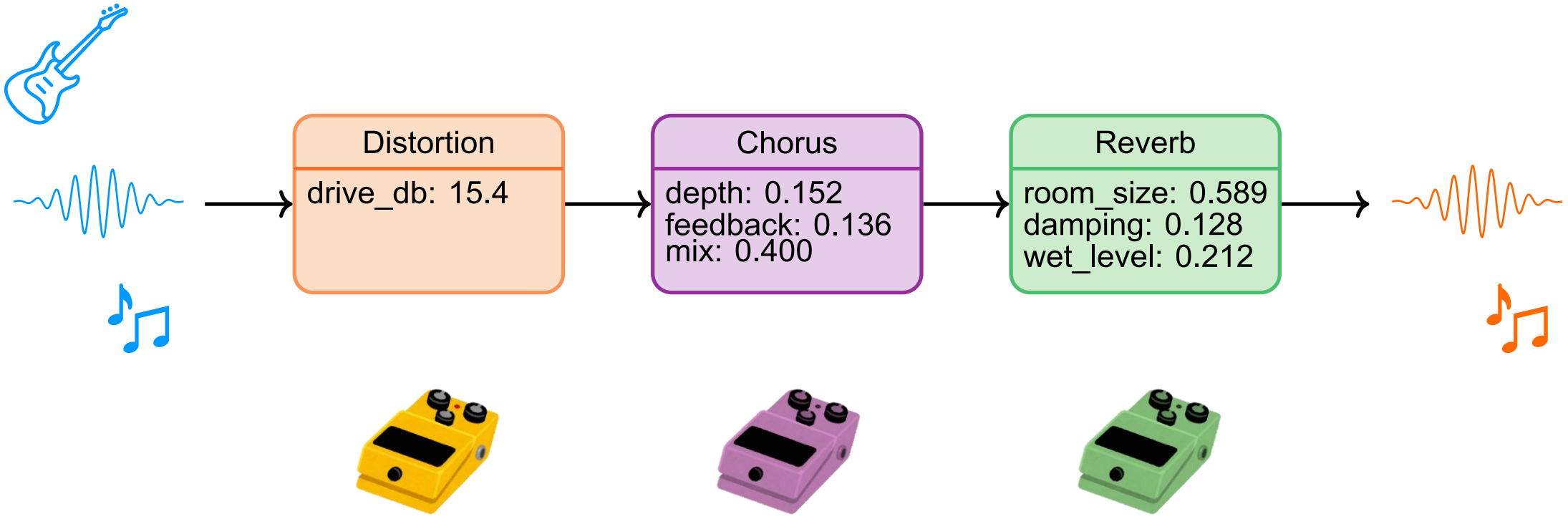
深層学習に基づく予測と探索アルゴリズムによるオーディオエフェクト推定

第145回音楽情報科学研究発表会

沖田 陽一　片寄 晴弘

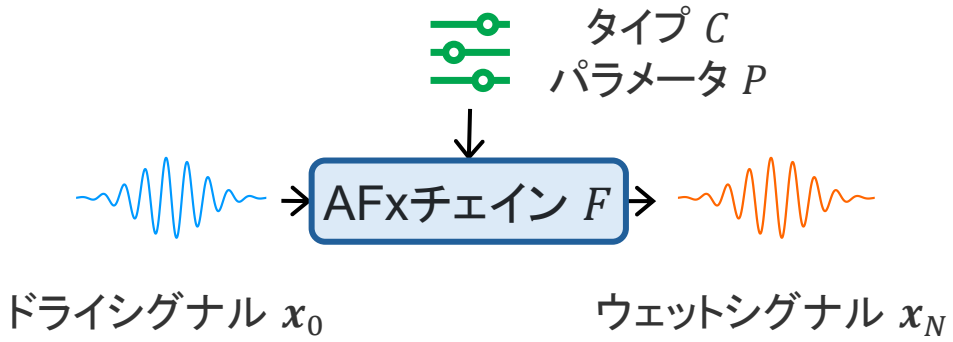
関西学院大学

オーディオエフェクト


Distortion
drive_db: 15.4
Chorus
depth: 0.152
feedback: 0.136
mix: 0.400
Reverb
room_size: 0.589
damping: 0.128
wet_level: 0.212

オーディオエフェクトの適用と推定

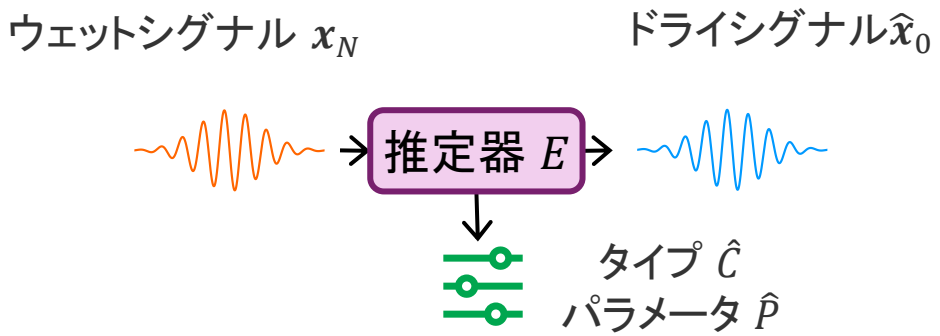
エフェクト適用



$$f_{c_N,\boldsymbol{p}_N} \circ \cdots \circ f_{c_1,\boldsymbol{p}_1}(\boldsymbol{x}_0) = F_{C,P}(\boldsymbol{x}_0) = \boldsymbol{x}_N$$

$$C = (c_1, \cdots, c_N)$$
$$P = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_N)$$

エフェクト推定



$$E(\boldsymbol{x}_N) = \left(\hat{C}, \hat{P}, \hat{\boldsymbol{x}}_0\right)$$

- ウェットシグナルからエフェクト構成（タイプ・パラメータ）の列とドライシグナルを推定

背景

- オーディオエフェクトは音楽や音声などのサウンドデザインにおいて広く用いられる
- しかし、複雑で多様なエフェクトを自在に駆使したサウンドデザインには技術面と芸術面の双方で高い専門性必要

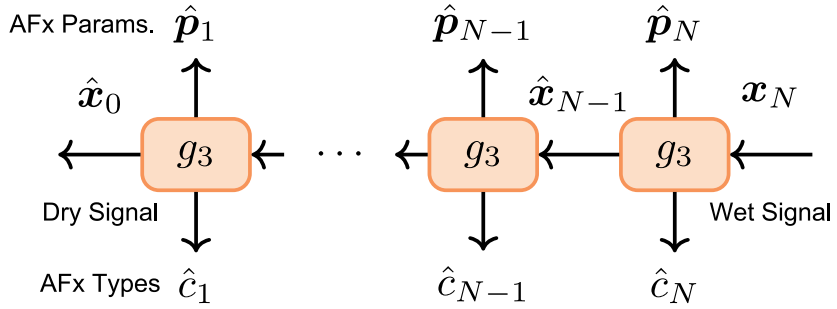
► 自動のオーディオエフェクト推定によって、既存の制作物からそのサウンドデザインの手法を効率的に学び、再利用することが可能に

