日本音響学会 2024年秋季研究発表会 3-6-4

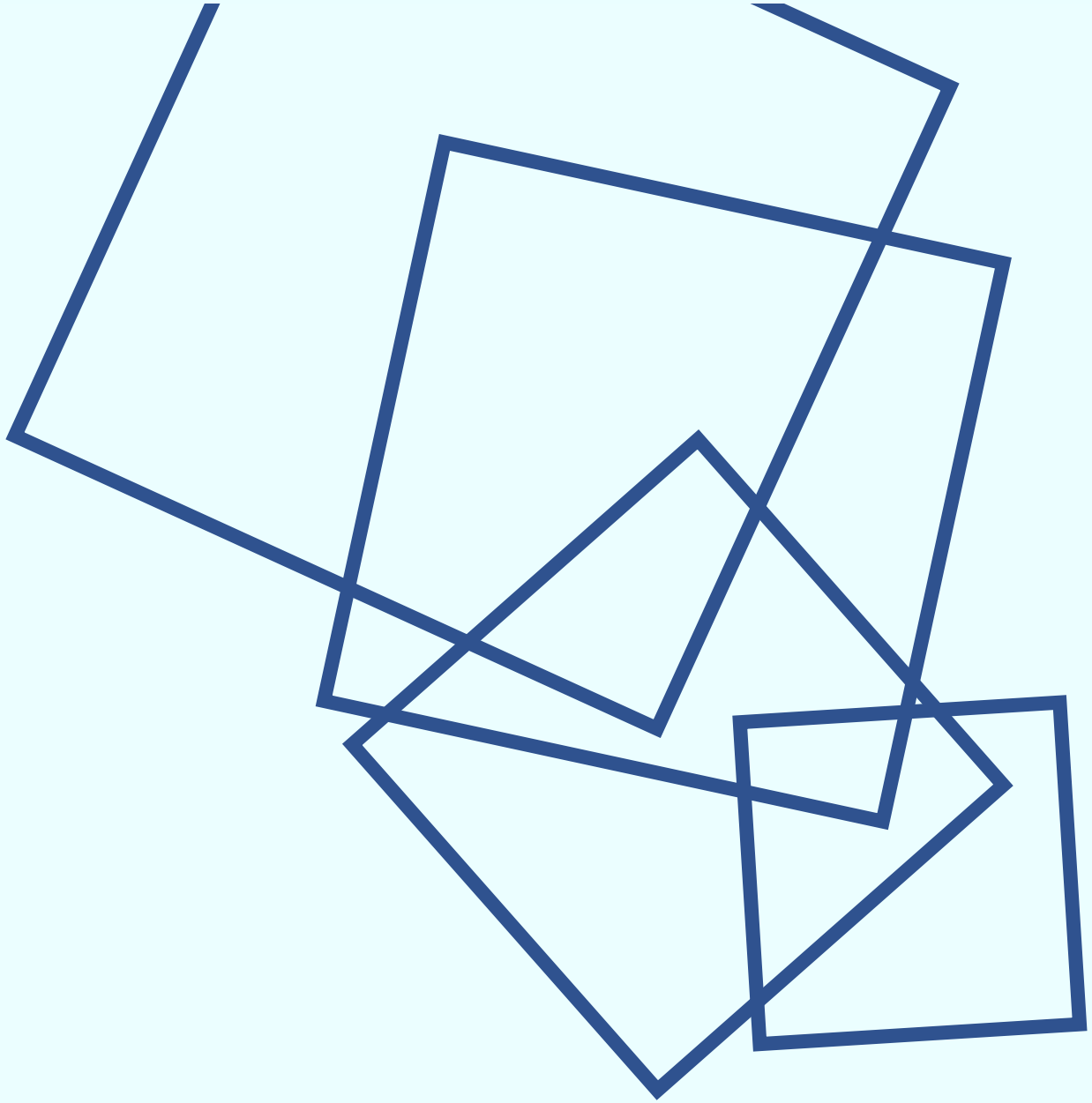

# UTMOSv2: 自然性MOS予測における スペクトログラム特徴量と SSL特徴量の統合的利用

☆馬場 凱渡，中田 亘，齋藤 佑樹，猿渡 洋（東京大学）

# Section1

## 研究概要 *Overview*

- 01. 研究概要
- 02. UTMOSv2
- 03. 実験的評価



# 研究背景



- 音声合成システムの評価：人間による主観評価に大きく依存
  Mean Opinion Score (MOS) テスト, AB テストなど

△ 時間的コスト
△ 金銭的コスト

→ 機械学習モデルによる自動化

# VoiceMOS Challenge 2024



- VoiceMOS Challenge (VMC) 2024：MOS予測の国際コンペ

・2022 [Huang+22], 2023 [Cooper+23a] に続く第3回目の開催

・本発表では, VMC2024 Track 1（高品質合成音声のMOS予測）を対象

# VMC 2024 Track1 Results



MSE: Mean Squared Error
LCC: Linear Correlation Coefficient
**太字**：1位 SRCC: Spearman's Rank Correlation Coefficient
<u>下線</u>：2位 KTAU: Kendall's tau

| | Zoom-in rate: 25% | | | | | | | | Zoom-in rate: 12% | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Utterance-level | | | | System-level | | | | Utterance-level | | | | System-level | | | |
| | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ |
| B01 | 1.154 | 0.508 | 0.509 | 0.358 | 0.998 | 0.750 | 0.745 | 0.539 | 0.741 | 0.422 | 0.417 | 0.285 | 0.589 | 0.608 | 0.609 | 0.444 |
| T01 | 0.358 | 0.490 | 0.491 | 0.342 | 0.162 | 0.703 | 0.700 | 0.495 | 0.296 | 0.402 | 0.421 | 0.286 | 0.127 | 0.540 | 0.513 | 0.333 |
| T02 | 0.355 | 0.550 | 0.555 | 0.394 | 0.141 | 0.789 | 0.769 | 0.578 | 0.410 | 0.520 | 0.544 | 0.377 | 0.240 | 0.745 | 0.702 | 0.467 |
| T03 | 0.406 | 0.471 | 0.488 | 0.344 | 0.200 | 0.674 | 0.672 | 0.483 | 0.264 | 0.481 | 0.514 | 0.357 | 0.082 | 0.779 | 0.768 | 0.539 |
| T04 | 0.718 | 0.421 | 0.391 | 0.272 | 0.494 | 0.728 | 0.698 | 0.512 | 0.457 | 0.311 | 0.281 | 0.191 | 0.270 | 0.571 | 0.549 | 0.364 |
| UTMOSv2 T05 | <u>0.287</u> | **0.686** | **0.679** | **0.497** | <u>0.073</u> | **0.944** | <u>0.940</u> | <u>0.773</u> | **0.203** | **0.654** | <u>0.658</u> | <u>0.464</u> | **0.056** | <u>0.834</u> | <u>0.788</u> | <u>0.610</u> |
| T06 | **0.255** | <u>0.658</u> | <u>0.676</u> | <u>0.493</u> | **0.047** | <u>0.931</u> | **0.943** | **0.793** | <u>0.221</u> | <u>0.629</u> | **0.695** | **0.505** | <u>0.057</u> | **0.890** | **0.952** | **0.824** |
| T07 | 0.753 | 0.377 | 0.359 | 0.254 | 0.446 | 0.724 | 0.675 | 0.493 | 0.517 | 0.300 | 0.294 | 0.207 | 0.251 | 0.676 | 0.672 | 0.444 |

