音声品質評価技術の進展と展望

安田裕介（国立情報学研究所）
2025年10月SP/SLP研究会

目次

1. 主観評価（20min）
   a. MOS評価とその問題点
   b. 主観比較評価
   c. 主観比較評価の応用
2. 自動品質評価（10min）
   a. MOS自動予測
   b. 比較自動予測
3. 大規模言語モデルのための主観・自動評価（10min）
   a. 強化学習による人間フィードバック（RLHF）
   b. 報酬モデルとしての自動品質評価
   c. 直接選好最適化（DPO）
   d. 音声言語モデルへの応用

- 比較評価を軸として音声品質評価を解説します。
- 音声品質評価の全体像の把握には、以下の文献を推奨：

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda and Junichi Yamagishi: A review on subjective and objective evaluation of synthetic speech, Acoust. Sci. & Tech. 45, 4 (2024)

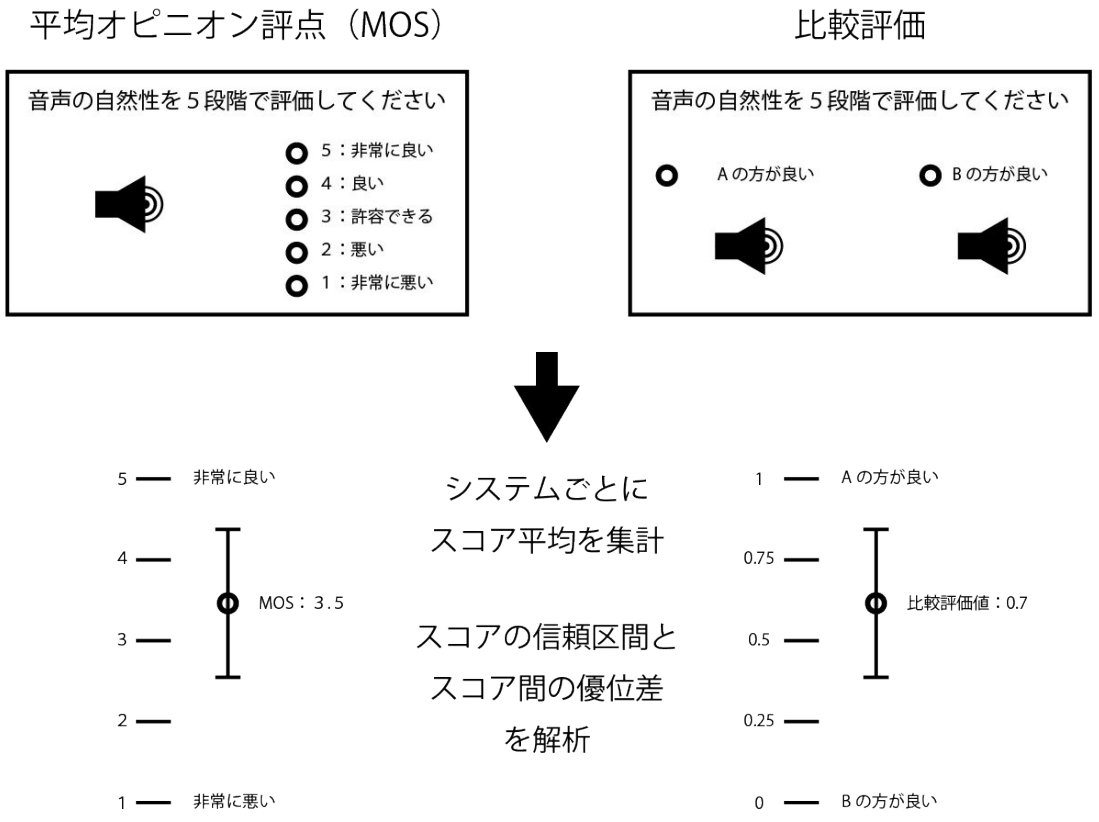
主観評価

- 音声の主観評価法
- MOS評価とその問題点
- 主観比較評価
- 主観比較評価の応用
  - ソートによる動的比較評価
  - マージに基づく継続可能評価
  - トーナメントに基づくアノテーション

この節の内容は以下の解説原稿も参照:
安田 裕介, 戸田 智基, "音声のMOS評価法の限界と大規模比較評価の新しい可能性 ," 日本音響学会誌, Vol. 80, No. 7, pp. 393-400, Aug. 2024.

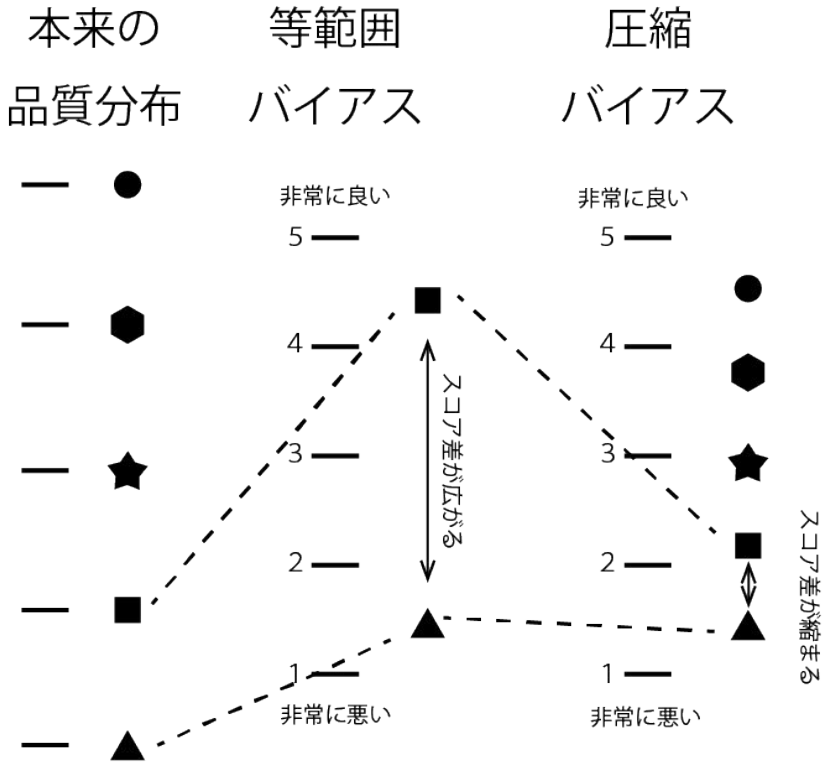
音声の主観評価法

- 主に、平均オピニオン評点（MOS）評価法と比較評価法がある
  - 他にもMUSHRAなど
- 評価の仕方とスコアが異なる
  - MOSは1つのサンプルに5段階スコアをつける
  - 比較評価は2つのサンプルの良い方を選ぶ
- スコアはシステムごとに平均化し、統計解析を行う



MOS評価とその問題点（1）

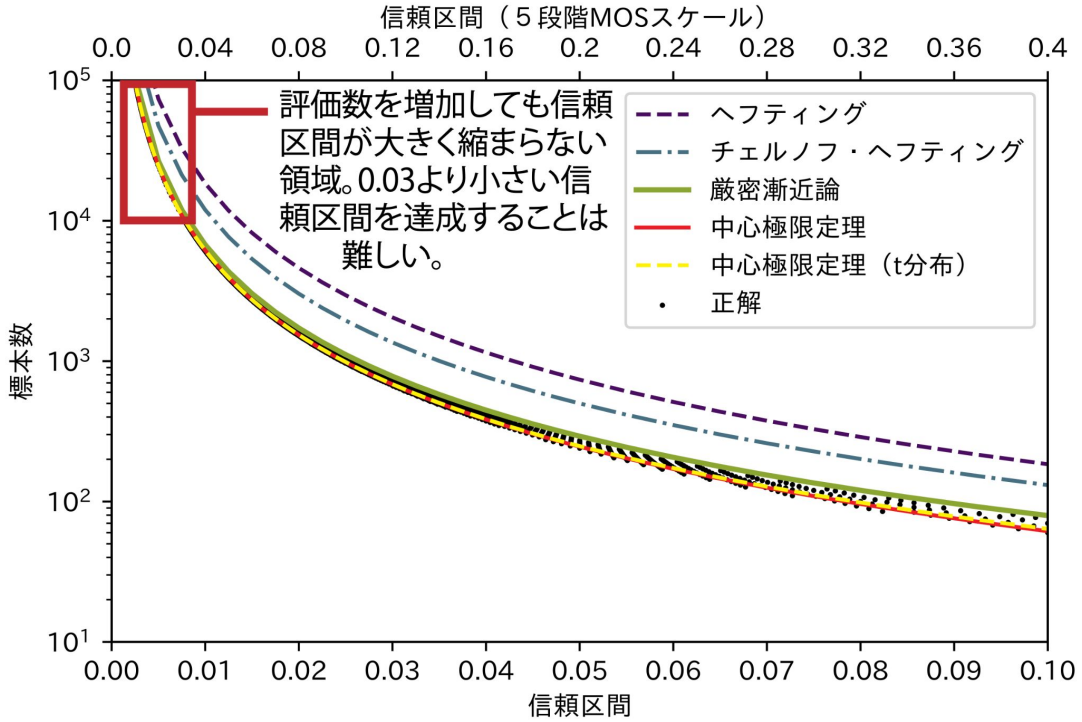
- 絶対評価値であるものの、**実際は相対的に評価**される
- 様々な**バイアス**がある
  - 等範囲バイアス：本来のスコア差よりも広がる
  - 圧縮バイアス：本来のスコア差よりも縮まる
- スコアに**一貫性がない**
  - 何と一緒に評価するかに左右される
  - 異なる実験間のスコアの比較が不可能