関連研究

- **オーディオエフェクト推定**
  - 予測的アプローチ [Hinrichs+ 2022, Wada+ 2025] ウェット → エフェクト
    - エフェクト構成の教師データによって、DNNなどの**機械学習モデル**を学習
    - 学習済みモデルによって、未知のウェットシグナルに対して推論
    - ► 複数エフェクトについて、順序やパラメータを含めた完全な構成を推定可能な手法は少ない
  - 探索的アプローチ [有山+ 2020, Yu+ 2025] ウェット・ドライ → エフェクト
    - ドライシグナルに推定したエフェクトを適用して**ウェットシグナルを再構成**
    - それと目標のウェットシグナルとの類似度を最大化するよう、エフェクト構成を動的に探索
    - ► 推定時にドライシグナル必要

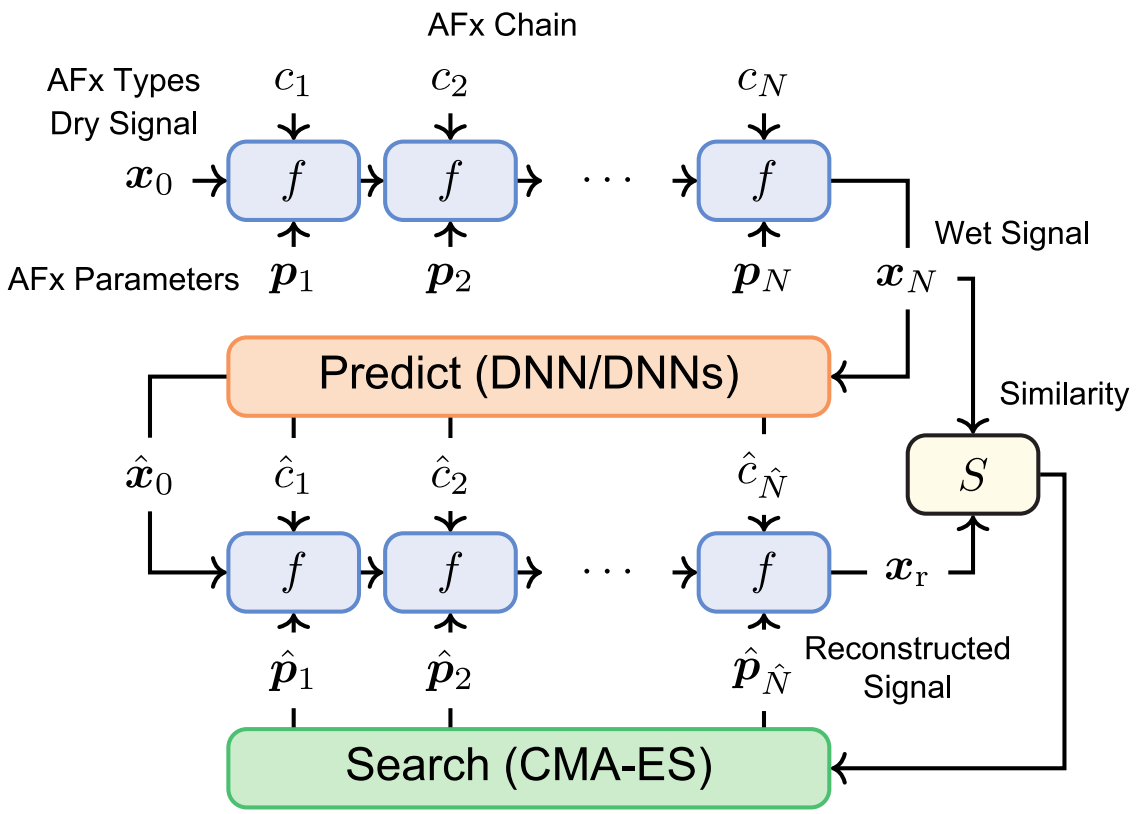
関連研究

- **オーディオエフェクト構成とドライシグナルの同時推定** [Take+ 2024] ウェット → エフェクト・ドライ
  - チェイン中の最後のエフェクトについて、その構成と入力信号（バイパスシグナル）を推定するDNNであるSunAFXiNetを構築（予測的アプローチ）
  - そのモデル自身にバイパスシグナルを再び入力する過程を繰り返すことで、チェイン全体の完全な構成とドライシグナルを推定
  - ▶ 反復推論における誤差累積は課題