# VMC 2024 Track1 Results



✨🏆7/16項目で1位，残り9項目で2位🏆✨

MSE: Mean Square Error
LCC: Linear Correlation Coefficient
SRCC: Spearman's Rank Correlation Coefficient
KTAU: Kendall's tau
太字：1位
下線：2位

| | Zoom-in rate: 25% | | | | | | | | Zoom-in rate: 12% | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Utterance-level | | | | System-level | | | | Utterance-level | | | | System-level | | | |
| | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ | MSE↓ | LCC↑ | SRCC↑ | KTAU↑ |
| B01 | 1.154 | 0.508 | 0.509 | 0.358 | 0.998 | 0.750 | 0.745 | 0.539 | 0.741 | 0.422 | 0.417 | 0.285 | 0.589 | 0.608 | 0.609 | 0.444 |
| T01 | 0.358 | 0.490 | 0.491 | 0.342 | 0.162 | 0.703 | 0.700 | 0.495 | 0.296 | 0.402 | 0.421 | 0.286 | 0.127 | 0.540 | 0.513 | 0.333 |
| T02 | 0.355 | 0.550 | 0.555 | 0.394 | 0.141 | 0.789 | 0.769 | 0.578 | 0.410 | 0.520 | 0.544 | 0.377 | 0.240 | 0.745 | 0.702 | 0.467 |
| T03 | 0.406 | 0.471 | 0.488 | 0.344 | 0.200 | 0.674 | 0.672 | 0.483 | 0.264 | 0.481 | 0.514 | 0.357 | 0.082 | 0.779 | 0.768 | 0.539 |
| T04 | 0.718 | 0.421 | 0.391 | 0.272 | 0.494 | 0.728 | 0.698 | 0.512 | 0.457 | 0.311 | 0.281 | 0.191 | 0.270 | 0.571 | 0.549 | 0.364 |
| UTMOSv2 T05 | 0.287 | **0.686** | **0.679** | **0.497** | 0.073 | **0.944** | 0.940 | 0.773 | **0.203** | **0.654** | 0.658 | 0.464 | **0.056** | 0.834 | 0.788 | 0.610 |
| T06 | **0.255** | 0.658 | 0.676 | 0.493 | **0.047** | 0.931 | **0.943** | **0.793** | 0.221 | 0.629 | **0.695** | **0.505** | 0.057 | **0.890** | **0.952** | **0.824** |
| T07 | 0.753 | 0.377 | 0.359 | 0.254 | 0.446 | 0.724 | 0.675 | 0.493 | 0.517 | 0.300 | 0.294 | 0.207 | 0.251 | 0.676 | 0.672 | 0.444 |

# VMC 2024 データ



- データ: Zoomed-in BVCC


  - Zoomed-in MOS予測の**学習セットは提供されない**

# VMC 2024 評価プロトコル



- MSE: 平均二乗誤差 (Mean Squared Error)
- LCC: 線形相関係数 (Linear Correlation Coefficient)
- SRCC: スピアマンの順位相関係数(Spearman's Rank Correlation Coefficient)
- KTAU: ケンドールの順位相関係数(Kendall's TAU)

- 発話ごとの評価 (Utterance-level)
- 音声合成システムごとの評価 (System-level)