S. Zielinski, F. Rumsey, and S. Bech, “On some biases encountered in modern audio quality listening tests- a review,” J. Audio Eng. Soc, vol. 56, no. 6, pp. 427–451, 2008.

MOS評価とその問題点（2）

- 微妙な品質差の評価に向かない
- →高品質な生成モデルの時代において限界が目立ってきた
  - MOSは品質差を間接的に評価するので、スコア差が出にくい
  - 信頼区間は評価数を増やすと縮まるが、限界がある
  - 近年の音声合成サンプルの品質差が信頼区間の限界に迫っている
  - 圧縮バイアスや評価数不足で、大規模MOSコーパスでの統計的有意差が乏しい



Y. Yasuda and T. Toda, “Analysis of Mean Opinion Scores in Subjective Evaluation of Synthetic Speech Based on Tail Probabilities,” in Proc. INTERSPEECH 2023, 2023, pp. 5491–5495.

古くて新しい評価法としての比較評価：なぜ比較評価か？

- **類似品質の識別が可能**：２つの対象の優劣を直接評価し、品質差の識別に優れ、より小さな評価数での評価が期待できる。
- **評価値の一貫性**：相対評価のため、他の評価対象の存在に影響されない。異なる実験の評価結果を混合可能。
- **バイアスの少なさ**：評価スケールを用いないため、被験者の解釈や他の評価対象の存在に影響されず、より評価が容易。
- **確立された数理モデル**：BTLモデルやPlacket-Luceモデルなど、比較確率やランキング確率の数理モデルによって、結果をモデル化できる。
- **統計解析の容易さ**：より厳密な信頼区間や統計解析法が利用できる。
- **モデルや条件の最適化への応用可能性**：どちらが良いかという、改善の方向を得られるため、モデルの最適化や最適条件の探索に応用可能。

比較評価の弱点

- **評価ペアの組み合わせの膨大さ**：比較評価の評価ペアの組み合わせは、評価対象の数に対して平方に増加し、大規模評価に向いていない。
  - →オンライン学習や能動学習で解決
- **評価値の解釈の難しさ**：比較評価のスコアは相対的な品質のため、MOSのような絶対的な品質スコアのように解釈が容易ではない。
  - →レイティング理論や効用理論で解決
- →解決法を適用すれば、**比較評価は強力な評価法** となる。

この発表で扱う比較評価の技術

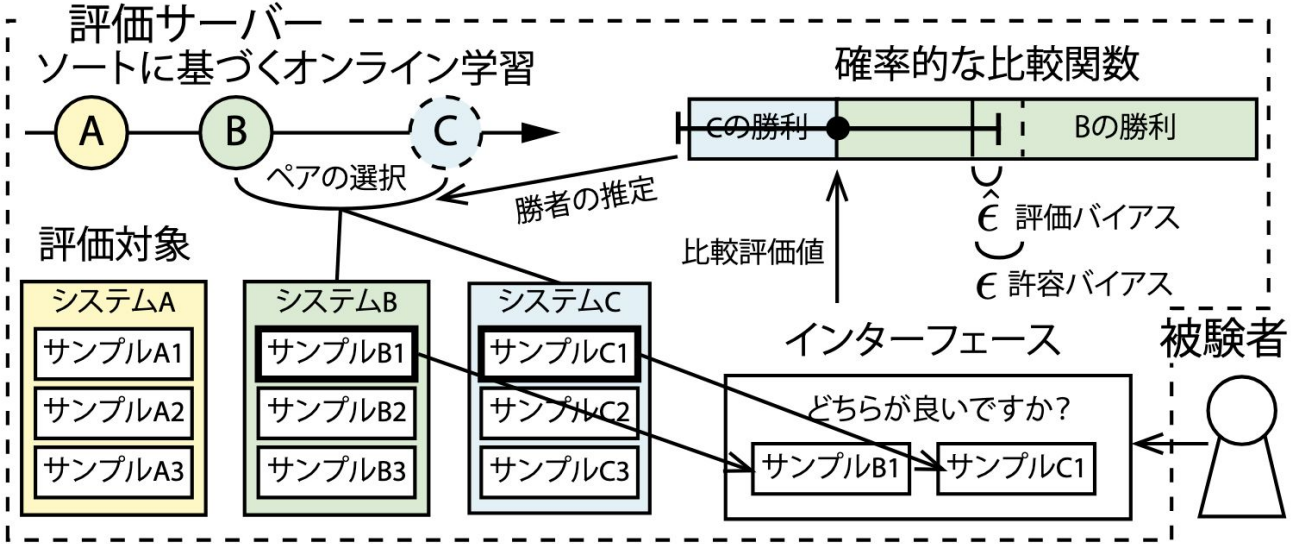
- **評価ペアの組み合わせの膨大さの解決:**
  - ソートに基づくオンライン学習によるペアを動的に選択する比較評価
- **評価値の一貫性を利用した継続可能評価:**
  - マージに基づくオンライン学習による複数の実験結果を統合する継続可能評価
- **評価値の解釈の難しさの解決:**
  - レイティング理論の基礎と自動品質評価への応用
  - 効用理論の基礎と自動品質評価への応用
- **モデルや条件の最適化への応用:**
  - (最適条件探索)トーナメントに基づくオンライン学習による正解ラベルのアノテーション
  - (モデル最適化)大規模言語モデルの人間フィードバックによる人間アラインメント学習

評価ペアの組み合わせの膨大さの解決：ソートに基づくオンライン学習によるペアを動的に選択する比較評価

- **目的**：比較評価によって評価対象のランキングを求める
- **解決する課題**：大規模評価では全ペアの組み合わせが大きすぎる
- **アイデア**：ソートアルゴリズムに基づいて評価するペアを選べば、少ないペア数の評価のみでランキングを求められる