- ▶ **提案アプローチ**：予測的アプローチに探索的アプローチを組み合わせ、性能向上を目指す

提案アプローチ



- オーディオエフェクト推定に対して、
  予測的アプローチと探索的アプローチを
  組み合わせた新たなアプローチを提案

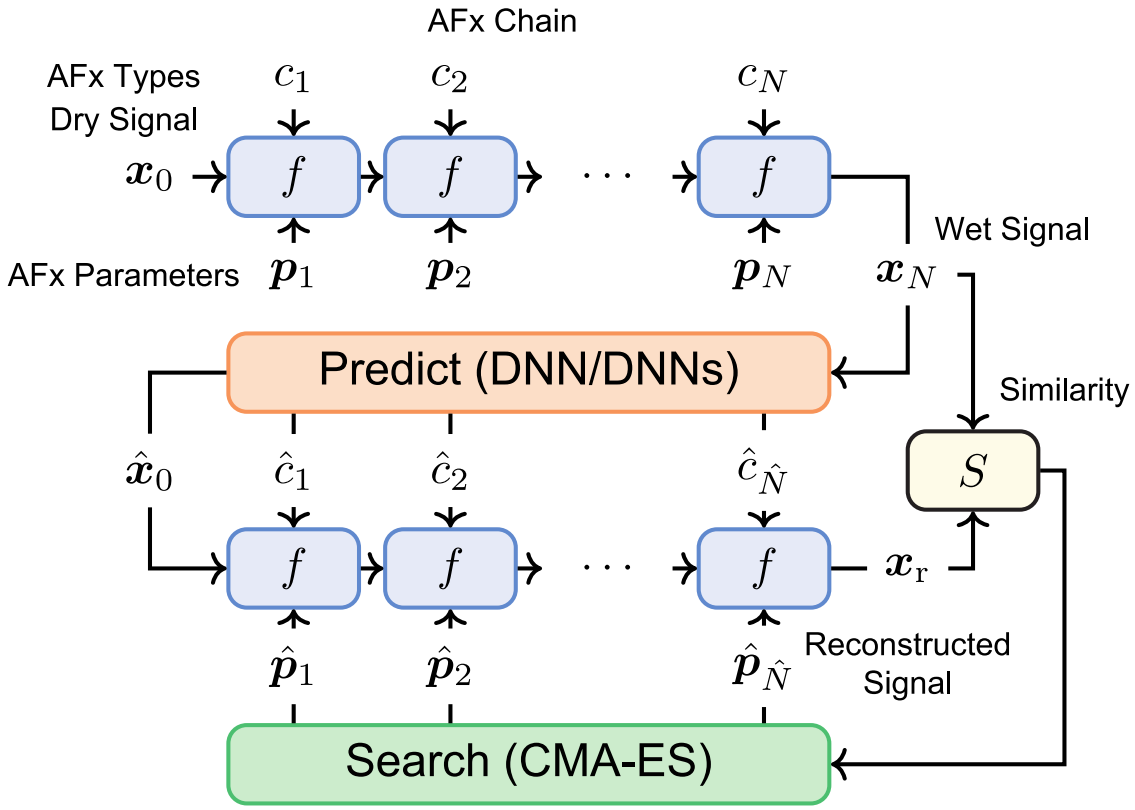
1. 予測段階：エフェクト構成・ドライシグナル
   を予測

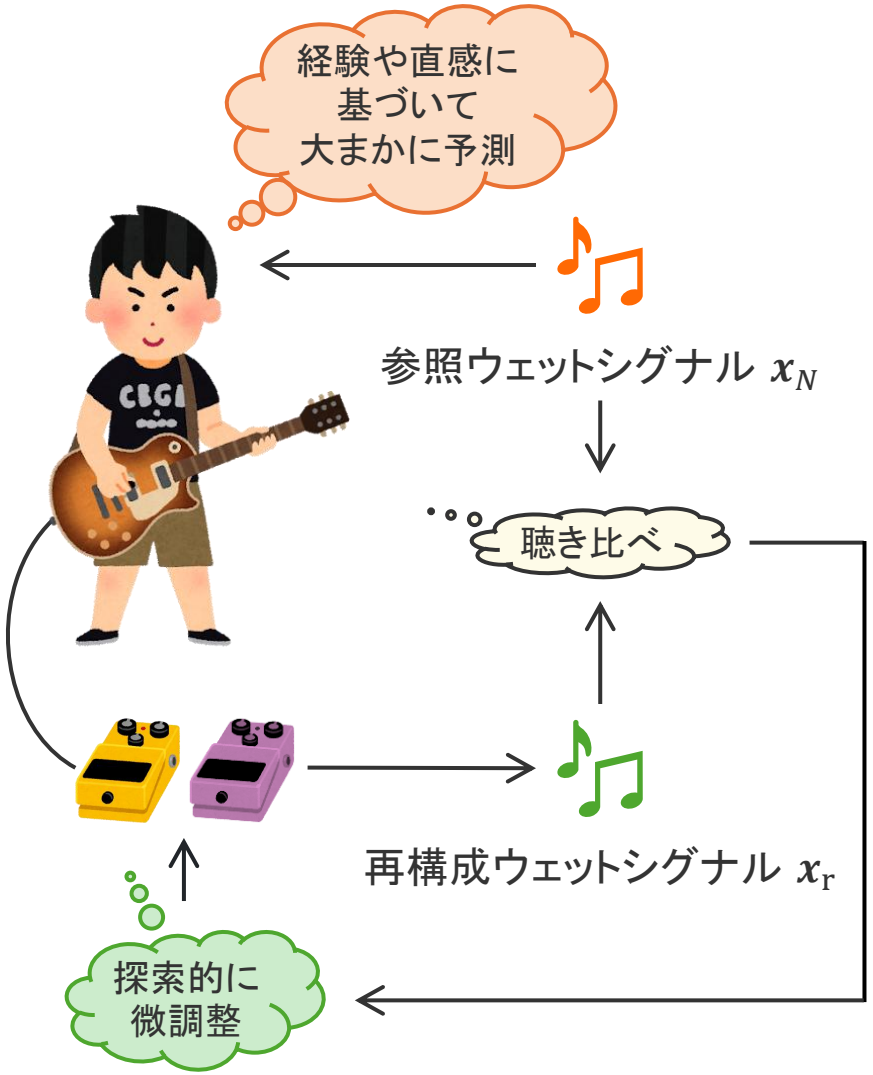
2. 探索段階：ウェットシグナル再構成類似度
   を最大化するようエフェクトパラメータを
   探索