▸ 本発表では12%🔍でのMSEとSRCCのみ報告

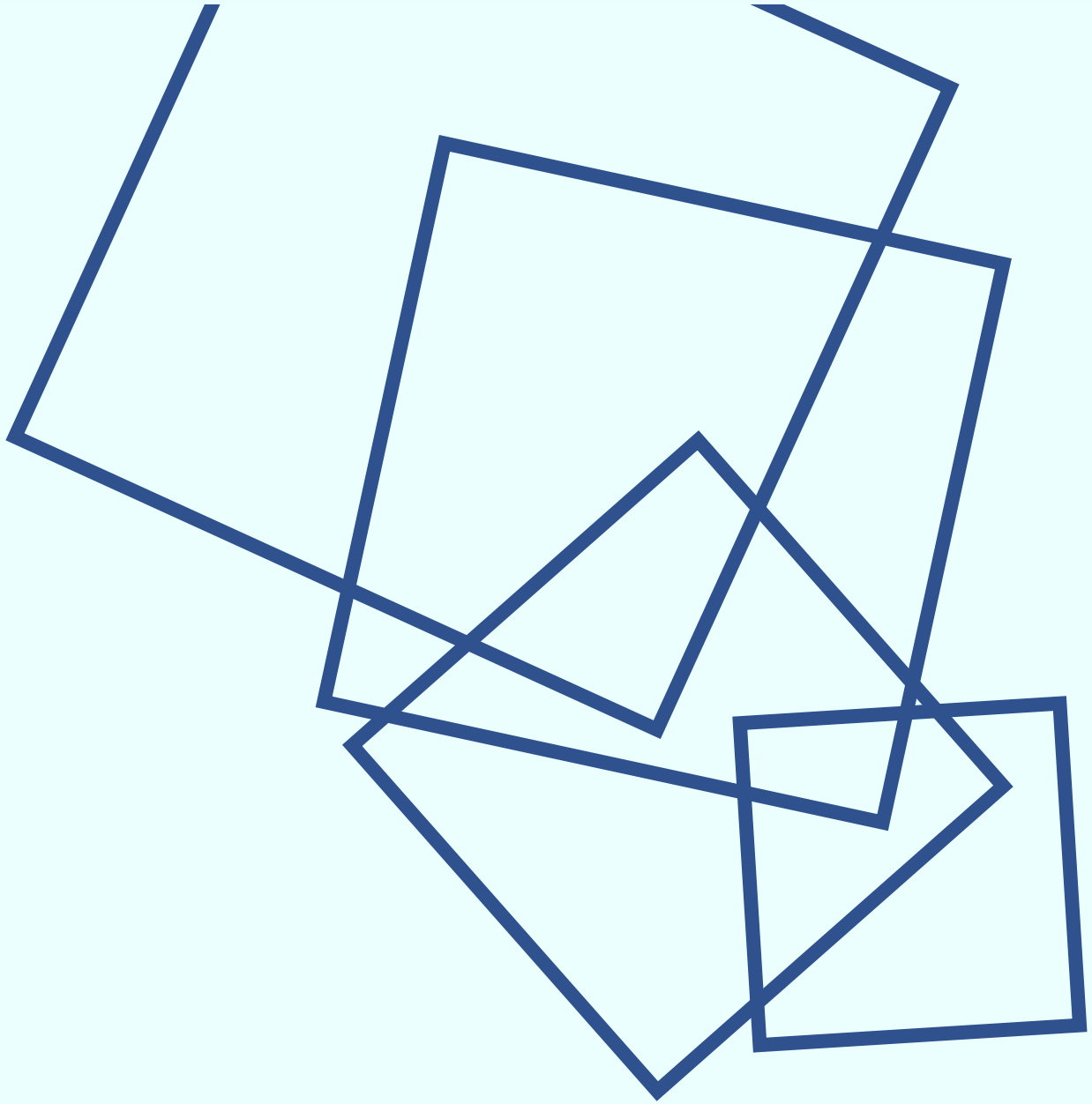
# Section2

## UTMOSv2 *UTMOSv2*


01. 研究概要

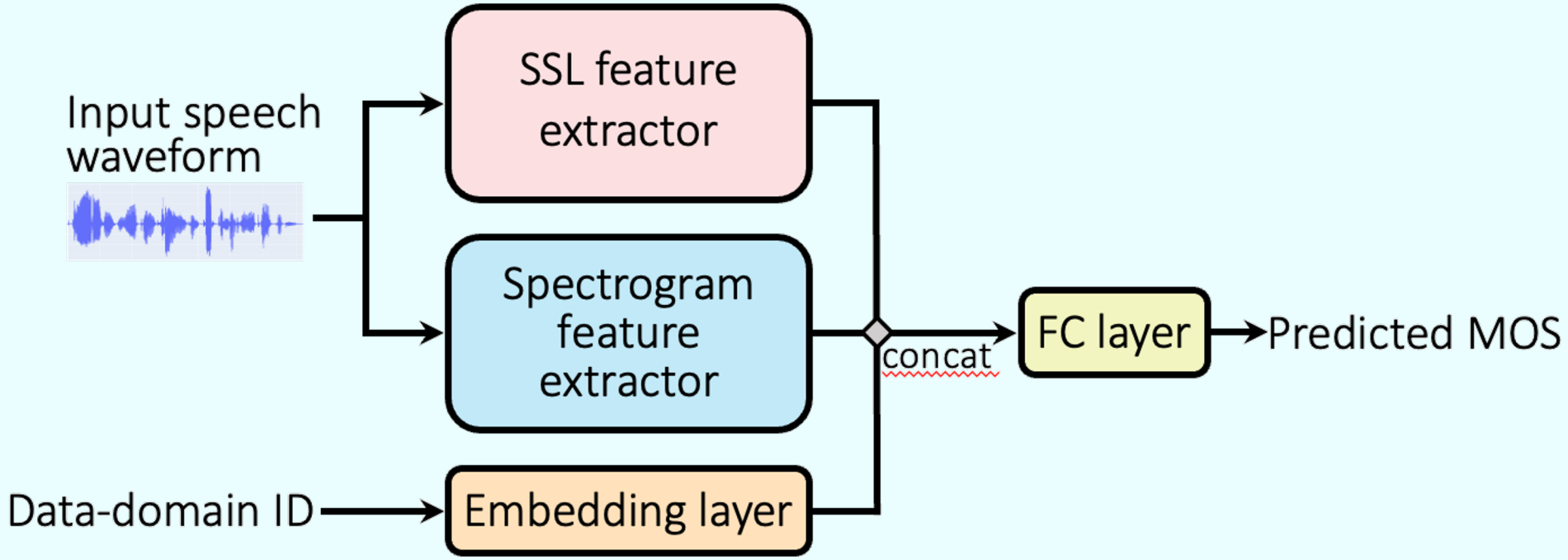
02. UTMOSv2

03. 実験的評価

# UTMOSv2の基本構造

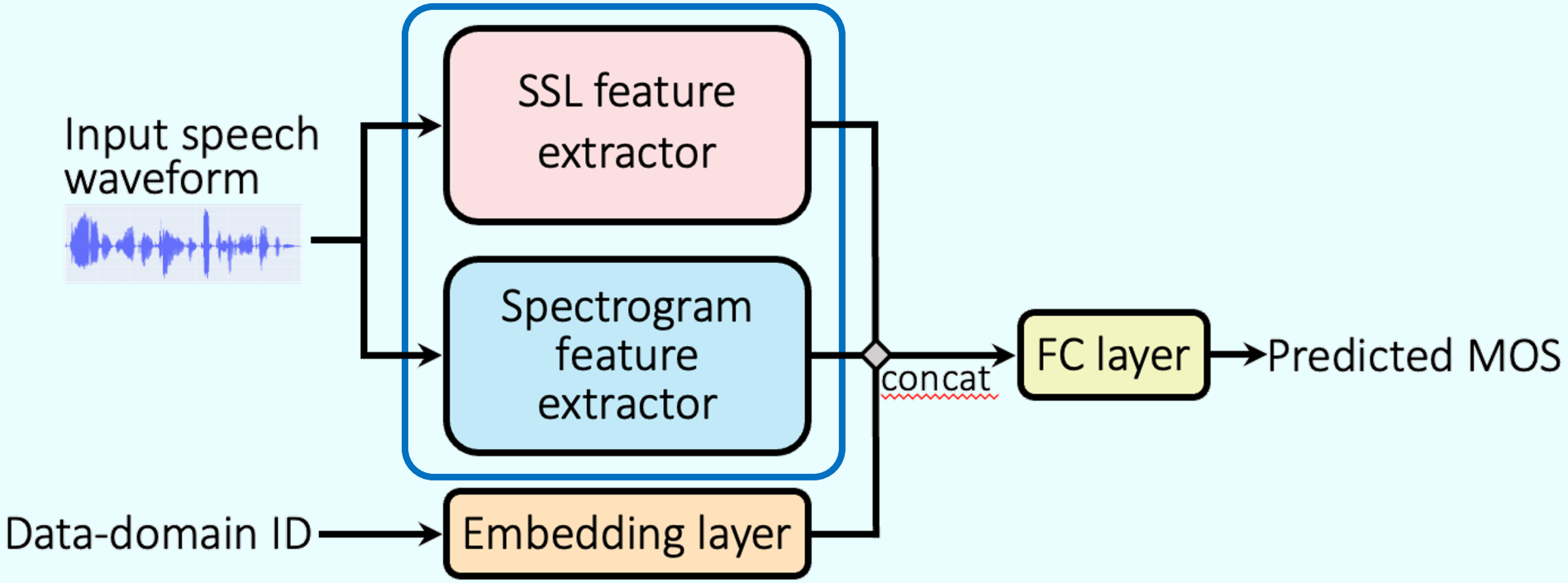




# UTMOSv2の基本構造



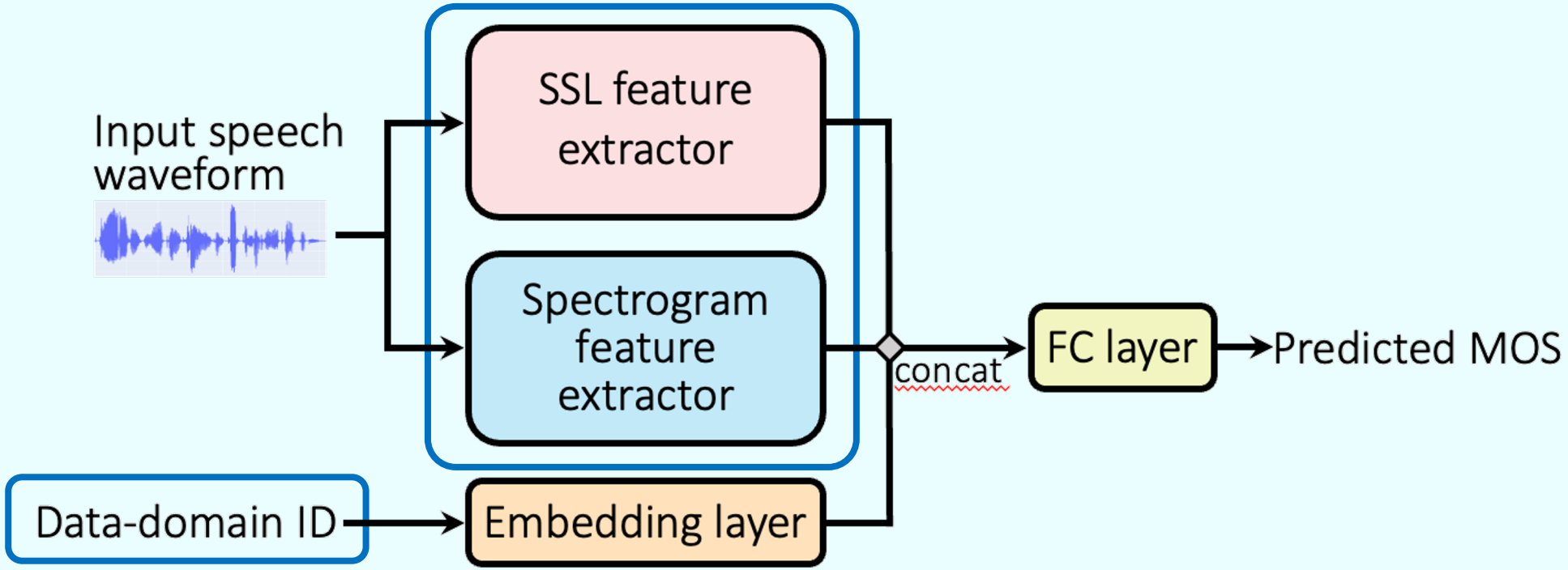
自己教師あり学習 (SSL) 特徴量と
スペクトログラム特徴量の fusion



# UTMOSv2の基本構造



自己教師あり学習 (SSL) 特徴量と
スペクトログラム特徴量の fusion

