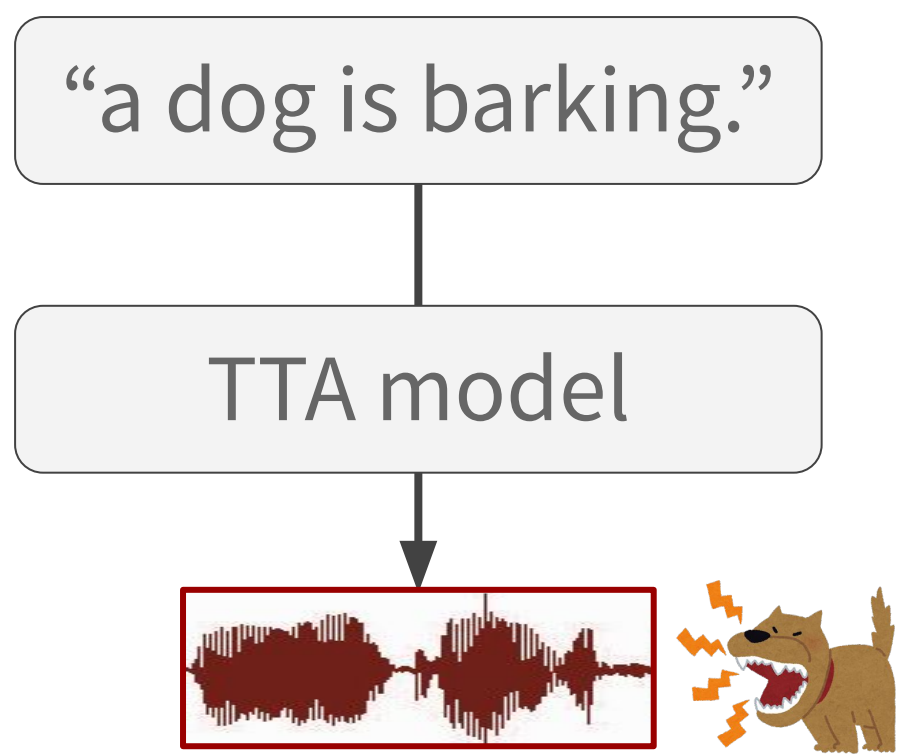


1. はじめに

1.1 研究背景

- 環境音合成（text-to-audio; TTA）は、アニメや映画に必要な音の作成に使われる[1]。
→環境音収録，検索コストを大幅に削減。
- TTAモデルの性能を上げるには何が必要？
→合成音の適切な評価指標。