▶ 予測段階でエフェクト除去を行うことで
再構成類似度の評価が可能に

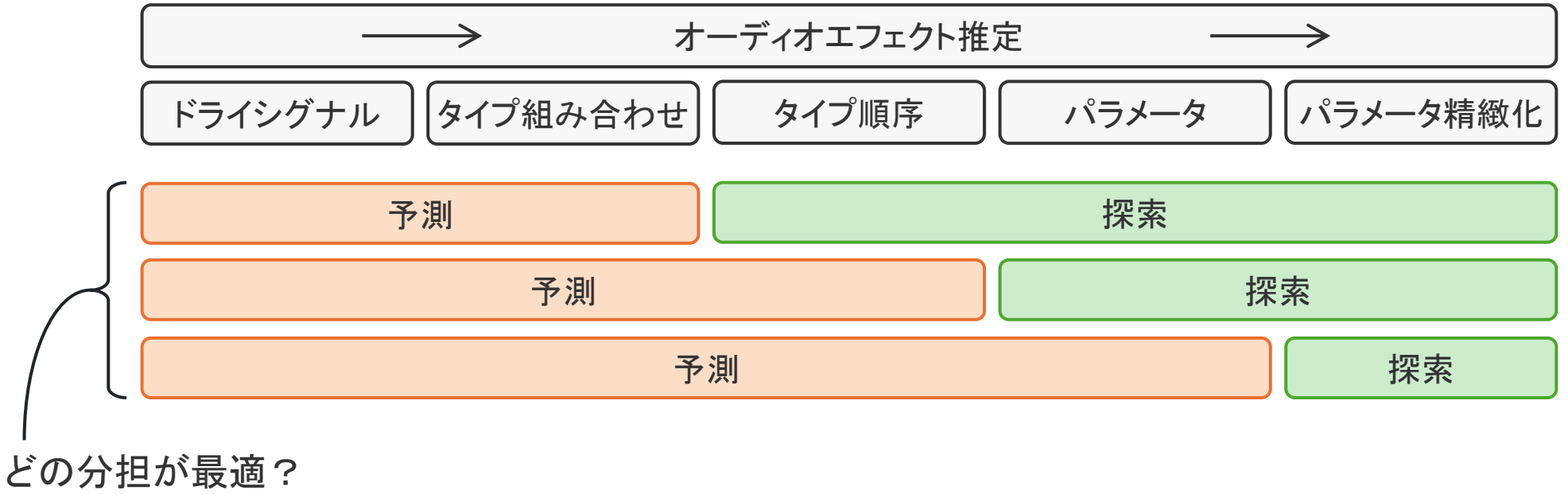
▶ 後段の探索で前段の予測を**補完・改善**

提案アプローチのアナロジー





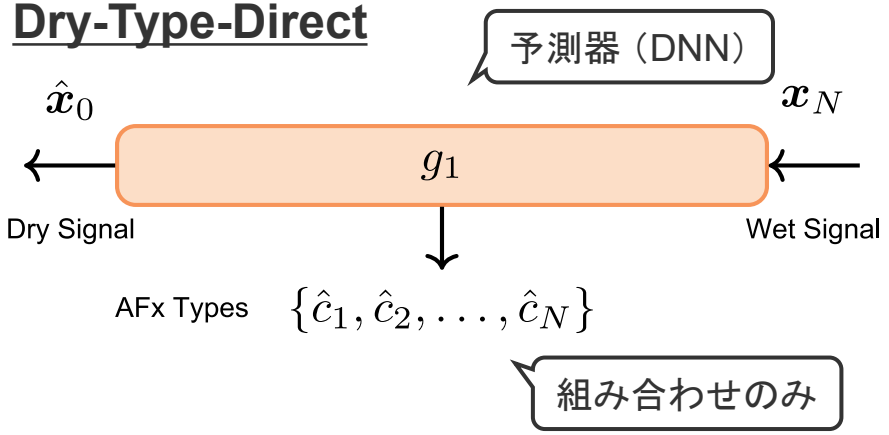
提案アプローチに基づく手法設計



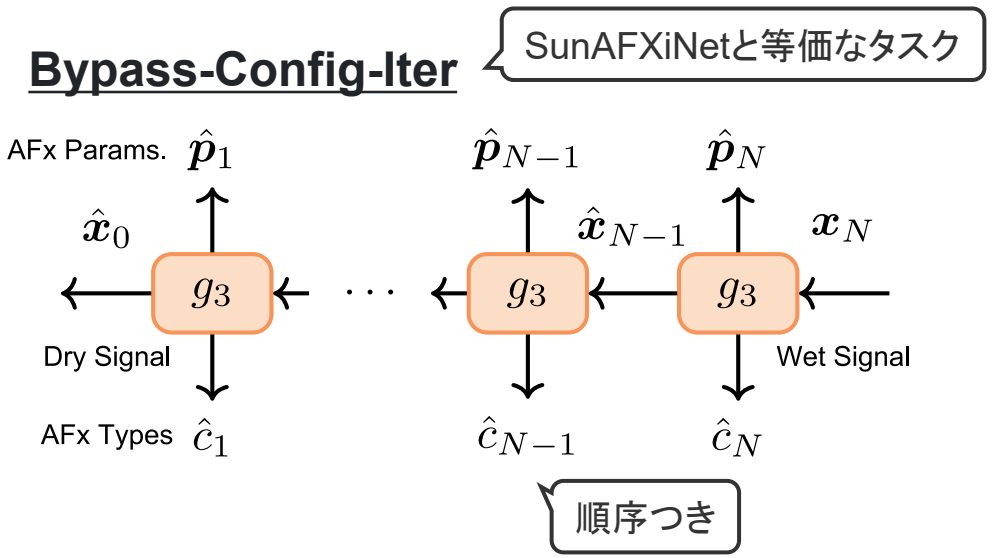
- 二段階のアプローチに基づいた手法設計では、各段階の間で**タスクの分担**は不可欠な選択
  - ► 全体として解くタスクは同一だがタスクの分担が異なる3手法を比較

深層学習に基づく予測

1. **Dry-Type-Direct**

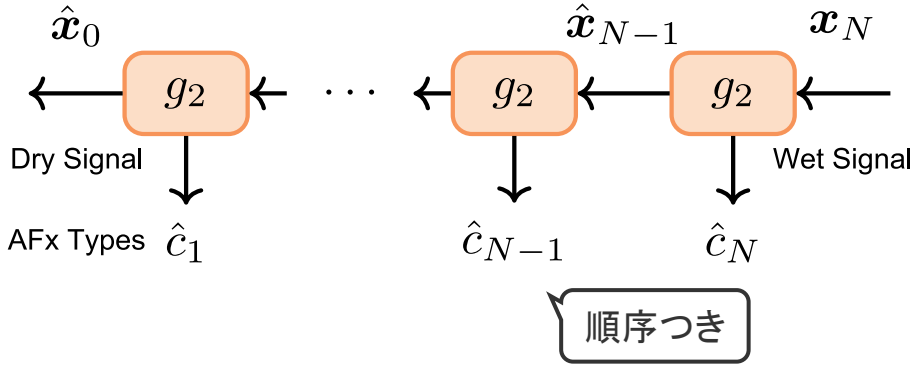


3. **Bypass-Config-Iter**



- 予測器はいずれもSunAFXiNet [Take+2024] ベースのアーキテクチャのDNN

2. **Bypass-Type-Iter**



- 残りのタスクは探索段階へ

ウェットシグナル再構成に基づく探索

- ウェットシグナル再構成類似度を最大化するようエフェクトパラメータを探索

予測段階で推定したドライシグナル

$$\hat{P} = \arg\max_{P} S\left(\underbrace{F_{\hat{C},P}(\hat{\boldsymbol{x}}_0)}_{\text{再構成ウェットシグナル}}, \boldsymbol{x}_N\right)$$