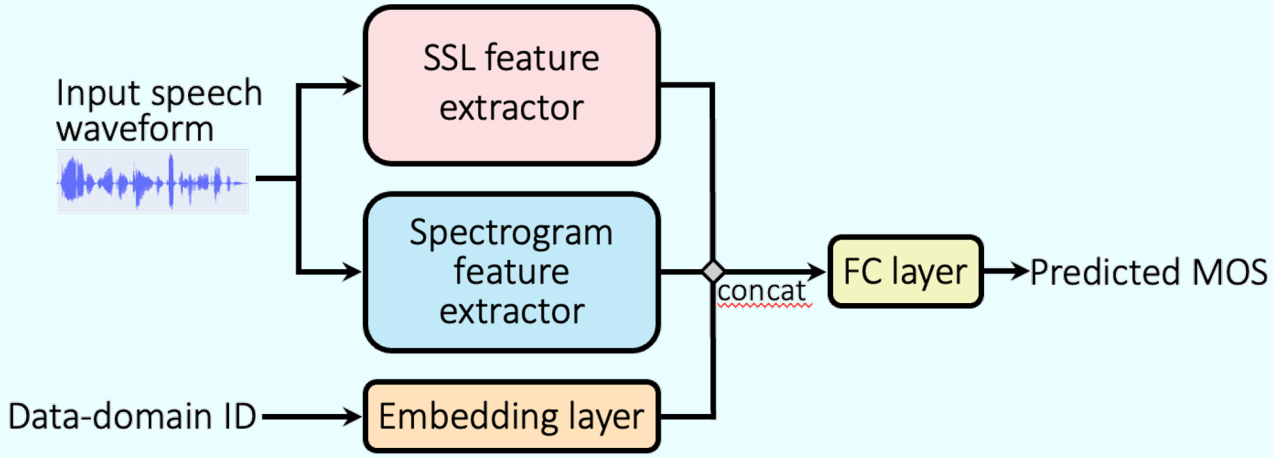



データドメインでの
条件付け

# UTMOSv2の3つの工夫



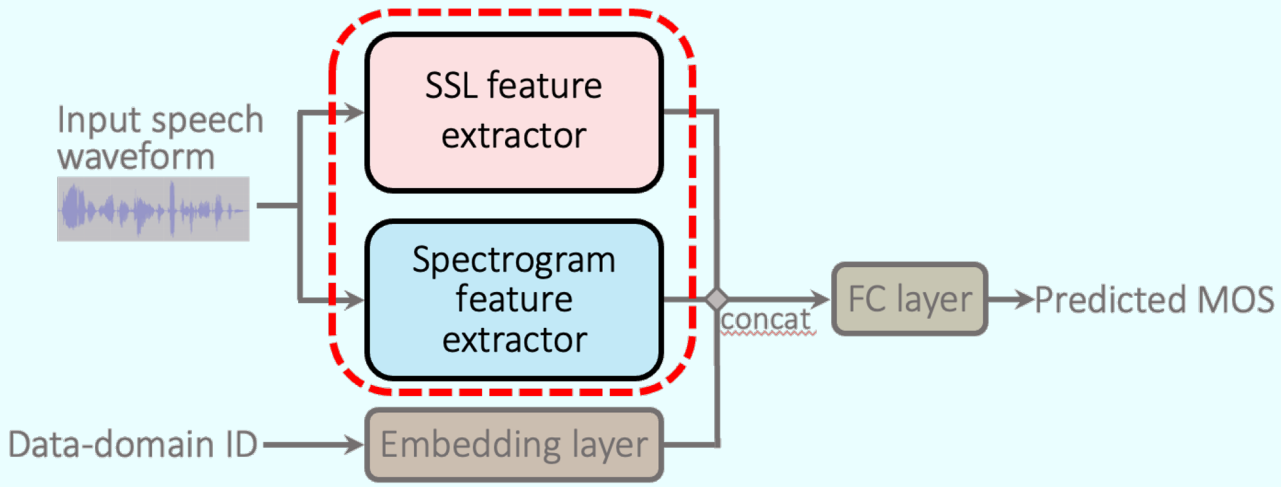
1. SSL/スペクトログラム特徴量のfusion
2. データドメインでの条件付け
3. モデルの多段階学習



# UTMOSv2の3つの工夫



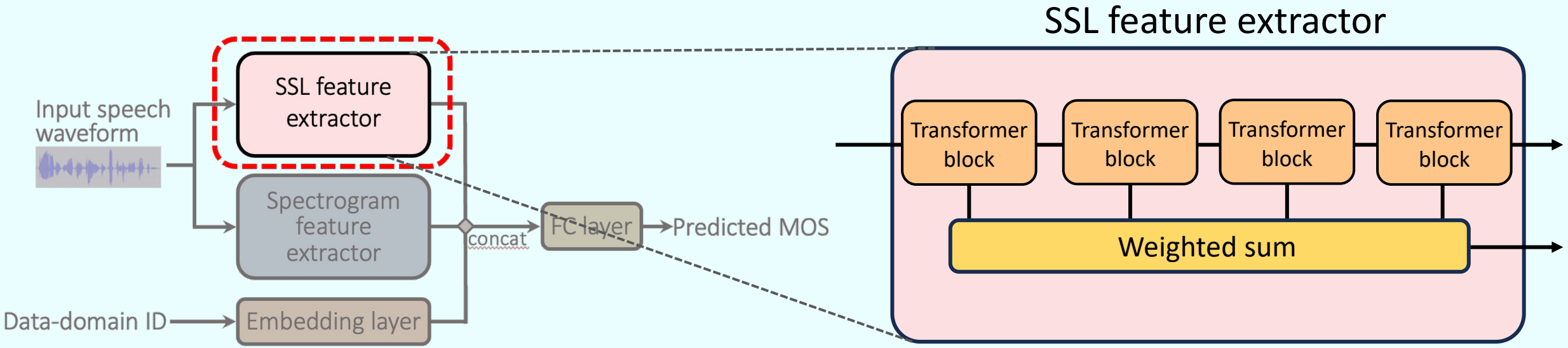
1. SSL/スペクトログラム特徴量のfusion
2. データドメインでの条件付け
3. モデルの多段階学習



# ① SSL/スペクトログラム特徴量のfusion



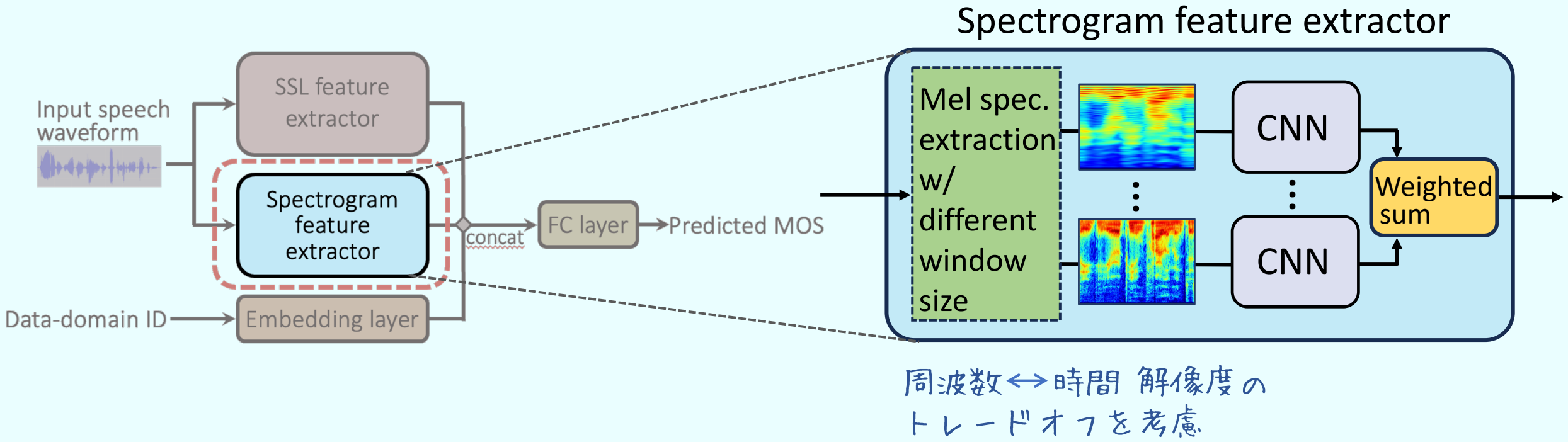
- SSL特徴量抽出器
  - UTMOS [Saeki+22] を参考に，事前学習済みのSSLモデルを用いて音声波形から汎用的特徴量を抽出