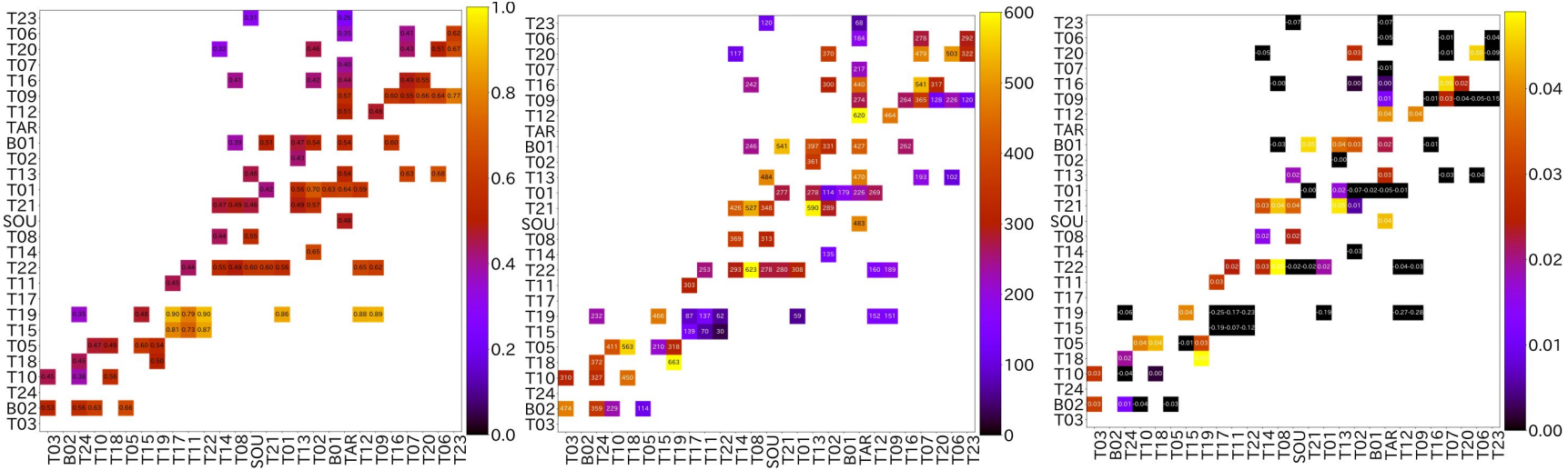
主観比較評価とその応用：ソートによる動的比較評価

- 比較評価のペアの組み合わせは膨大
- ランキングが目的なら、全ペアの組み合わせの評価は不要
- ソートに基づくオンライン学習を用いて、評価ペアを決定



Yusuke Yasuda, Tomoki Toda, Automatic design optimization of preference-based subjective evaluation with online learning in crowdsourcing environment, Computer Speech & Language, Volume 96, 2026

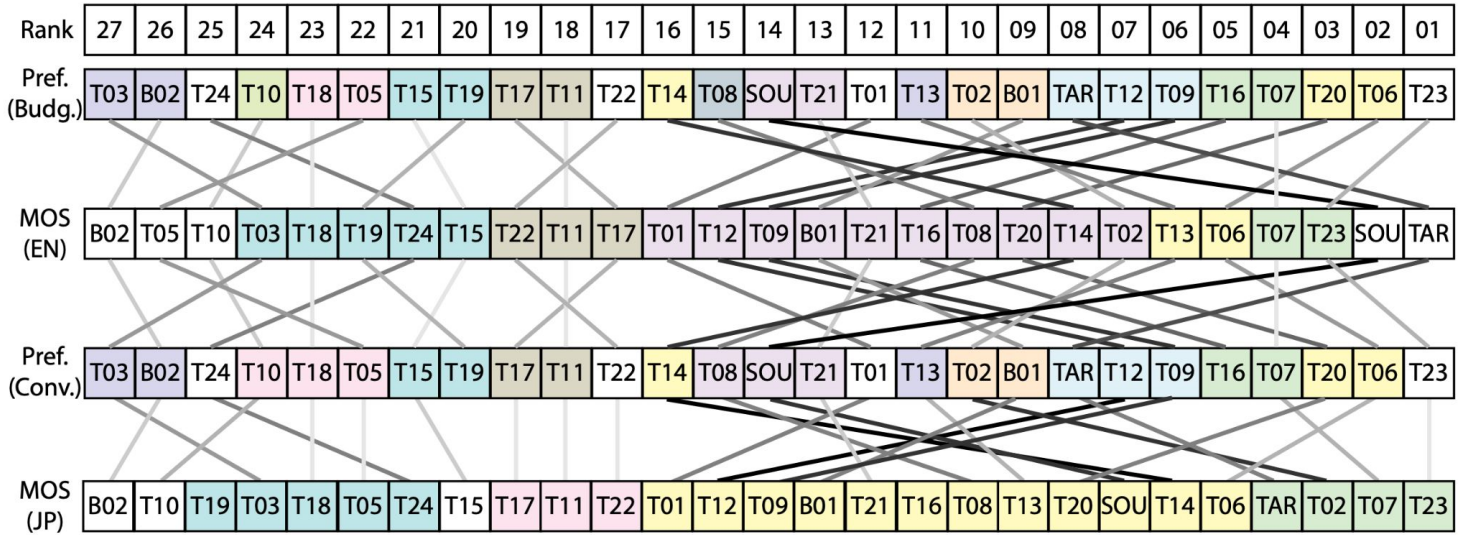
ソートによる動的比較評価の特徴：評価結果



(a) Preference score $\bar{p}_{ij}$　(b) Number of evaluations, $\bar{r}_{ij}$　(c) Error bias $\hat{\epsilon}_H$

- 品質の似たペアのみを評価する
  - 全組み合わせ381ペア→81ペアのみ評価
- 品質差のあるペアの評価数は少ない
- 品質が近いペアの評価数は多い
- 評価誤差は目標値以下