S: 音響信号間の類似度（今回はSI-SDRを使用）

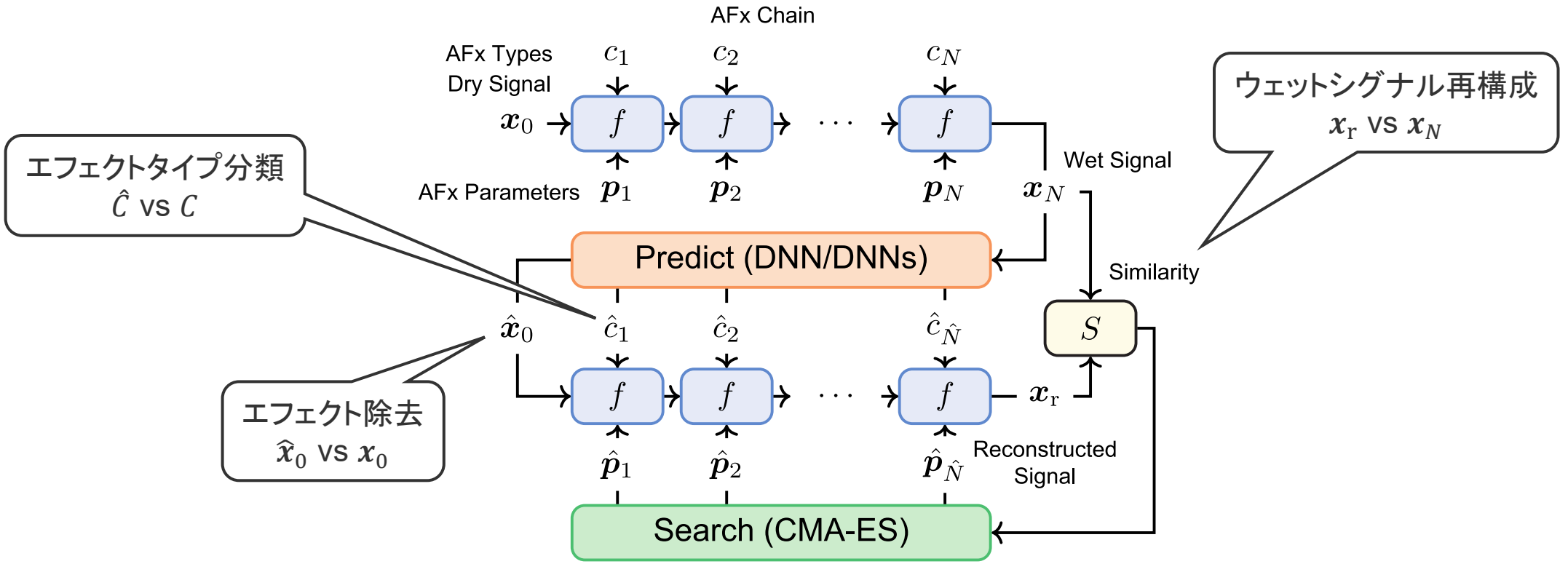
- 目的関数内のエフェクトFは、多くの場合、実装不明または微分不可能
  - ブラックボックス最適化（今回は進化的アルゴリズムCMA-ES [Hansen+ 1996] を使用）

- 予測段階でCの順序を推定しないDry-Type-Directでは二段階の探索
  1. 予測組み合わせからなる全順列について探索し**順序決定**
  2. 推定順序で**さらなる精緻化**のためP探索（前段の$\hat{P}$を**初期解**に）

- 予測段階で$\hat{P}$を推定するBypass-Config-Iterでもそれを探索の**初期解**に

評価実験

- 3つの側面から提案手法の性能を評価
- チェイン長Nごとの傾向も分析



評価実験に用いたデータセット

- ドライシグナル
  - 既存データセットから収集
  - ギターで演奏された音楽的抜粋
  - 10.0 sチャンク × 2231個

- ウェットシグナル
  - ドライシグナルにPedalboardライブラリのエフェクト適用
  - 各タイプが高々1回現れる全15通りのチェインを構成
  - 可変パラメータは表の範囲でランダムに設定
  - 中間の信号を含め合計205 h

▶ 予測モデルの学習・検証と手法の評価へ

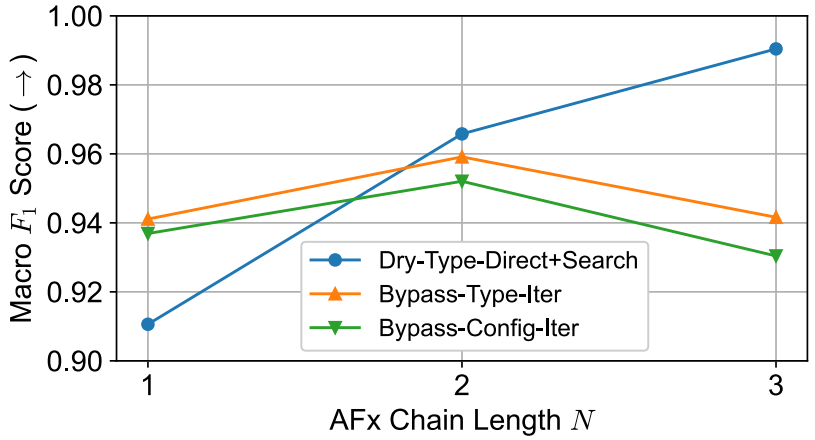
タイプ	可変パラメータ
Chorus	depth feedback mix
Distortion	drive_db
Reverb	room_size damping wet_level

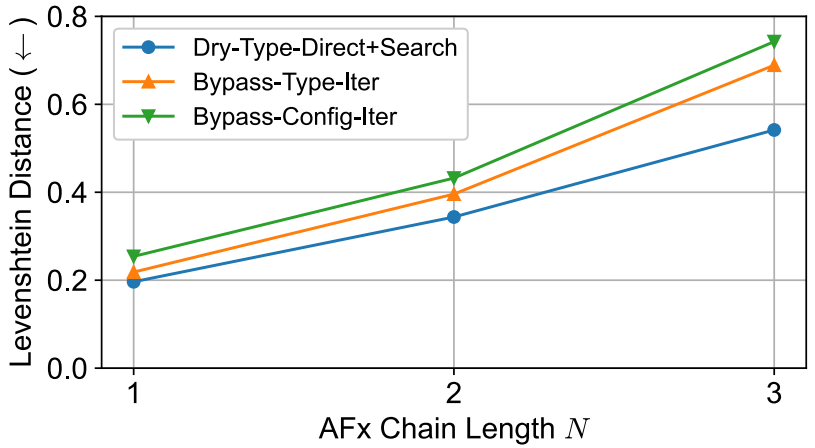
オーディオエフェクトタイプ分類の評価

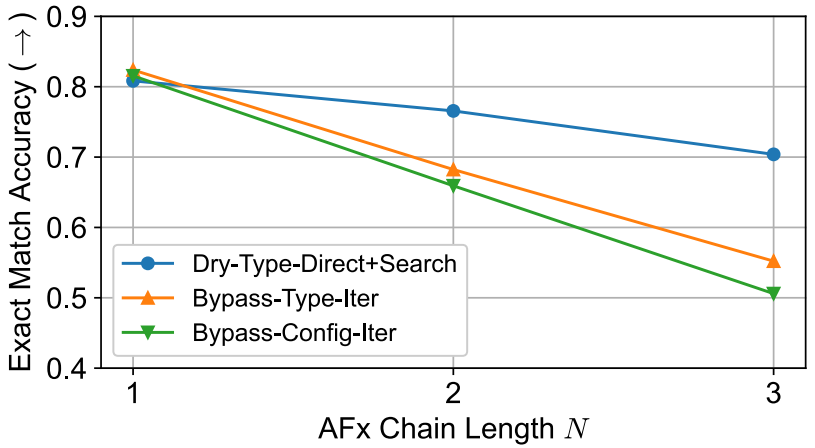
全指標で“Dry-Type-Direct + Search”最良

全体			
Method	Macro F_1 (↑)	LD (↓)	EMA (↑)
Dry-Type-Direct + Search	**0.958**	**0.313**	**0.774**
Bypass-Type-Iter	0.949	0.369	0.723
Bypass-Config-Iter	0.942	0.408	0.702