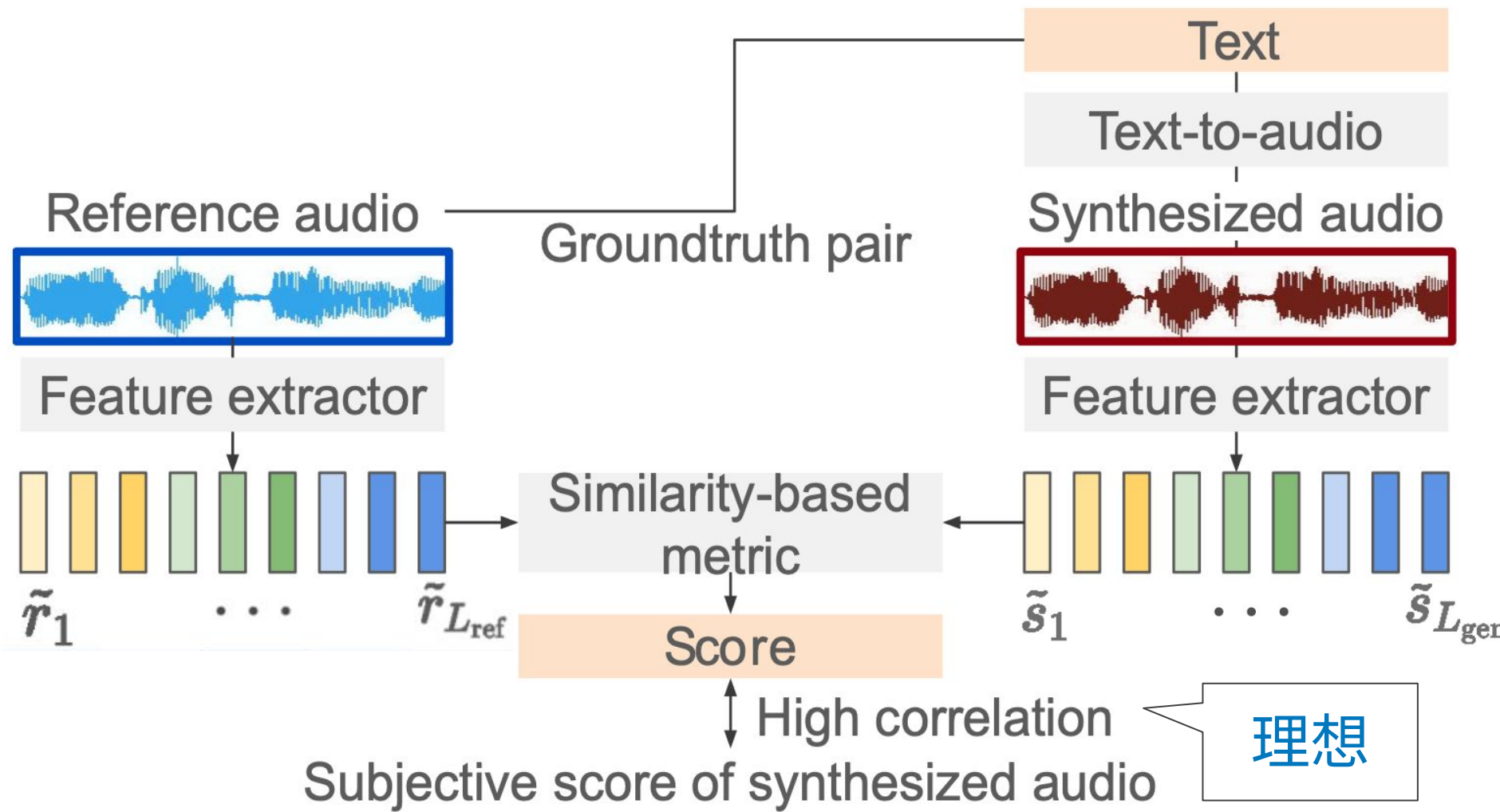


1.2 従来の評価指標

- 主観評価（MOS）[2]：人間が1~5の5段階で評価した平均。
 - OVL（overall quality）：合成音の品質・自然さ
 - REL（relevance）：入力テキストと合成音の関連度→課題：時間的・金銭的成本がかかる。
- 自動評価：合成音と{テキスト・正解音}を用いて合成音を自動で評価。
→課題：従来の自動評価（MCD[3]等）は主観評価値と相関が低い[4]。
→人間の評価に強く相関する自動評価指標を作りたい。