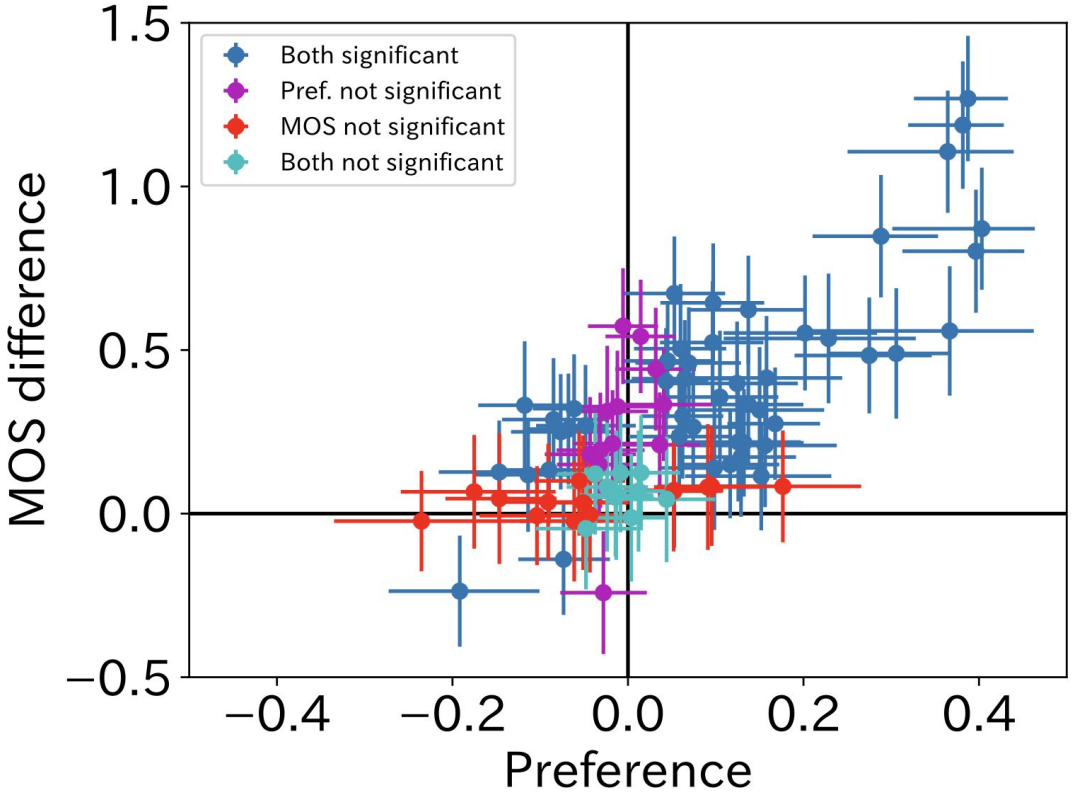
ソートによる動的比較評価の特徴：MOSとの比較（1）



- MOSと似たようなランキングが得られる
- MOSには隣接ペアに統計的有意差がない大きな領域がある（圧縮バイアス）
- 比較評価には、そのような領域がない

ソートによる動的比較評価の特徴：MOSとの比較（2）

- MOS差と比較評価値には相関がある。
- 評価数あたりの統計的有意ペア数では、比較評価よりMOSの方がまだ効率が良い。

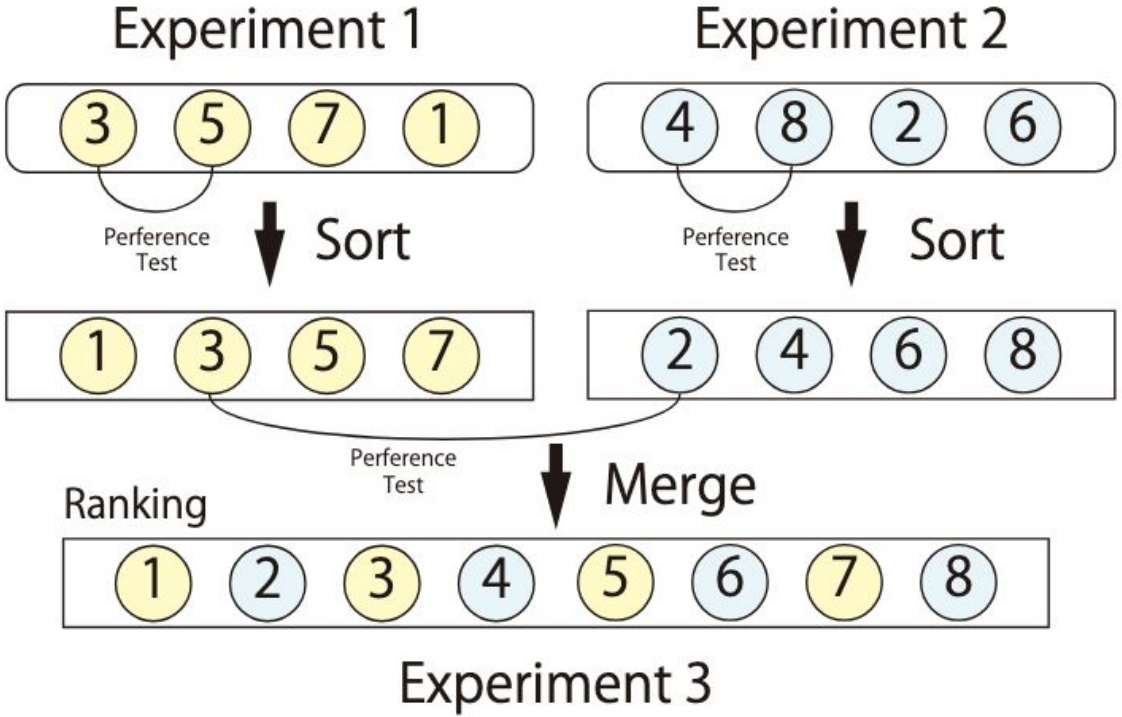


評価値の一貫性を利用した継続可能評価：マージに基づくオンライン学習による複数の実験結果を統合する継続可能評価

- **目的**：異なる複数の比較評価実験の結果を単一のランキングに統合し、段階的に主観評価コーパスを構築する。
- **解決する課題**：自動品質評価の学習コーパスとして用いられるMOSコーパスは、一度の実験で評価する必要があり、データの規模に限界がある。
- **アイデア**：マージアルゴリズムのオンライン学習を用いれば、複数のソート済みの比較評価実験のランキングを、1つのランキングに統合できる。

主観比較評価とその応用：継続的主観評価

- MOSは評価内容に依存するので、異なる実験の結果を混ぜ合わせることはできない
  - コーパス構築に不向き
- 比較評価であれば、文脈非依存なので、異なる実験を統合できる
- ソートとマージアルゴリズムを組み合わせて、異なる実験を1つのランキングに統合する継続可能評価を実現



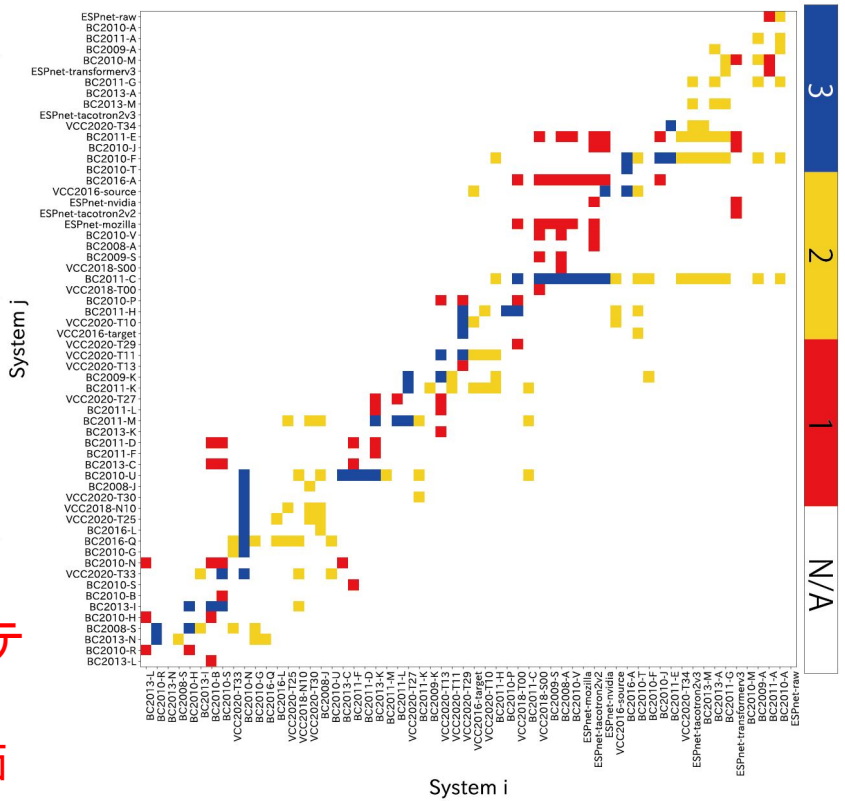
Yasuda, Y., Yamagishi, J., Toda, T. (2025) Continual Subjective Evaluation Method of Speech by Merging Sort-based Preference Tests Towards Ever-Expanding Corpus of Human Ratings. Proc. 13th edition of the Speech Synthesis Workshop, 14-20

継続的主観評価の特徴(1):実験の設計と結果

Experiment No.	1	2	3
Sort Algorithm	Insert Rank	Merge Rank	-
Merge Algorithm	-	Merge	Merge
#Sort Systems	30	30	-
#Merge Systems	-	10	50
#Scores in Budget	24,960	24,960	15,540
#Convergence Cost	14,977	19,658	9,761
#Evaluated Pairs	70	98	48
#Significant Pairs	28	49	21
#Max Cost per Pair	528	413	465
#Min Cost per Pair	219	60	127

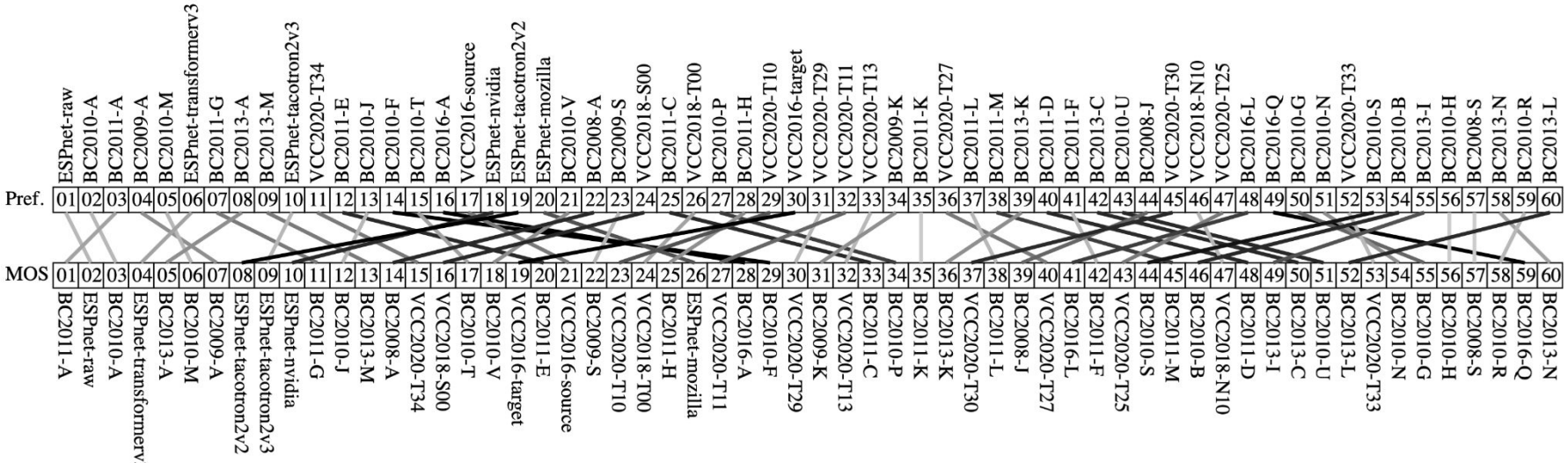
Table 1: *Settings and results of three experiments.*