- Macro F_1 Score: 組み合わせのみ
- Levenshtein Distance: 順序考慮し、部分的正しさも評価
- Exact Match Accuracy: 完全一致のみ評価







F_1のみN増加に伴って性能良化する場合あり

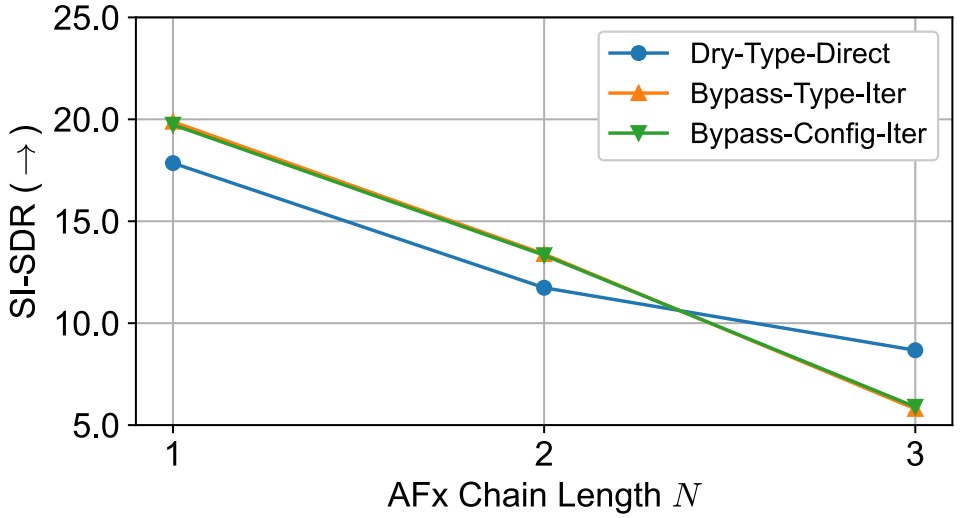
長いチェインほど“Dry-Type-Direct + Search”有利

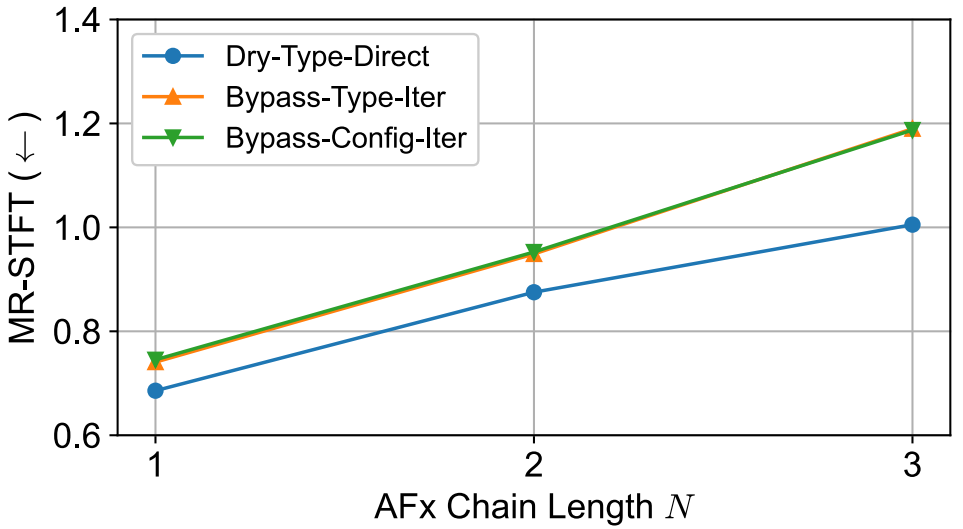
オーディオエフェクト除去の評価

時間領域と時間周波数領域の指標で評価分かれる

全体

Method	SI-SDR (↑)	MR-STFT (↓)
Dry-Type-Direct	13.96	**0.813**
Bypass-Type-Iter	**14.95**	0.898
Bypass-Config-Iter	14.88	0.902





$N = 3$では“Dry-Type-Direct”が逆転

ウェットシグナル再構成の評価

- ウェットシグナル再構成を通してエフェクト構成推定の総合的な性能を評価
- エフェクト除去性能とは独立に、主要な目的であったエフェクト構成推定性能を評価するため、評価時の再構成には真のドライシグナルを利用

$$\boldsymbol{x}_{\mathrm{r}} = F_{\hat{C},\hat{P}}(\boldsymbol{x}_0) \quad \text{vs} \quad \boldsymbol{x}_N$$

- ベースライン：予測段階でエフェクト構成全体を予測するBypass-Config-Iterで、精緻化のためであった探索を行わない手法
  - ► **探索を加える提案アプローチの有効性を検証**