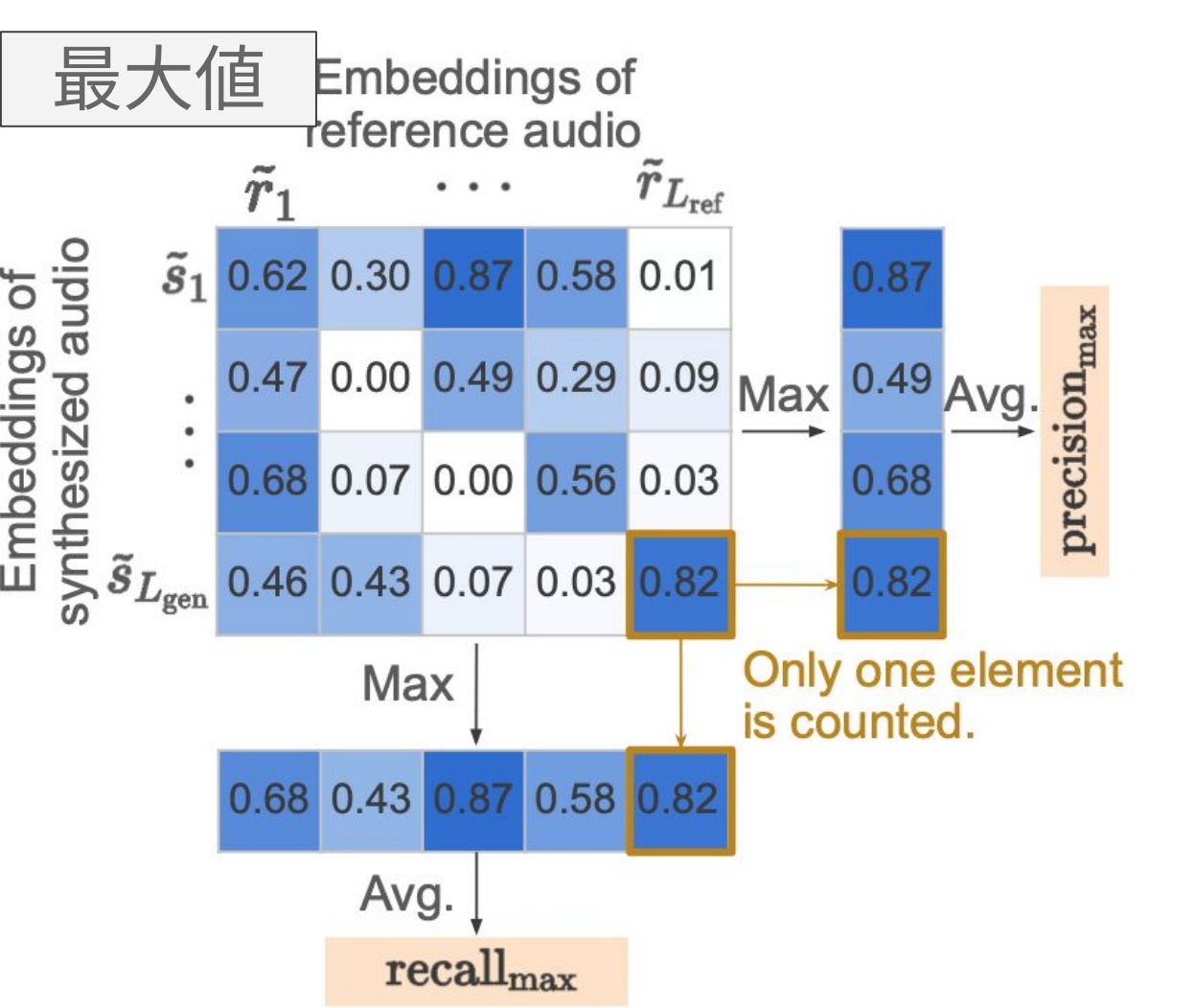
2. 提案法（AudioBERTScore）：合成音，正解音それぞれの埋め込みベクトルの類似度計算によるスコア