- 3つの実験(2つのソートと1つのマージ)で60システムを1つのランキングにマージ
- 全組み合わせ1770ペアのうち216ペアのみを評価



実験番号ごとの評価ペアの分布

継続的主観評価の特徴(2):ランキング



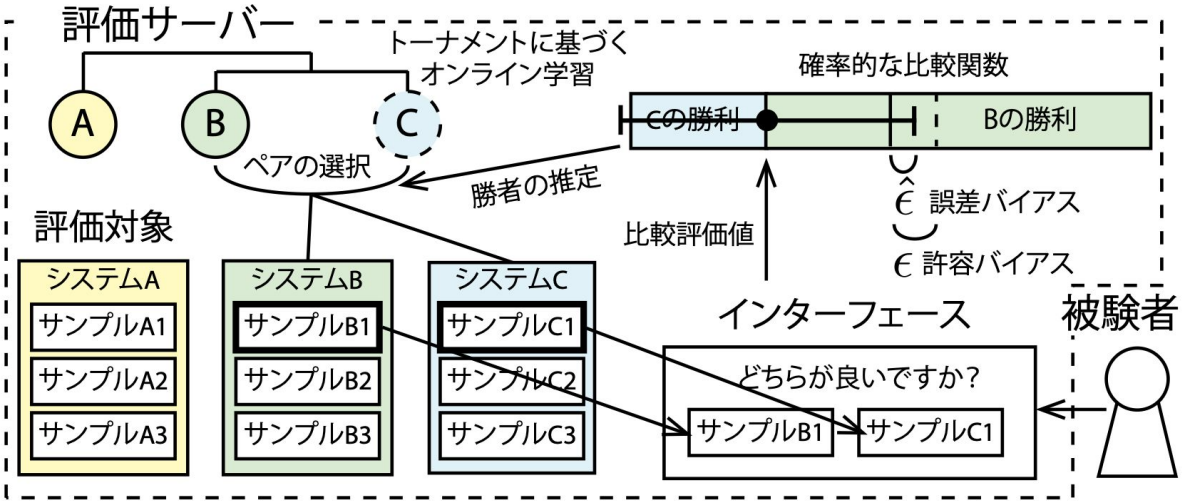
- MOSと類似のランキングが得られる
  - Spearman相関係数:0.943
  - Kendall's tau係数:0.798
- MOSとは異なり**分割して評価が可能**なので、**より詳細な評価**が可能
  - MOSよりも統計的有意なペア数が多い
  - 圧縮バイアスの回避

最適条件探索への応用：
トーナメントに基づくオンライン学習による正解ラベルのアノテーション

- **目的**：アクセントのアノテーションのコストを下げ、かつ一般の被験者によるアノテーションを可能にし、大規模に正解ラベルを収集するする。
- **解決する課題**：アクセントのアノテーションには言語の専門家が必要で、コストと時間がかかりスケールしない。
- **アイデア**：
  - アクセント制御が可能な音声合成器によって全アクセントパターンのサンプルを合成し、その中から最も良いアクセントを評価することで、正解アクセントラベルを探す。
  - アノテーションを比較評価という単純なタスクに変換 し、専門家への依存性をなくす。
  - トーナメントアルゴリズムによって最小の比較ペア数 で正解ラベルを推定する。

主観比較評価とその応用:トーナメントによるアノテーション

- 目的は、主観評価をもとに、正解ラベルを見つけること
- 最も良いサンプルが、正解ラベル
- 全アクセントパターンを合成したサンプルをトーナメントアルゴリズムで比較評価することで、効率的に正解ラベルを見つける

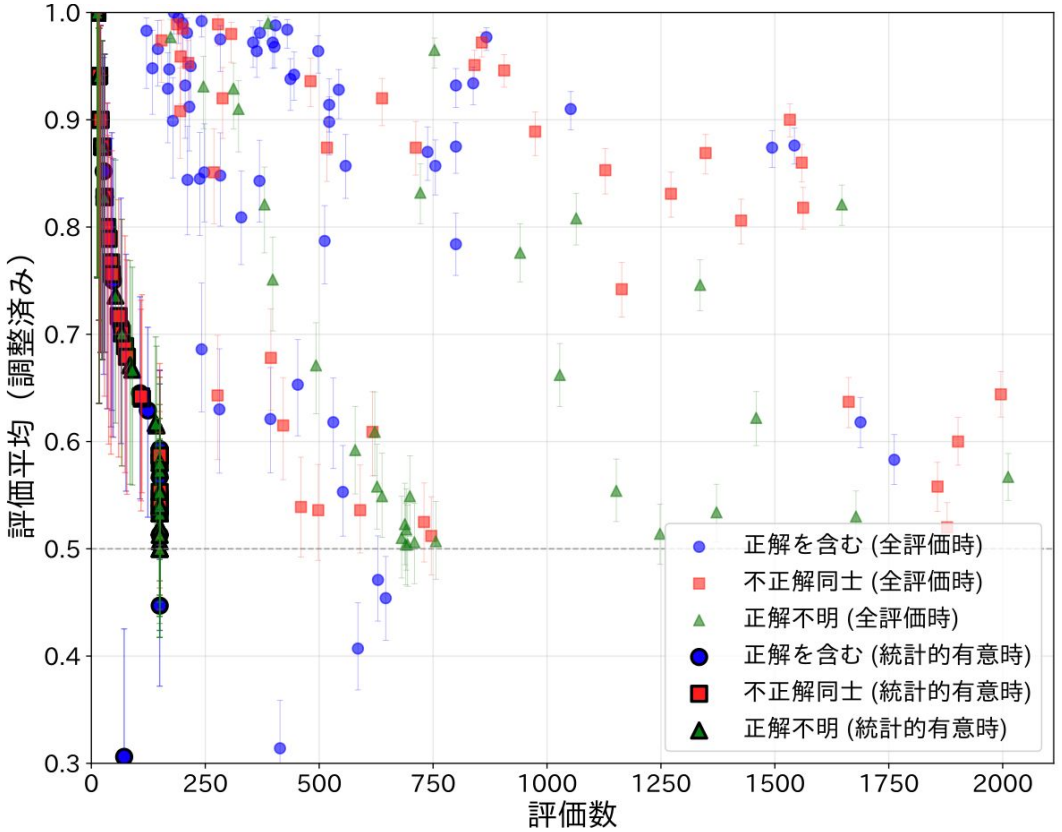


- 音声の優劣を聞いて評価することは、言語学に基づく解析よりも容易
- 非専門家でも評価が可能なため、クラウドソーシングで大規模にアノテーションが可能

安田 裕介, 井本 桂右, 深山 覚, 戸田 智基, "アクセント制御音声合成と主観比較評価最適化による専門家非依存アクセントアノテーション法," 音講論, 1-1-3, pp. 1123-1126, Sep. 2025.