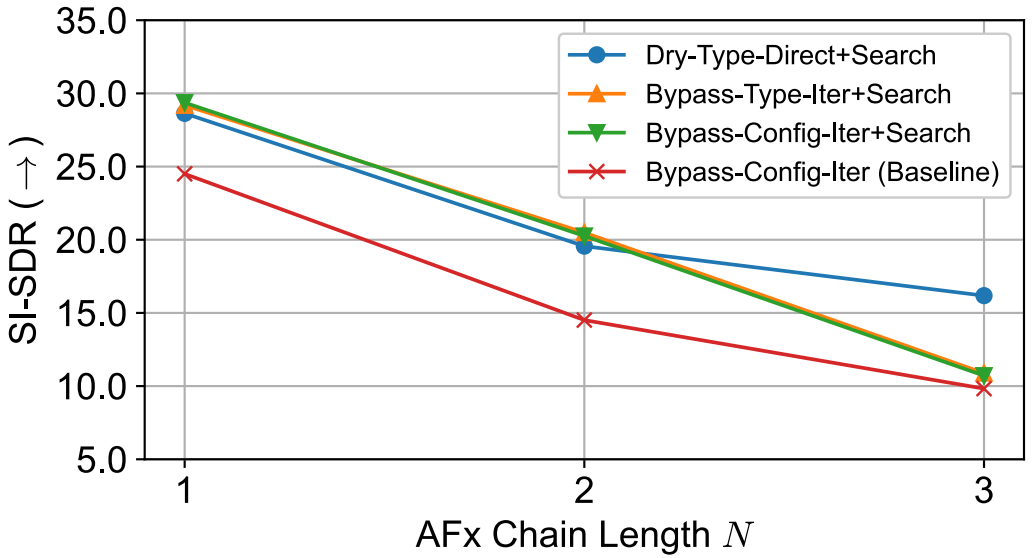
ウェットシグナル再構成の評価

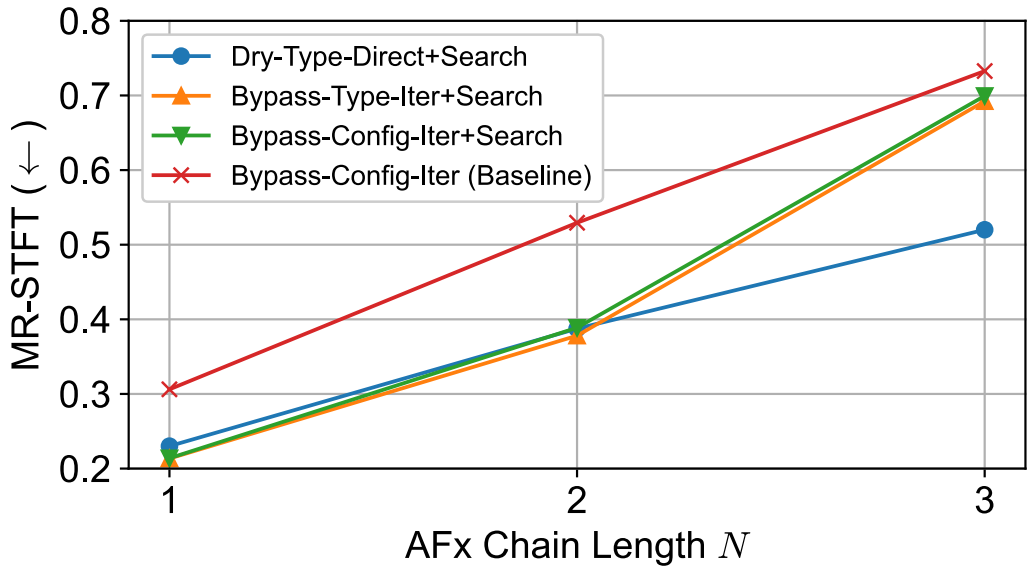
全提案手法がベースライン上回る

全体

Method	SI-SDR (↑)	MR-STFT (↓)
Bypass-Config-Iter (Baseline)	18.18	0.465
Dry-Type-Direct + Search	**23.07**	**0.340**
Bypass-Type-Iter + Search	22.68	0.361
Bypass-Config-Iter + Search	22.64	0.366

