペア埋め込みモデル

- **目的: 自由記述による声質操作**
  - 多様な要素を制御できる操作手段
  - 画像等での成功
- **声質表現文からの音声合成**
  - 声質表現文による声質操作の台頭 [1-3]
    - 少数話者→声質の多様性に欠ける →**大規模コーパスの必要性**
  - 声質表現文→潜在表現エンコーダ
    - 大規模コーパスで学習可能な手法
    - 画像分野: 対照学習モデル
- **Coco-Nutコーパス[4]**
  - 声質を表現する文と音声のペアコーパス
    - 例: 元気な男性が明るい声で、テンション高く発表をするように喋っている。
  - Webから収集したデータにより構築
    - **多様な声質・記述**
- **提案: 対照学習による埋め込み**
  - 声質表現文→潜在表現エンコード可
  - CLAP[5]に基づく対照学習
    - 大規模データから記述-音声対応学習
  - 特徴予測学習の併用
    - 声質の特徴を明示的に学習



手法: 対照学習 + 特徴予測学習により声質表現文-音声関係を学習する埋め込みモデル

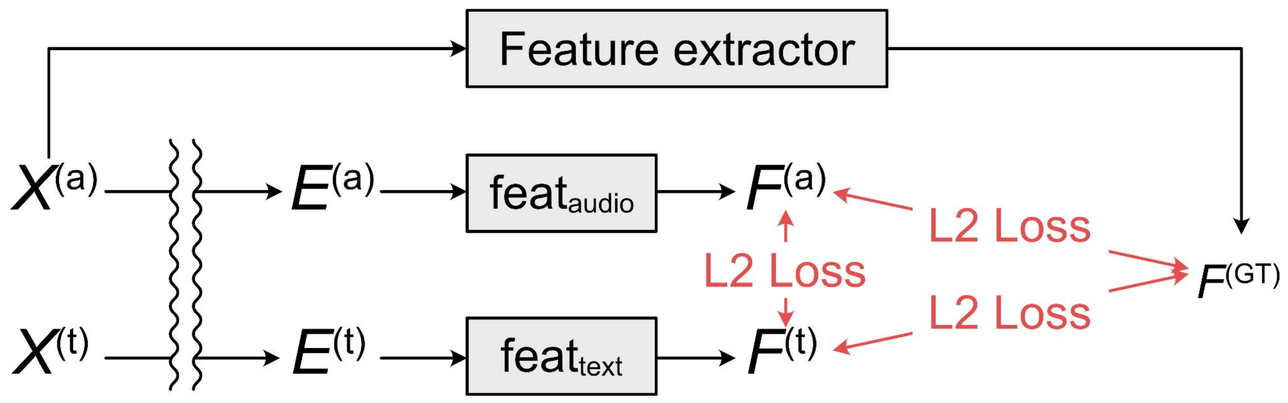
CLAP [5] ベース対照学習



- 音声$X^{(a)}$と声質表現文$X^{(t)}$を同一空間の埋め込み$E^{(a)}$、$E^{(t)}$に変換
  - f_{audio}: 事前学習済みHuBERT [6]
  - f_{text}: 事前学習済み日本語RoBERTa [7]
- 潜在空間上で正例ペアの距離が近くなるよう学習
- **声質表現文に対応した表現を得られる保証はない**

特徴予測学習

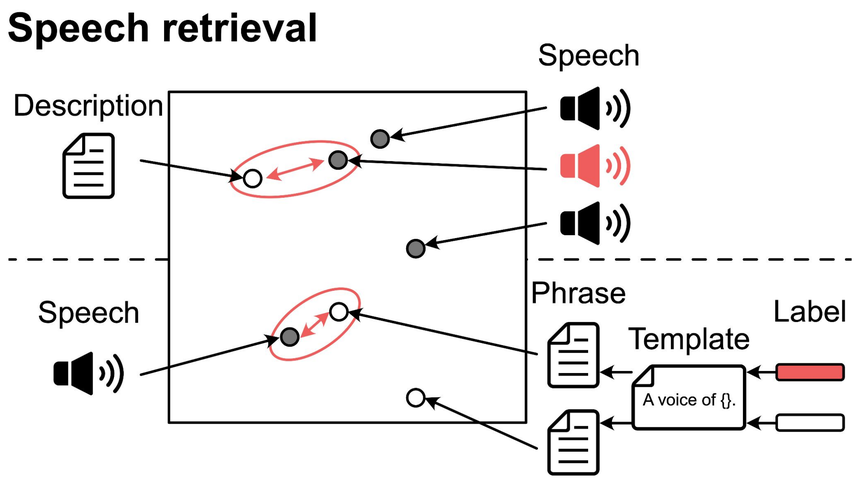
- 対照学習における補助タスクの有効性 [8] →**声質学習効率化の補助タスク**