2.1 提案法概要

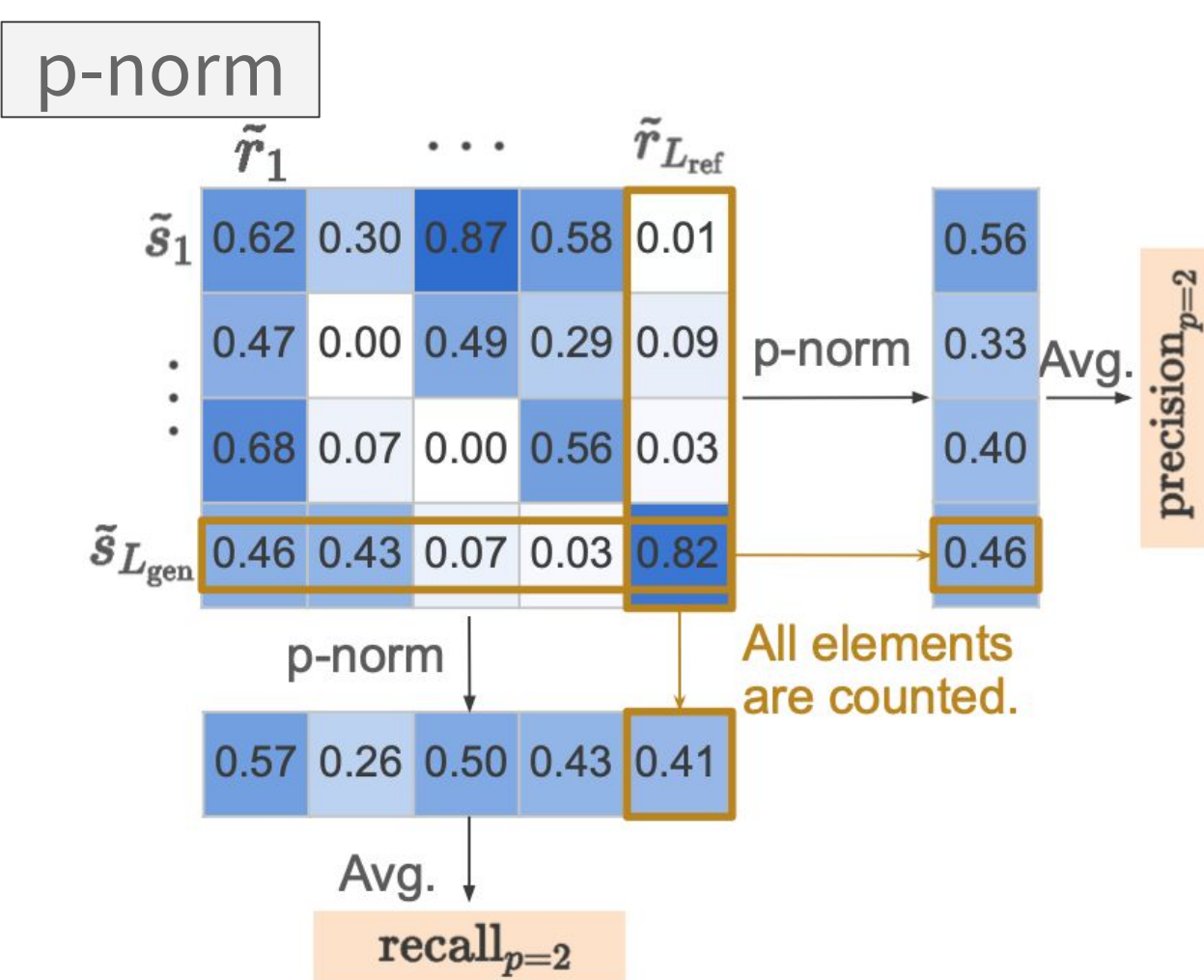


- 合成音と参照音（テキストと対の正解音）からそれぞれ埋め込みベクトル系列を獲得
→類似度計算によりスコアを算出。
- そのスコアと主観評価の相関が高くなることが望ましい。
- 主観評価値との相関を上げる際に重要な点
 - Feature extractorが適切に音をベクトル系列で表現できるか。
 - 合成音と正解音ベクトルの類似度計算をどのように行うか。

2.2 類似度計算の詳細



- 類似度行列の計算
 - $\tilde{\mathbf{r}}_*$ (行) は正解音ベクトル系列， $\tilde{\mathbf{s}}_*$ (列) は合成音ベクトル系列
 - 各要素は合成音と正解音ベクトルのコサイン類似度
- 2種類のノルム計算
 - 最大値：自然言語処理の既存手法 BERTScore [5] と同様。局所フレームの対応を仮定。
 - p-norm：時間遍在する環境音に対応するよう新たに定式化。p->∞で最大値ノルム。