トーナメントによるアノテーションの特徴

- ペアの正解アクセント型の推定精度：90%以上の高精度
  - 全評価数を用いた場合 93.1%
  - 統計的有意が確定した評価数の時点：96.6%
- 評価数が少なくても正しいアクセントを推定
  - 最小の評価数の場合：わずか13回の評価数で正解アクセント型を推定
  - 被験者は正解アクセントに強い選好を示した
  - 不正解アクセント同士の比較では、どちらが勝っても推定精度に影響なし
- 評価数を増やしたとしても，推定精度が向上するわけではなかった
  - 信頼区間は縮小するが、推定精度に寄与しない



自動品質評価

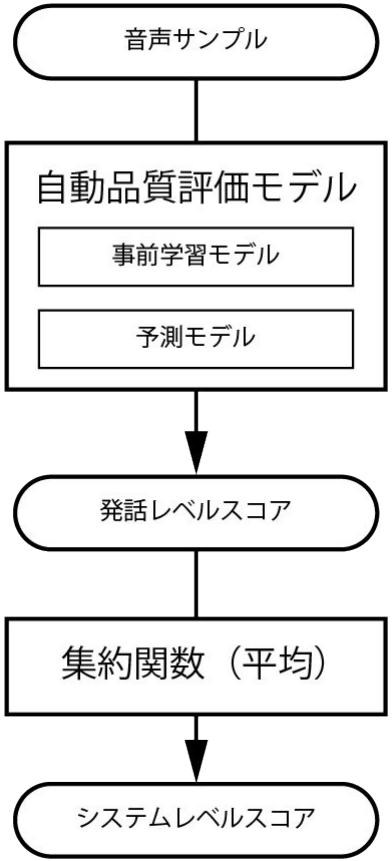
- 自動品質評価の目的
- MOS自動予測
- レイティング理論の基礎と自動品質評価への応用
- 効用理論の基礎と自動品質評価への応用
- 比較的学習基準を活用した自動品質評価

自動品質評価の目的と問題

- 自動品質評価の目的:
  - 評価のコスト(費用や時間)をゼロ にする
  - 大量の評価対象を評価する
- 自動品質評価の問題:
  - 予測値の信頼性が限定的
    - 教師ラベル自体にノイズが多い
  - 適用できないドメインがある
    - 汎化性能が低いので、学習データのドメイン外データに適用しにくい

MOSを予測する自動品質評価の基礎

- 音声サンプルからMOSを予測する
- 予測の単位には発話レベルとシステムレベルがある
  - システムレベルスコアは発話レベルスコアを集約（平均値の計算）することで求めらる
- 事前学習モデルを用いて特徴抽出し、予測モデルを用いてスコアを予測する
  - 事前学習モデルを用いることが予測精度に重要



比較評価値の解釈の難しさの解決：
レイティング理論の基礎と自動品質評価への応用

- **目的**：MOS自動予測と同じ手法で比較評価値を予測できるようにする。
- **解決する課題**：比較評価値は相対スコアであり、MOSのような絶対スコアのように容易な解釈や扱いができない。
- **アイデア**：比較評価データから絶対的なスコアを導出するためにレイティング理論を適用する。システムレベルのスコアが計算できるので、MOS自動予測と同じ手法で学習・推論できる。

例：Colleyレイティング（1）

- レイティング：チーム（音声品質評価ではシステムに相当）の勝敗から計算できるチームの強さの指標
- Colleyの手法は、勝率に基づき、レイティングを計算する。
- 特徴として保存特性がある：
  - すべてのチームの初期レイティングは 1/2から始まる
  - 全レイティングの平均値は常に 1/2である。
  - あるチームのレイティングが上がれば、他のチームのレイティングは下がる。
  - レイティングが高いチームが低いチームに負けると、レイティングが大きく下がる。
- 保存特性を満たすために、勝率の計算方法を修正する：

通常の勝率

$$r_i = \frac{w_i}{t_i}$$

修正した勝率

$$r_i = \frac{1+w_i}{2+t_i}$$

r_i：チームiのレイティング

w_i：チームiの勝利した試合の数

l_i：チームiの敗北した試合の数

t_i：チームiの参戦した試合の合計数

この修正は、n回の試行でs回成功したとき、次の回が成功する確率が$\frac{s+1}{n+2}$となるという、Laplaceの継起の法則に由来する。

例：Colleyレイティング（2）

- 修正した勝率に基づくColleyレイティングは、次の線形方程式を解くことで計算できる：

$$\mathbf{Cr} = \mathbf{b}$$

$\mathbf{r_{n\times 1}}$：Colleyのレイティングベクトル

$\mathbf{b_{n\times 1}}$：$b_i = 1 + \frac{1}{2}(w_i - l_i)$として定義される右辺ベクトル

n_{ij}：チームiとチームjがお互いに対戦した回数

$\mathbf{C_{n\times n}}$：以下で定義されるColleyの係数行列

$$C_{ij} = \begin{cases} 2 + t_i & i = j \\ -n_{ij} & i \neq j \end{cases}$$

r_i：チームiのレイティング

w_i：チームiの勝利した試合の数

l_i：チームiの敗北した試合の数

導出：

（1）チームiの勝利数を分解

$$w_i = \frac{w_i - l_i}{2} + \frac{w_i + l_i}{2}$$
$$= \frac{w_i - l_i}{2} + \frac{t_i}{2}$$
$$= \frac{w_i - l_i}{2} + \sum_{j=1}^{t_i} \frac{1}{2}$$

（2）保存特性により、$1/2$をレイティングの平均に置き換え

$$w_i \approx \frac{w_i - l_i}{2} + \sum_{j \in O_i} r_j$$

（3）等式とみなしてColleyの手法の式に代入し、線型方程式を得る

$$r_i = \frac{1 + w_i}{2 + t_i}$$
$$= \frac{1 + \frac{w_i - l_i}{2} + \sum_{j \in O_i} r_j}{2 + t_i}$$

比較評価値の解釈の難しさの解決:
効用理論の基礎と自動品質評価への応用

- **目的**: 比較評価データを直接自動品質モデルの学習に利用する。
- **解決する課題**: 比較評価値はペアに対するスコアであるため、比較評価データから直接学習するためには、モデルの学習基準もペアに対するものにする必要がある。一方、自動品質予測の推論時は、単一のサンプルのスコアを計算したいため、単一のサンプルのスコアを出力する必要がある。
- **アイデア**: 比較確率が絶対的な品質である効用の差から生じると考える効用理論を用いる。選好確率をBTLモデルでモデル化し、効用をモデルで予測する。

BTLモデルによる比較確率のモデル化と効用の予測

- 比較確率はBTLモデルでモデル化できる。
- BTLモデルでは、**比較確率は効用（品質）の比**になる。
- 自動品質予測モデルはサンプルの**効用を予測**する。
- 比較評価値を学習ラベルとして用い、**尤度最大化**を用いて最適化できる。
  - MOS予測のように、L1誤差が学習基準ではない。

比較確率のモデル化：

$$P(y_1 \succ y_2) = \frac{\exp(r_\phi(y_1))}{\exp(r_\phi(y_1)+r_\phi(y_2))} = \sigma(r(y_1) - r(y_2))$$

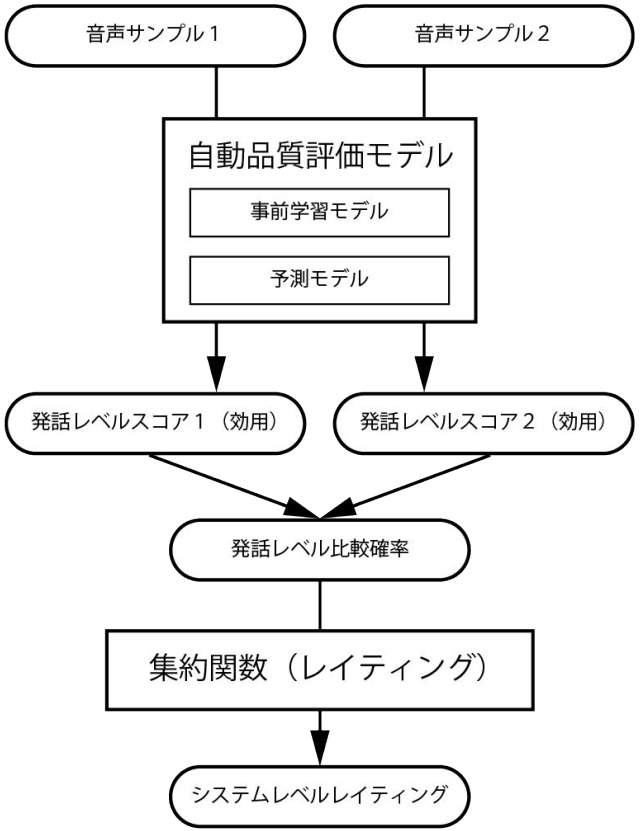
学習基準：

$$\mathcal{L}r(\phi) = -\mathbb{E}_{(y_1,y_2,y_1 \succ y_2)\sim\mathcal{D}}[\log \sigma(r_\phi(y_1) - r_\phi(y_2))]$$

- r_ϕ：パラメータϕの効用（品質）予測モデル
- $\sigma(\cdot)$：シグモイド関数
- y_1: 音声サンプル1
- y_2: 音声サンプル2
- $\mathcal{D}$: データセット。音声サンプルy_1, y_2と比較結果ラベル$y_1 \succ y_2$の３つ組$(y_1, y_2, y_1 \succ y_2)$の集合。

比較的学習基準を活用した自動品質評価

- MOS予測に比較的な学習基準を追加
- 比較確率はBTLモデルを用いてモデル化
- 比較評価ラベルはサンプルペアのMOSから計算
  - 比較評価値を学習ラベルとして用いることもできる
- MOSのバイアスに頑健な予測が可能になる
- システムレベルスコアとしてMOSの代わりにレイティングを計算可能



C.-H. Hu, Y. Yasuda, T. Toda, "Preference-based training framework for automatic speech quality assessment using deep neural network," Proc. INTERSPEECH, pp. 546-550, Dublin, Ireland, Aug. 2023.