# ① SSL/スペクトログラム特徴量のfusion



- スペクトログラム (spec.) 特徴量抽出器　　kaggleでもよく見る手法！
  - パラ言語情報認識タスク [Szep+20][Amiriparian+17] における成功例をもとに
    スペクトログラムから画像認識モデルで特徴量を抽出

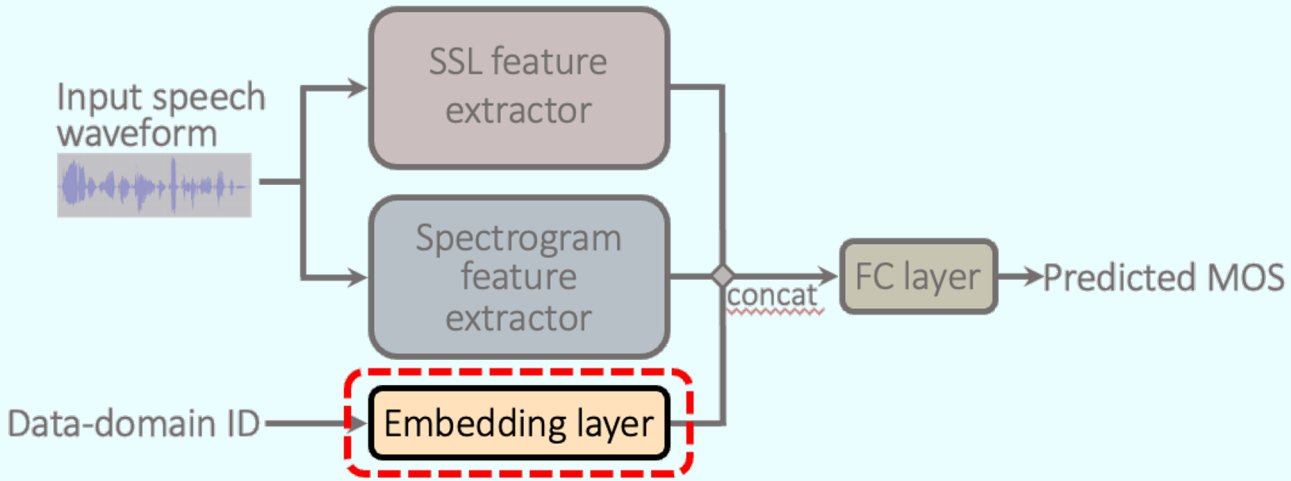



周波数↔時間 解像度の
トレードオフを考慮

# UTMOSv2の3つの工夫



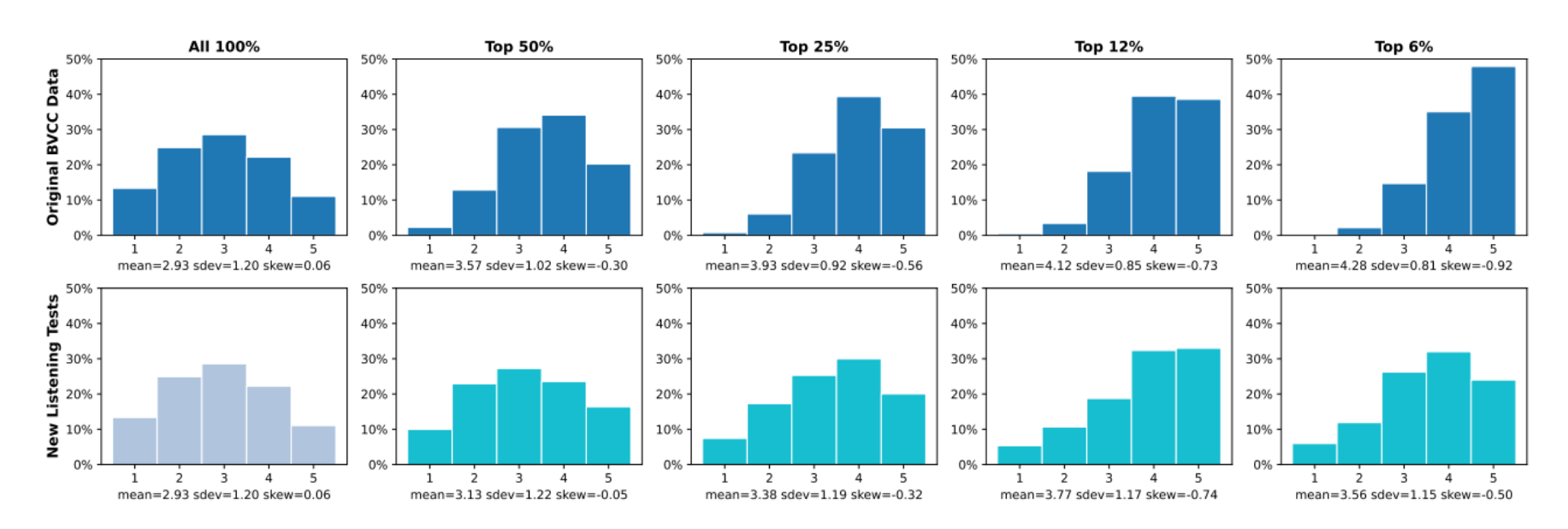
1. SSL/スペクトログラム特徴量のfusion
2. データドメインでの条件付け
3. モデルの多段階学習



# ② データドメインでの条件付け



- Range-equalizing bias
  - 比較する音声合成システムによって評価結果が異なる
    → 単純に混ぜて学習すると学習が不安定に (Range-equalizing bias [Cooper+23])




[Cooper+23]より引用

# ② データドメインでの条件付け



- 学習に用いたデータセットの ID を表す embedding を追加

  UTMOS [Saeki+22] を参考

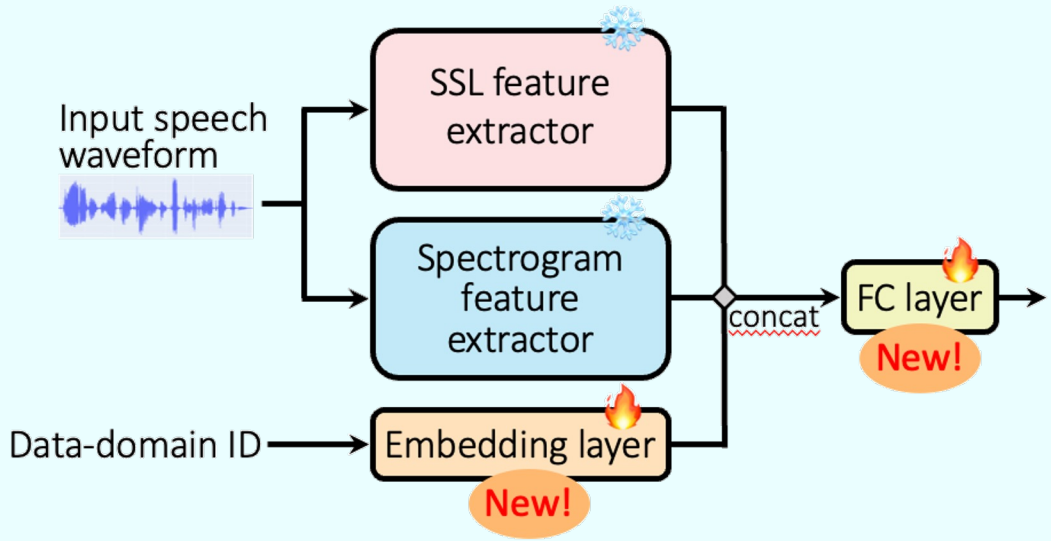
  - 推論時には学習済みデータセットIDのどれかを入力して予測

- 本研究で用いたデータセット
  - BVCC: VMC2022 [Huang+22] のデータセット
  - Blizzard Challenge 2008 ~ 2011: 過去のTTS国際コンペ
  - SOMOS [Maniati+22]: ニューラルTTSのみ
  - sarulab-data: 我々が独自に収集した 50%🔍BVCC
    - Prolific でMOSテストを実施し, 304名の評価者を募集
      (各合成音声サンプルを異なる8名が評価)