- **声質に関係する特徴量を埋め込みから予測**
  - 音声から信号処理で正解特徴量が得られるもの
- **L2ノルムを損失関数として学習**
  - 声質表現文埋め込み予測-正解、音声埋め込み予測-正解誤差
  - 声質表現文埋め込み予測-音声埋め込み予測間の誤差

評価タスク

- モデル性能を検証するための下流タスクを設定



声質表現文による音声取得

1. 取得対象音声を全て埋め込む
2. 取得に用いる声質表現文を埋め込む
3. 声質表現文と**コサイン類似度が最も高い**音声を取得結果とする

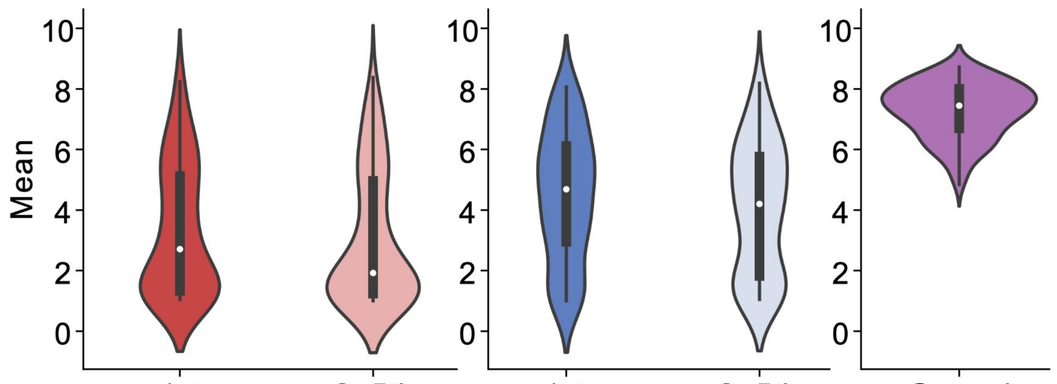
Zero-shot音声分類

1. 全分類対象ラベルの声質表現フレーズの作成
   - ラベルとテンプレートから自動生成
     - ラベル例: 性別、感情等
     - テンプレート例: ○○の声
2. 声質表現フレーズを全て埋め込む
3. 分類対象音声を埋め込む
4. 音声と**コサイン類似度が最も高い声質表現フレーズのラベル**へ分類する

評価: 声質取得性能の主観・客観評価

声質表現文による音声取得についての主観評価

- **声質表現文-取得された音声の対応度合いに対しての主観による評価**
  - 取得対象音声: Coco-Nut評価セット内全音声
  - 取得に使用する声質表現文: Coco-Nut評価セット内声質表現文からランダム100件
  - 評価: クラウドワーカーによる聴取評価
    - 1(全く対応していない)~9(とてもよく対応している)の9段階評価
    - 各声質表現文-音声ペアについて20人分の評価を平均
  - 特徴量予測使用モデル・不使用モデルについて評価
    - 1st candidate: 声質表現文-コサイン類似度が最も高い音声のペア
    - 2~5th candidate: コサイン類似度が2-5番目に高い音声からランダム1件選んだペア
  - 比較として正例ペア(Ground Truth)の対応度合いについても評価



- **評価結果(評価分布)**
- 特徴量予測なし:
  - 低評価側に山あり
- 特徴量予測あり:
  - 平均の向上

→特徴量予測により極端な不一致を防ぎ、**性能が向上**

→ただし**正例ペアの一致度には及ばない**

Zero-shot音声分類についての客観評価

- **性別ラベルについてのZero-shot分類性能**
  - 分類対象音声: JVS [9] parallel100セット内全音声
  - ラベル: 男性/女性
  - テンプレート:「○○が喋っている。」
- **評価結果(分類正答率)**
  - 特徴量予測を導入した方が若干精度が向上

特徴量予測	正解性別	正答率
なし	男性	0.998
	女性	0.941
あり	男性	0.986
	女性	1.000

展望

- 提案モデルの更なる性能向上、および提案モデルを組み込んだTTSの構築

Reference

[1] Z. Guo et al., ICASSP, 2023.
[2] D. Yang et al., arXiv, 2023.
[3] Y. Zhang et al., ASRU, 2023.
[4] A. Watanabe et al., ASRU, 2023.
[5] B. Elizalde et al., arXiv, 2023.
[6] WN. Hsu et al., IEEE TASLP, 2021.
[7] Y. Liu et al., arXiv, 2019.
[8] CF. Yeh et al., ASRI, 2023.
[9] S. Takamichi et al., AST, 2020.