“Dry-Type-Direct + Search”が最良

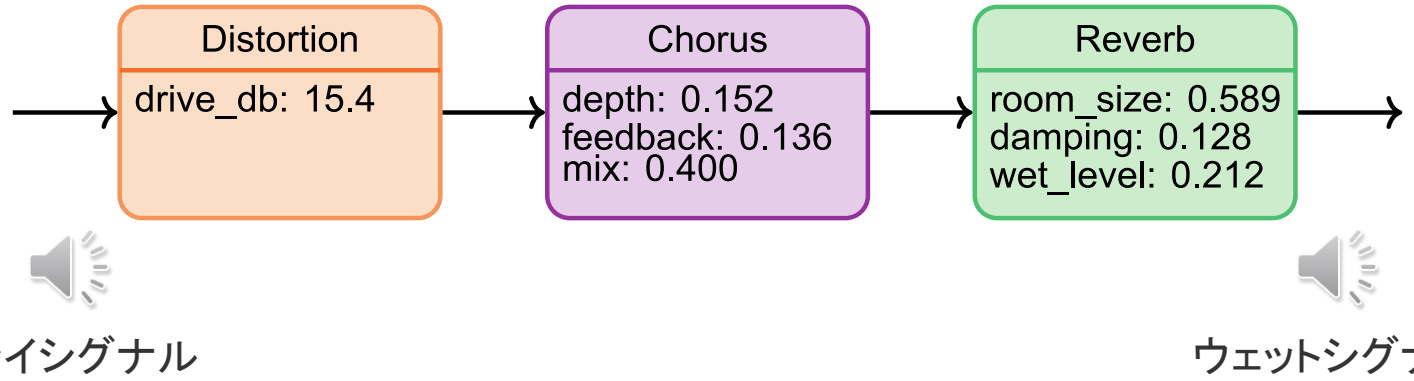




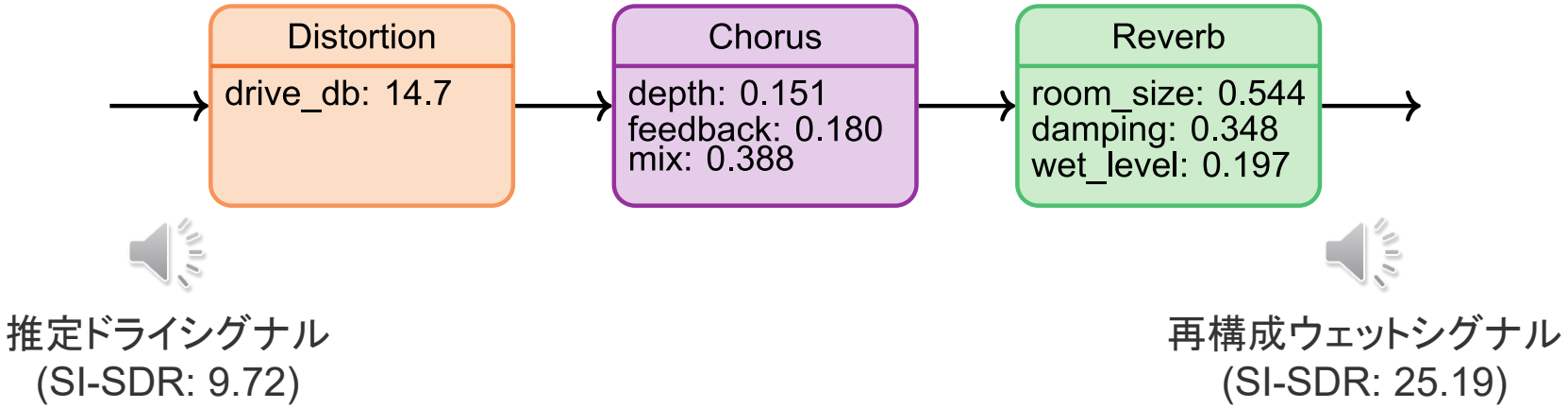
長いチェインほど“Dry-Type-Direct + Search”有利

推定結果の例

- Ground-Truth



- Dry-Type-Direct + Search



まとめ

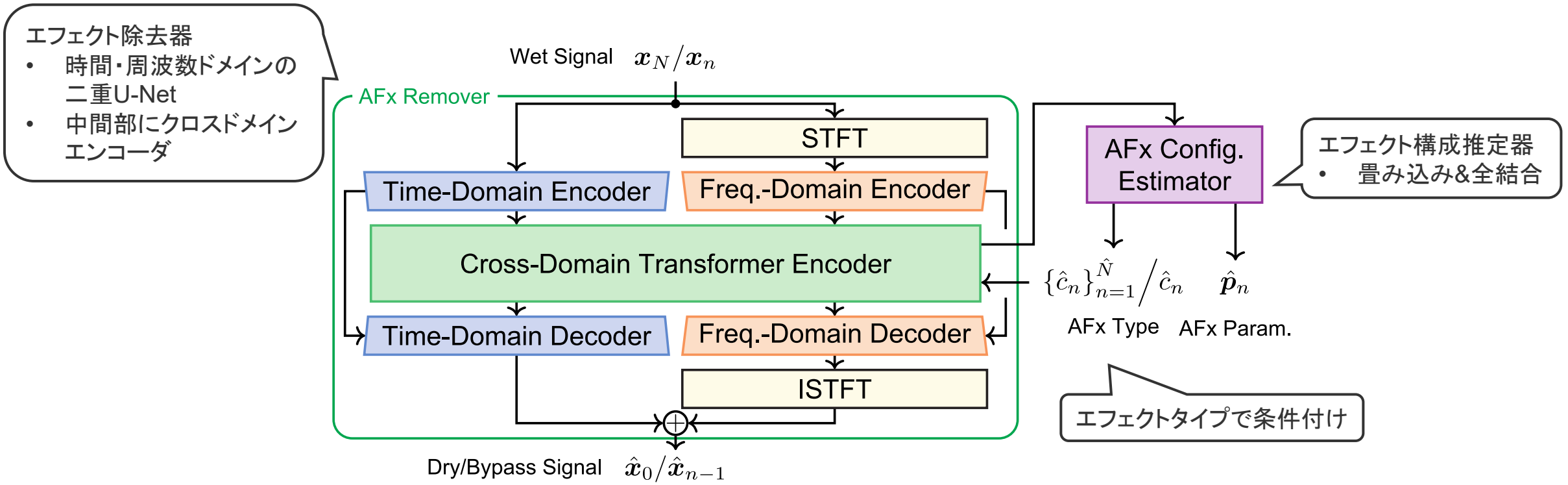
- 取り組んだこと
  - オーディオエフェクト推定に対して、深層学習に基づく予測と探索アルゴリズムを組み合わせた新たなアプローチを提案
  - 予測段階でドライシグナルを推定することで、ウェットシグナル再構成類似度を目的関数として、予測の補完や改善が可能に

- 評価実験
  - 提案手法が予測段階のみによる手法を上回る性能を示す
  - 予測段階でエフェクトタイプ組み合わせのみを予測し、探索段階でその順序を推定するタスク分担が最適であることが示唆される

デモページ

- 課題と展望
  - タイプごとの傾向や各指標のばらつきなど、さらなる性能分析の余地あり
  - 推定ドライシグナルに過剰適合しないよう探索規模をチューニング
  - 扱ったエフェクトの多様性（タイプやパラメータ）が限定的、これを拡張し汎用性を向上

予測モデルのアーキテクチャ



- SunAFXiNet [Take+ 2024] ベースのアーキテクチャ
- エフェクト除去器はHTDemucs [Rouard+ 2023] ベースのアーキテクチャ
- 予測タスクによって出力部などを変更

課題と展望：さらなる性能分析

- 修士論文で報告した性能分析
  - 評価用データセット全体にわたる平均
  - チェイン長ごとの傾向

↓

- さらなる性能分析
  - 性能のばらつきや信頼区間
  - エフェクトタイプごとの傾向

▶ 手法のより正確な評価

▶ エフェクト推定の性質に関する知見を蓄積

課題と展望：探索規模のチューニング

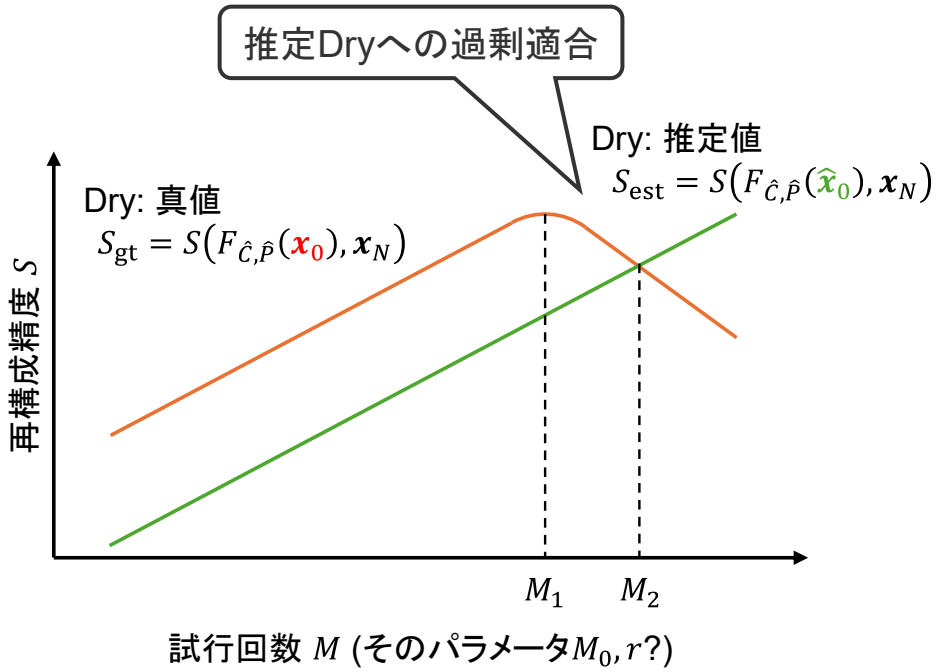
- 現行手法における探索段階の総試行回数の設定：

$$M = \lfloor M_0 d^r \rfloor$$

  - M_0: 定数パラメータ
  - d: 探索次元（エフェクトパラメータ数）
  - r: 定数パラメータ

- 誤差を含む推定ドライシグナル$\hat{\boldsymbol{x}}_0$を用いた最適化では$\hat{\boldsymbol{x}}_0$への過剰適合が起こりうる

$$\hat{P} = \arg\max_P S\left(F_{\hat{C},P}(\hat{\boldsymbol{x}}_0), \boldsymbol{x}_N\right)$$



- ▶ 過剰適合の挙動を確認し知見を蓄積
- ▶ 過剰適合を起こさないよう探索規模（パラメータM_0, r）をチューニングし精度向上を図る

課題と展望：エフェクトの多様性

- 現行手法で対象としたエフェクトの多様性（タイプ、パラメータ、個数）は、実用的なサウンドデザインを考えると限定的
  - ► 対象を拡大し汎用性を向上

- ただし、現行手法では探索が困難なパラメータが存在
  - パラメータの小さい変化によって波形が大きく変化
  - 再構成類似度（SI-SDR）のパラメータ空間における地形が多峰的、急峻
    - ► パラメータと相関のある特徴量や類似度の構築

現行手法で対象としたエフェクト

タイプ	可変パラメータ
Chorus	depth feedback mix
Distortion	drive_db
Reverb	room_size damping wet_level

現行手法で探索困難なパラメータ

タイプ	パラメータ
Chorus	rate_hz
Phaser	rate_hz
Delay	delay_seconds