# UTMOSv2の3つの工夫



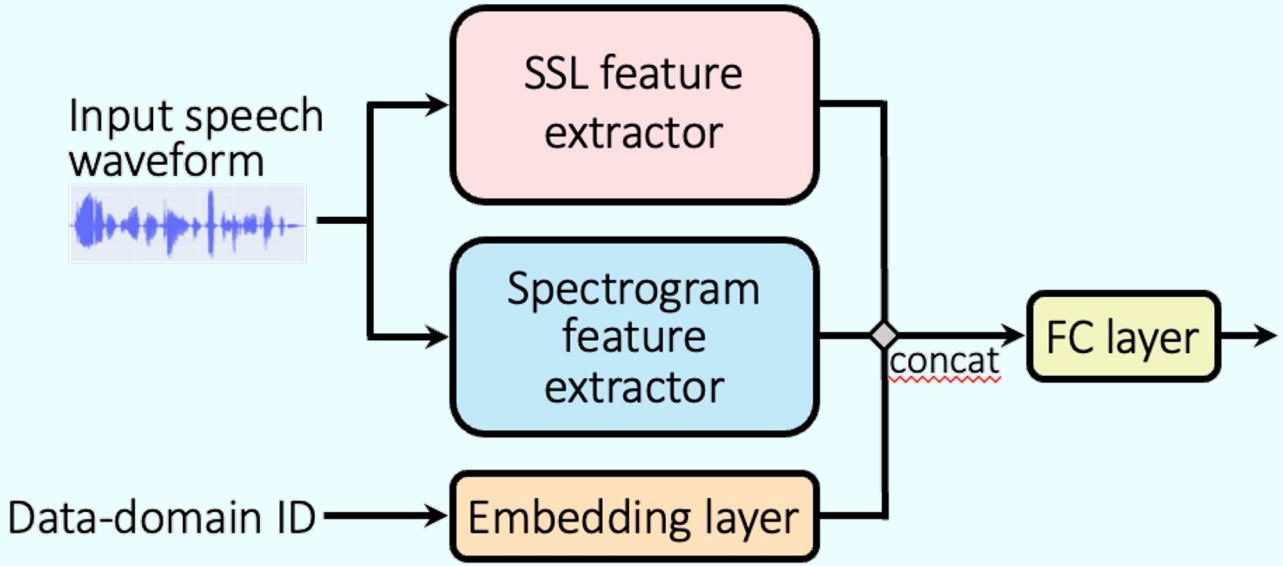
1. SSL/スペクトログラム特徴量のfusion
2. データドメインでの条件付け
3. モデルの多段階学習



# ② モデルの多段階学習



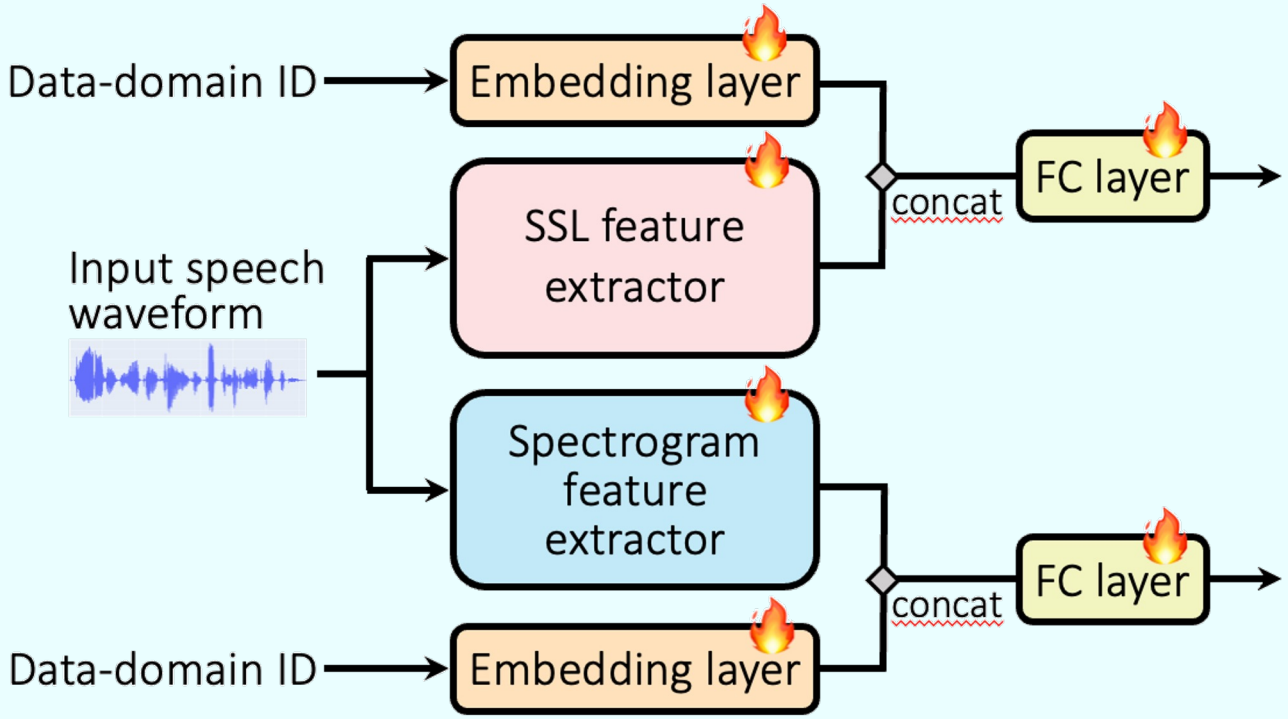
- 各モジュールをMOS予測に最適化し精度向上を狙う
- 損失関数： MSE loss と contrastive loss [Saeki+22] の重み付き和



# ② モデルの多段階学習



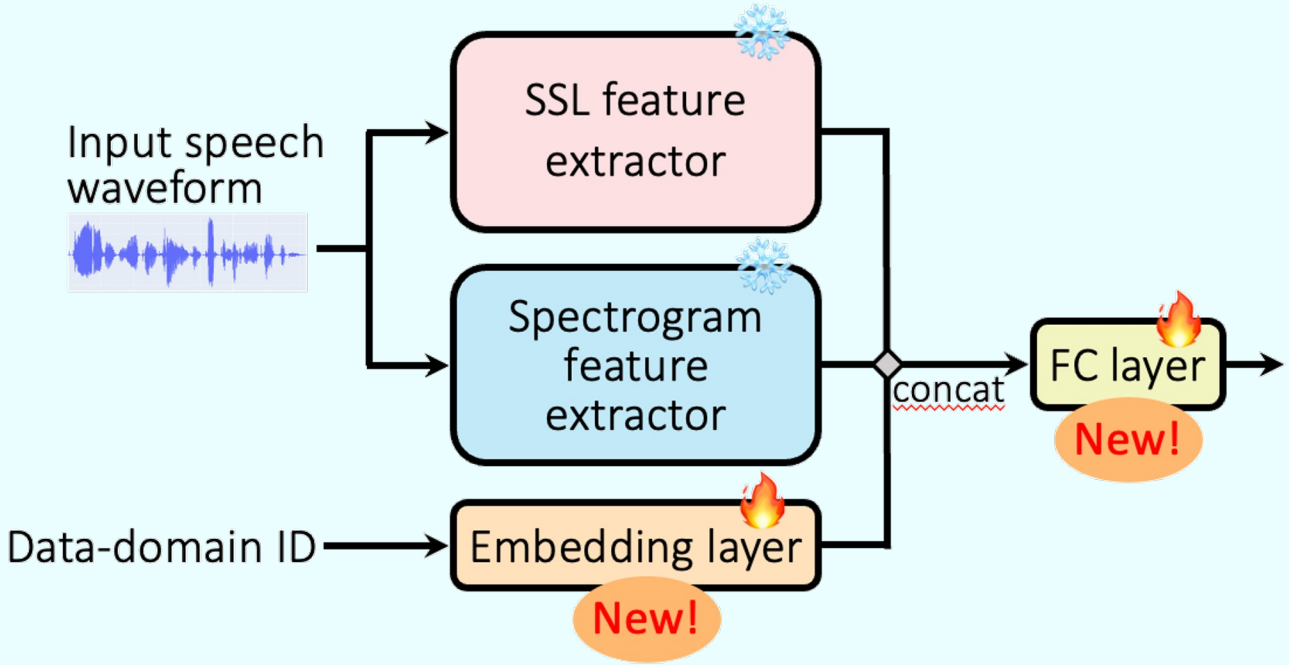
- Stage #1：各特徴量抽出器を個別に学習
  - 各特徴量抽出器ごとにデータドメイン埋め込み層とFC層を用意