モデル最適化への応用：
大規模言語モデル（LLM）の人間フィードバックによる
人間アラインメント学習

- **目的**：LLMの応答が協力的、誠実、無害な口調になるように、LLMの応答を人間の好みに合わせる（アラインメント）。
- **解決する課題**：人間の好む応答には明確な正解応答や、設計可能な報酬関数が存在するわけではなく、通常の教師あり学習や強化学習が使えない。
- **アイデア**：応答の候補を比較し、好ましい方を人間が選択することによって、強化学習の枠組みで人間の選好を学習する。

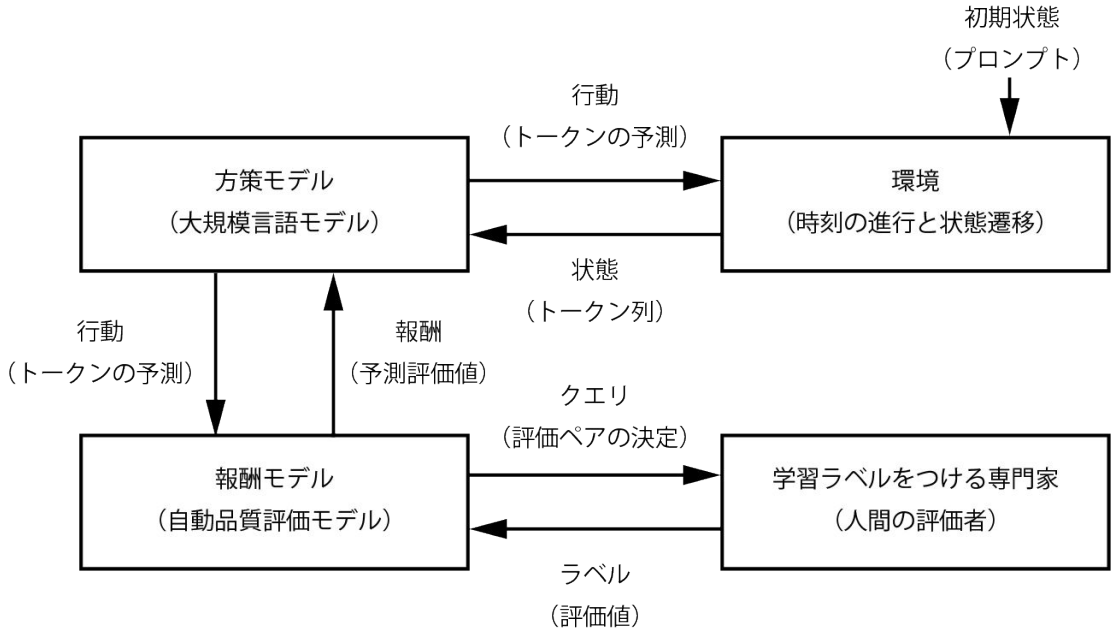
大規模言語モデルのための主観評価

- 強化学習による人間フィードバック（RLHF）
- 報酬モデルとしての自動品質評価
- 直接選好最適化（DPO）
- 音声言語モデルへの応用

強化学習による人間フィードバック（RLHF）の構成要素

RLHFの構成要素には、**音声品質評価の技術要素が登場**する：

- フィードバック
  - 主観評価の方法
- ラベル収集
  - どのペアを評価するかクエリ
- 報酬モデル学習
  - 自動品質評価の学習
- 方策モデル学習
  - 言語モデル学習



Timo Kaufmann, Paul Weng, Viktor Bengs, Eyke Hüllermeier:
A Survey of Reinforcement Learning from Human Feedback. Trans. Mach. Learn. Res. 2025 (2025)

人間フィードバック：なぜ比較評価を用いるのか？

- 強化学習では、スカラ（数値）の報酬を用いる。
- しかし、RLHFでは、MOSのようなスカラ（数値）フィードバックを用いず、比較評価を用いて、報酬モデルを学習する。
- なぜか？
  - スカラフィードバックは、人間が 一貫性をもって評価できる形式ではない 。
  - スカラフィードバックは、最適解の付近で最高点に飽和してしまうため、 改善方向の情報に乏しい 。

報酬モデル学習

- 選好確率はBTLモデルでモデル化できる。
- BTLモデルでは、比較確率は効用の比になる。
- 報酬モデルはLLMの応答の効用を予測する。
- 効用は方策学習のスカラ報酬となる

$$P(y_1 \succ y_2) = \frac{\exp(r_\phi(y_1))}{\exp(r_\phi(y_1)+r_\phi(y_2))}$$
$$= \sigma(r(y_1) - r(y_2))$$

- r_ϕ：パラメータϕの効用予測モデル
- $\sigma(\cdot)$：シグモイド関数
- y_1: 応答1
- y_2: 応答2
- $\mathcal{D}$: データセット。応答1,2と比較結果ラベルの３つ組$(y_1, y_2, y_1 \succ y_2)$の集合。

方策モデル学習：強化学習

- 代理アドバンテージ関数を最適化する。
  - アドバンテージ関数：状態sにおいて行動aをとる良さを相対的に表す関数
- 最適化基準は、LLMの離散生成により微分不可能。よって、強化学習によって最適化する。
- 報酬関数がスカラ報酬を予測するため、通常の強化学習アルゴリズムを用いることができる。
- 強化学習アルゴリズムには、例えば近接方策最適化（PPO）を用いることができる。

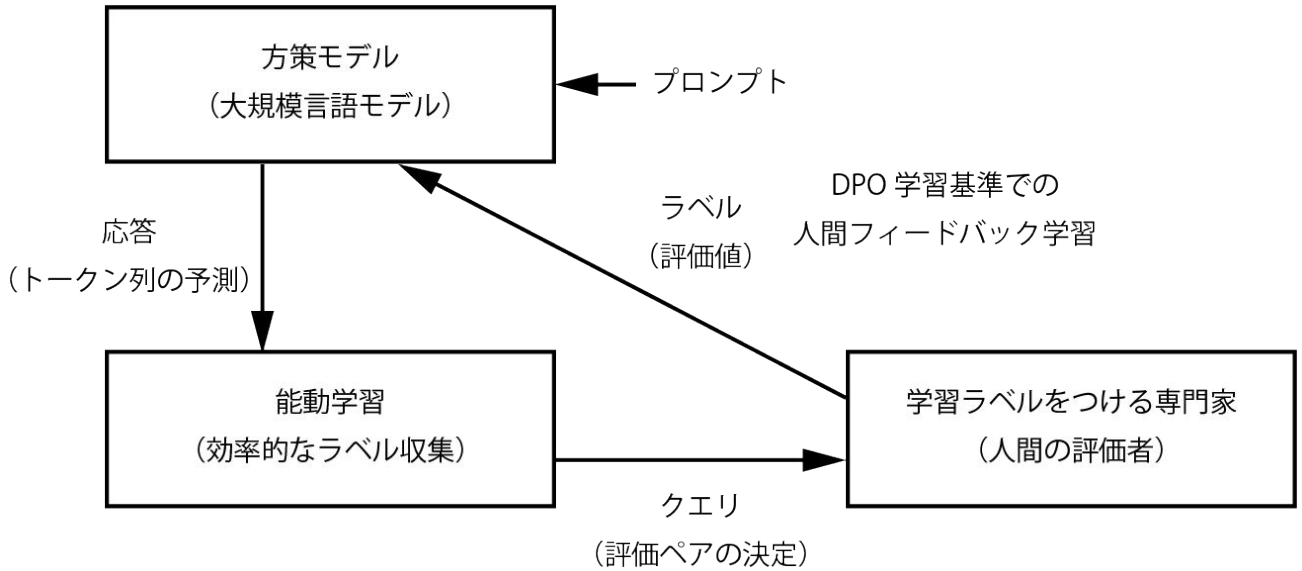
$$J_\pi(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[r_\phi(x,y) - \beta \log \frac{\pi_\theta(\cdot|x)}{\pi_{\text{ref}}(\cdot|x)}\right]$$