- 3種類のスコア
 - precision：合成音の特徴を正解音がどのくらい含むか
$$\text{precision}_{\lambda,p} = \lambda \cdot \text{precision}_{\text{max}} + (1 - \lambda) \cdot \text{precision}_p$$
 - recall：正解音の特徴を合成音がどのくらい含むか
$$\text{recall}_{\lambda,p} = \lambda \cdot \text{recall}_{\text{max}} + (1 - \lambda) \cdot \text{recall}_p$$
 - F1：precisionとrecallの調和平均
$$\text{F1}_{\lambda,p} = 2 \times \frac{\text{precision}_{\lambda,p} \times \text{recall}_{\lambda,p}}{\text{precision}_{\lambda,p} + \text{recall}_{\lambda,p}}$$

3. 実験的評価：提案法は人間の評価値とどのくらい相関する？

3.1 実験条件

- 評価セット = PAM[5]テストセット332組
 - 参照音（正解音）
 - テキスト
 - 合成音
 - 主観評価値：OVL, REL の5段階MOS値
- 特徴抽出器
 - BYOL-A[7]，ATST-Frame[8]，AST[9]
- 提案手法スコアと人間評価値の相関の計算
 - ピアソンの線形相関係数（LCC）
 - スピアマンの順位相関係数（SRCC）

3.2 各特徴抽出器と各層の比較実験

- 主観評価値と相関が高くなる特徴抽出器の探索。
 - 特徴抽出器はそれぞれ層を持つ。各層を用いて計算したスコアと主観評価値との相関。
-
- TransformerベースであるAST-13層目が主観評価値と最も高い相関。
 - BYOL-A（CNNベース）→ local層は音響特徴，global層は文脈情報をそれぞれ保持。
 - AST，ATST-Frame（Transformerベース）→ 後半の層が文脈情報を含んでいる。

3.3 OVLとRELの相関値の相関

- 特徴抽出器各層の{OVL, REL}-LCCの相関。
-
- 全体的にOVL-LCC < REL-LCC。
→音質より関連度の定量化に優れる。
 - 相関値の相関 > 0.8
→片方の性能向上でもう一方も向上

3.4 p，λの値の比較実験

- 主観評価値との相関が高くなるp, λの探索。
 - p=106, λ=-3.5で最も高い相関。
-
-
-

3.5 他の自動評価指標との比較

- 参照音を用いる従来指標MCD・WARP-Q[10]よりも高い相関。
→環境音の自動評価に強く貢献。
- | | OVL | | REL | |
|--|--------------|--------------|--------------|--------------|
| | LCC | SRCC | LCC | SRCC |
| Compared metrics | | | | |
| MCD | 0.004 | 0.008 | 0.029 | 0.05 |
| WARP-Q | 0.241 | 0.228 | 0.202 | 0.213 |
| Proposed AudioBERTScore (AST, 13-th layer) | | | | |
| F1 _{max} | 0.368 | 0.397 | 0.490 | 0.515 |
| F1 _{λ=0, p=106} | 0.401 | 0.420 | 0.523 | 0.545 |
| F1 _{λ=-3.5, p=106} | 0.424 | 0.433 | 0.546 | 0.567 |
| Other metrics | | | | |
| PAM | 0.595 | 0.604 | 0.529 | 0.556 |
| CLAPScore | 0.337 | 0.323 | 0.487 | 0.475 |
- p-norm計算導入により相関向上。

Reference

[1] T. Marrinan, arXiv, 2024
[2] Y. Okamoto, " APSIPA ASC, 2022.
[3] R. Kubichek, PACRIM,1933
[4] 高野, 日本音響学会, 2025
[5] T. Zhang, ICLR, 2020
[6] T. Saeki, Interspeech, 2024
[7] D. Niizumi, TASLP, 2023
[8] Y. Gong, Interspeech, 2021
[9] X. Li, arXiv, 2023
[10] W. A. Jassim, ICASSP, 2021
[11] F. Xiao, DCASE Workshop, 2024