# ② モデルの多段階学習



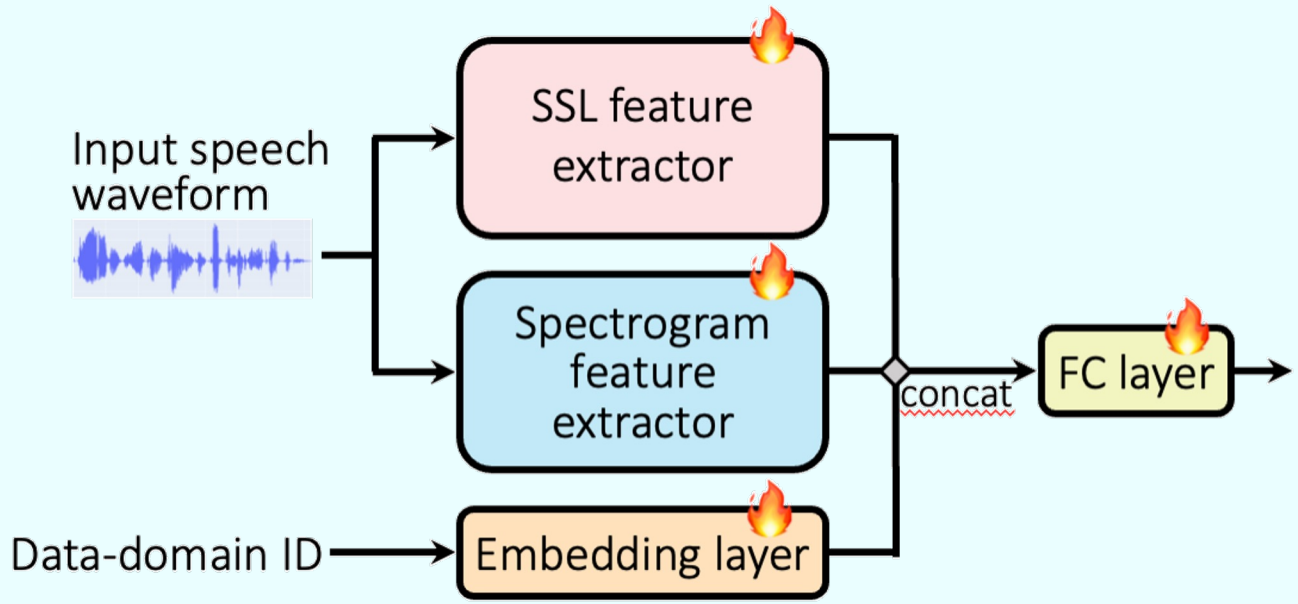
- Stage #2：特徴量 fusion 層の学習
  - Stage #1 で学習した特徴量抽出器のパラメータを固定
  - 新たに用意した fusion 用のデータ埋め込み層とFC層を学習



# ② モデルの多段階学習



- Stage #3：モデル全体の fine-tuning
  - モデル全体を小さい学習率・少ないエポック数で fine-tuning



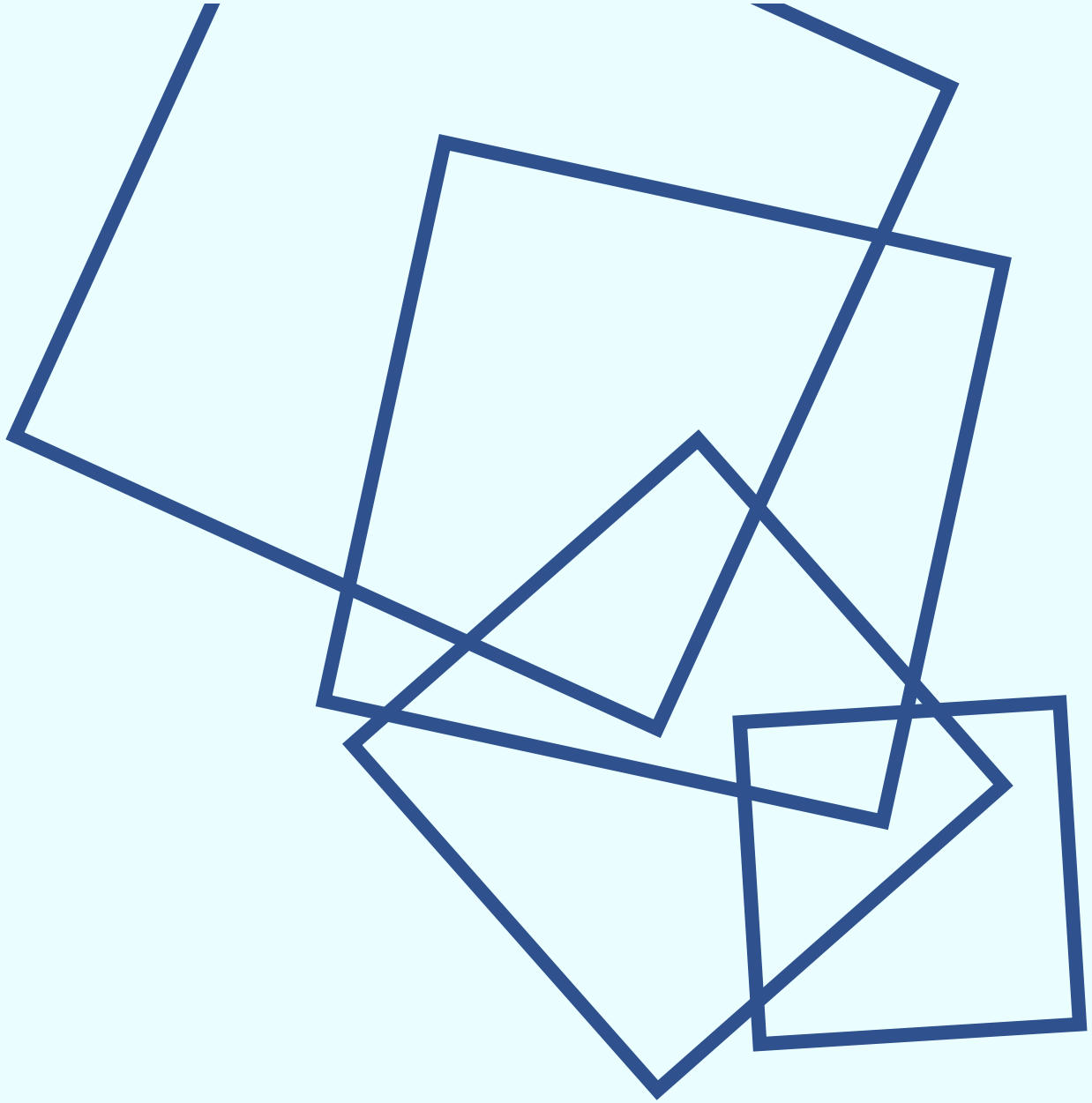
# Section3

## 実験的評価 *Experiments*

01. 研究概要

02. UTMOSv2

03. 実験的評価



# 共通の実験条件（詳細は原稿参照）



| | |
|---|---|
| 特徴量抽出器 | SSL: wav2vec 2.0 base [Baevski+20] (事前学習 w/ LibriSpeech [Panayotov+15])<br>Spec.: EfficientNetV2 [Tan+21] (事前学習 w/ ImageNet [Deng+09]) |
| 最適化アルゴリズム | AdamW [Loshchilov+19] w/ cosine annealing [Loschchilov+17] |
| 学習率 | 多段階学習の各 Stage ごとに調整 |
| データ拡張 | Mixup [Zhang+18] |