β：方策が参照方策から離れすぎないようにする正則化項の強さ

$\pi_{\text{ref}}(y|x)$：参照方策。FTモデル$\pi^{\text{SFT}}(y|x)$で初期化する。

方策モデル学習：直接選好最適化（DPO）

- DPOは報酬を最適方策に結びつけることによって、RLHFの目的関数を**閉じた形式で解**を求める。
- DPOでは、**報酬モデルが不要**で、かつ**強化学習も不要**で、**尤度最大化**で方策モデルを学習できる。



直接選好最適化（DPO）の洗練された点

- 最適方策の解析解で表した報酬には、分配関数Z(x)が現れ、計算は不可能である。
- 効用理論では、比較確率は効用（報酬・品質）の差の関数として与えられた。
- 2つの候補の効用（報酬・品質）の差を取ると、分配関数Z(x)が打ち消される。

報酬rが与えられたときの最適方策は以下となる。

$$\pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

最適方策の式を報酬関数に関して解くと以下になる。

$$r(x,y) = \beta\log\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta\log Z(x)$$

この報酬関数を、選好確率モデルに代入すると以下を得る。

$$P(y_w \succ y_l|x) = \frac{\exp(r(x,y_w))}{\exp(r(x,y_w)+r(x,y_l))}$$

$$= \sigma(r(x,y_w) - r(x,y_l))$$

$$= \sigma\left(\beta\log\frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

直接選好最適化（DPO）の導出（1）

以下の強化学習の目的関数から始める。

$$\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}[r_\phi(x,y)] - \beta\mathbb{D}_{\mathrm{KL}}[\pi_\theta(y|x)\|\pi_{\mathrm{ref}}(y|x)]$$

このKL正則化を施した報酬の最大化に対する最適方策は以下の形式で与えられることが分かっている。

$$\pi_r(y|x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp(\frac{1}{\beta}r(x,y))$$

$Z(x) = \sum_y \pi_{\mathrm{ref}}(y|x)\exp(\frac{1}{\beta}r(x,y))$：分配関数

$\pi_r(y|x)$：最適方策

- x：プロンプト
- y：応答
- $\pi(y|x)$：プロンプトxから応答yを予測する方策
- $r(x,y)$：報酬関数

直接選好最適化（DPO）の導出（2）

分配関数の計算は困難であるため、この式を最適方策$\pi_r(y|x)$、参照方策$\pi_{\text{ref}}(y|x)$、分配関数に関して再配置を行う。

最適方策の式を報酬関数に関して解くと以下になる。

$$\pi_r(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

$$\log \pi_r(y|x) = \log \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

$$= -\log Z(x) + \log \pi_{\text{ref}}(y|x) + \frac{1}{\beta}r(x,y)$$

$$\frac{1}{\beta}r(x,y) = \log \pi_r(y|x) + \log Z(x) - \log \pi_{\text{ref}}(y|x)$$

$$r(x,y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$$

直接選好最適化（DPO）の導出（3）

効用理論とBTLモデルの良いところは、選好確率は報酬の差に依存するという点である。よって、選好確率において分配関数は打ち消される。

この報酬関数を、選好確率モデルに代入すると以下を得る。

$$P(y_w \succ y_l|x) = \frac{\exp(r(x,y_w))}{\exp(r(x,y_w)+r(x,y_l))}$$

$$= \sigma(r(x, y_w) - r(x, y_l))$$

$$= \sigma\left(\beta \log \frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

直接選好最適化（DPO）の導出（4）

この選好確率をもとに、DPO は以下の負の対数尤度を最適化することで言語モデルの方策を直接最適化する。この目的関数は、方策のパラメータによってパラメータ化されている。

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right)\right]$$

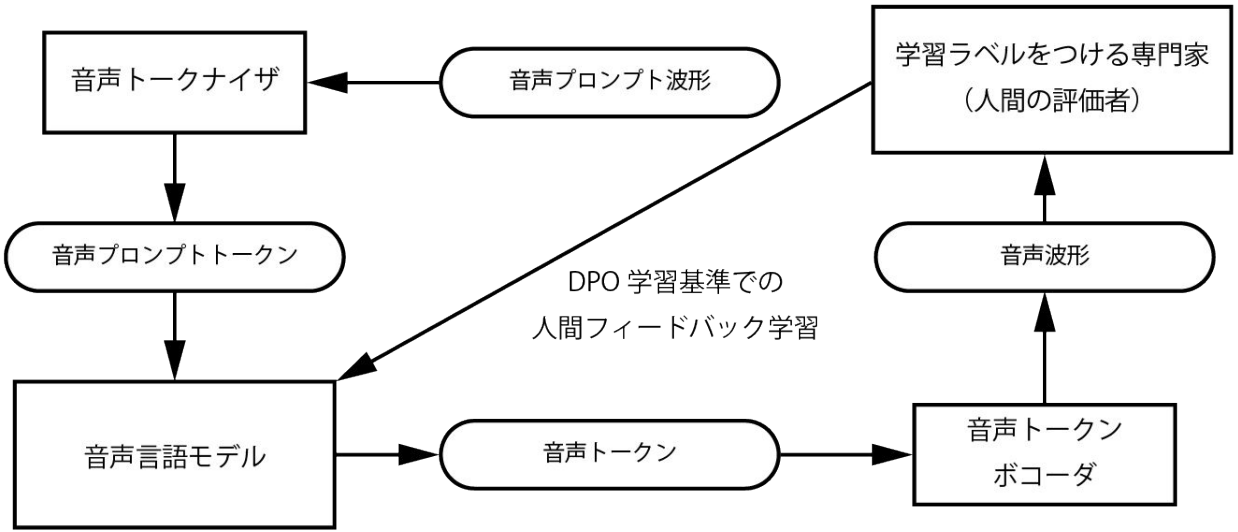
音声言語モデルにおけるRLHF

- 音声言語モデルは意味の一貫性や関連性に乏しく、冗長な応答を生成する傾向にある。
  - 音声言語モデルのトークン単位は主に音素といったセグメンタルな情報 。
  - トークン単位が小さいため、長期の意味の一貫性を保つことが困難。
  - 意味の一貫性、応答への関連性、応答の冗長さ（繰り返し）が課題
- 教師あり学習では長期の情報を学習することが困難。
- RLHFを用いて、人間による意味の一貫性等の判断をもとに、学習する。
- 言語モデルの他にも、TTSの品質改善にRLHFが用いられる例もある。

- Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, Xipeng Qiu: SpeechAlign: Aligning Speech Generation to Human Preferences. NeurIPS 2024
- Chen Chen, Yuchen Hu, Wen Wu, Helin Wang, Eng Siong Chng, Chao Zhang: Enhancing Zero-shot Text-to-Speech Synthesis with Human Feedback. CoRR abs/2406.00654 (2024)

例：音声言語モデルのRLHF

- 直接選好最適化（DPO）により、人間の評価をもとに、プロンプトに対する**応答の一貫性や関連性を改善**
- この論文では、人間の比較評価の代わりに、ASRで書き起こしてLLMで評価している
  - 音声評価はテキスト評価よりも評価者のバイアスが大きいため



Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, Ivan Bulyko:
Align-SLM: Textless Spoken Language Models with Reinforcement Learning from AI Feedback. ACL (1) 2025: 20395-20411

まとめ

- 比較評価には欠点があるが、それを克服すれば強力な評価手法となる。
  - 類似品質の識別が可能
  - 評価値の一貫性
  - バイアスの少なさ
  - 確立された数理モデル
  - 統計解析の容易さ
  - モデルや条件の最適化への応用可能性
- 比較評価のペア組み合わせ爆発はオンライン学習や能動学習で克服できる。
- 比較評価は様々な応用が可能。
  - 継続可能評価
  - アノテーション
  - 人間フィードバック学習
- 大規模言語モデルや音声言語モデルにも、比較評価は活用されている。