# 実験1：特徴量fusionに関する評価



- UTMOSv2 の特徴量抽出器を1つだけ使うケースとの比較
  - Baseline: B01 (SSL-MOS [Cooper+22]) と UTMOS [Saeki+22]

| | Utterance-level | | System-level | |
|---|---|---|---|---|
| | MSE ↓ | SRCC ↑ | MSE ↓ | SRCC ↑ |
| UTMOSv2 | 0.459 | 0.579 | 0.288 | 0.854 |
| w/o SSL | 0.357 | 0.516 | 0.188 | 0.770 |
| w/o spec. | 0.673 | 0.529 | 0.497 | 0.793 |
| B01 | 0.741 | 0.417 | 0.589 | 0.609 |
| UTMOS | 0.541 | 0.300 | 0.378 | 0.367 |

# 実験1：特徴量fusionに関する評価



- UTMOSv2 の特徴量抽出器を1つだけ使うケースとの比較
  - Baseline: B01 (SSL-MOS [Cooper+22]) と UTMOS [Saeki+22]

| | Utterance-level | | System-level | |
|---|---|---|---|---|
| | MSE ↓ | SRCC ↑ | MSE ↓ | SRCC ↑ |
| UTMOSv2 | 0.459 | 0.579 | 0.288 | 0.854 |
| w/o SSL | 0.357 | 0.516 | 0.188 | 0.770 |
| w/o spec. | 0.673 | 0.529 | 0.497 | 0.793 |
| B01 | 0.741 | 0.417 | 0.589 | 0.609 |
| UTMOS | 0.541 | 0.300 | 0.378 | 0.367 |

✔ Baseline よりも高い zoomed-in MOS 予測性能を達成

# 実験1：特徴量fusionに関する評価



- UTMOSv2 の特徴量抽出器を1つだけ使うケースとの比較
  - Baseline: B01 (SSL-MOS [Cooper+22]) と UTMOS [Saeki+22]

| | Utterance-level | | System-level | |
|---|---|---|---|---|
| | MSE ↓ | SRCC ↑ | MSE ↓ | SRCC ↑ |
| UTMOSv2 | 0.459 | 0.579 | 0.288 | 0.854 |
| w/o SSL | 0.357 | 0.516 | 0.188 | 0.770 |
| w/o spec. | 0.673 | 0.529 | 0.497 | 0.793 |
| B01 | 0.741 | 0.417 | 0.589 | 0.609 |
| UTMOS | 0.541 | 0.300 | 0.378 | 0.367 |

✓ 統合的利用 (fusion) によりSRCCが改善

※ MSEは w/o SSL が最良

# 実験2：多段階学習に関する評価



- Stage 2 を除外する場合と Stage 1&2 を除外する場合を比較
  - 即ち, 特徴量抽出器をMOS予測で事前学習することの有効性を調査

| | Utterance-level | | System-level | |
|---|---|---|---|---|
| | MSE ↓ | SRCC ↑ | MSE ↓ | SRCC ↑ |
| UTMOSv2 | 0.459 | 0.579 | 0.288 | 0.854 |
| w/o Stage 2 | 0.342 | 0.505 | 0.108 | 0.816 |
| w/o Stage 1&2 | 0.293 | 0.423 | 0.097 | 0.672 |

✓多段階学習によりSRCC改善

※MSEは劣化

# 実験3：学習データセットに関する評価



MOS 予測の推論時に指定したデータドメインID

| Train data | BVCC MSE ↓ | BVCC SRCC ↑ | BC MSE ↓ | BC SRCC ↑ | SOMOS MSE ↓ | SOMOS SRCC ↑ | -data MSE ↓ | -data SRCC ↑ |
|---|---|---|---|---|---|---|---|---|
| All datasets | 0.288 | 0.854 | 0.088 | 0.851 | 0.056 | 0.844 | 0.058 | 0.838 |
| w/o BVCC | - | - | 0.343 | 0.832 | 0.128 | 0.846 | 0.101 | 0.836 |
| w/o BC | 0.145 | 0.819 | - | - | 0.069 | 0.823 | 0.122 | 0.805 |
| w/o SOMOS | 0.224 | 0.696 | 0.221 | 0.682 | - | - | 0.221 | 0.700 |
| w/o -data | 0.282 | 0.647 | 0.102 | 0.661 | 0.186 | 0.690 | - | - |

✓ 基本的に、全データセットで学習した場合に最良

(System-level の結果のみ掲載)

# 実験3：学習データセットに関する評価



MOS 予測の推論時に指定したデータドメインID

| Train data | BVCC MSE ↓ | BVCC SRCC ↑ | BC MSE ↓ | BC SRCC ↑ | SOMOS MSE ↓ | SOMOS SRCC ↑ | -data MSE ↓ | -data SRCC ↑ |
|---|---|---|---|---|---|---|---|---|
| All datasets | 0.288 | 0.854 | 0.088 | 0.851 | 0.056 | 0.844 | 0.058 | 0.838 |
| w/o BVCC | - | - | 0.343 | 0.832 | 0.128 | 0.846 | 0.101 | 0.836 |
| w/o BC | 0.145 | 0.819 | - | - | 0.069 | 0.823 | 0.122 | 0.805 |
| w/o SOMOS | 0.224 | 0.696 | 0.221 | 0.682 | - | - | 0.221 | 0.700 |
| w/o -data | 0.282 | 0.647 | 0.102 | 0.661 | 0.186 | 0.690 | - | - |

⚠SOMOS or -dataを除外した場合性能が劣化

(System-level の結果のみ掲載)

# 実験3：学習データセットに関する評価



MOS 予測の推論時に指定したデータドメインID

| Train data | BVCC MSE ↓ | BVCC SRCC ↑ | BC MSE ↓ | BC SRCC ↑ | SOMOS MSE ↓ | SOMOS SRCC ↑ | -data MSE ↓ | -data SRCC ↑ |
|---|---|---|---|---|---|---|---|---|
| All datasets | 0.288 | 0.854 | 0.088 | 0.851 | 0.056 | 0.844 | 0.058 | 0.838 |
| w/o BVCC | - | - | 0.343 | 0.832 | 0.128 | 0.846 | 0.101 | 0.836 |
| w/o BC | 0.145 | 0.819 | - | - | 0.069 | 0.823 | 0.122 | 0.805 |
| w/o SOMOS | 0.224 | 0.696 | 0.221 | 0.682 | - | - | 0.221 | 0.700 |
| w/o -data | 0.282 | 0.647 | 0.102 | 0.661 | 0.186 | 0.690 | - | - |

(System-level の結果のみ掲載)

MSEについて:
SOMOS寄与大

# まとめ



- 本発表：UTMOSv2 の紹介　https://github.com/sarulab-speech/UTMOSv2
  - 高品質合成音声の予測に向けた新たなMOS予測システム
  - SSL/スペクトログラム特徴量の統合的利用 + 多段階学習
  - VoiceMOS Challenge 2024 Track 1 の
    7/16評価指標で1位，残り9項目で2位を達成🏆✨
- 実験的評価結果から得られた知見
  - 特徴量 fusion はSRCCの改善に有効
  - 多段階学習によりSRCCが向上（ただしMSEは劣化）
  - 高品質な合成音声のMOS値を正確に予測したい場合, 以下が重要
    - (1) 低品質な音声合成システムを可能な限り除外
    - (2) なるべく最先端の音声合成システムを含むように学習セットを構成
- 今後: 自然性以外の評価への拡張 & モデルの軽量化を検討