2022年9月 第40回日本ロボット学会学術講演会

OS3「ロボット聴覚およびその展開」

そのエージェントの声、合っていますか？

- 声質変換技術と印象適合・人工感制御 -

齋藤 大輔（東京大学）

身近になった対話エージェント

- **音声で人とやりとりするコンピュータ：実体を伴わない**
  - スマートフォンに搭載のエージェント（e.g. Siri）
  - スマートスピーカ
- **対話ロボット：実体を伴う**
  - 物理的な実体を伴うエージェント
  - 人との多様なインタラクション
    - 身振り手振り
    - 音声会話

ロボホン

https://robohon.com

Romi

https://romi.ai

エージェントの声をどのように決めるか

- **聞き取りやすい声**
  - 言語的な了解性に着目
  - 伝えたい言語内容が伝わるかどうか
- **自然な声**
  - 音声の自然性に着目
  - 人間が話していると感じるかどうか
- **エージェントの身体性と調和する声**
  - エージェントの見た目との調和性に着目
  - そのエージェントが話していると感じるかどうか

オーディション、する？

根拠のある音声デザインのために

- **エージェントに調和した音声にむけて**
  - 吹き替える話者を選ぶのではなく調整して合わせる
- **声質変換技術による話者性の操作**
  - 声質変換によって話者情報を操作する：存在する話者へ
- **人工感制御技術による人工感の操作**
  - エージェントが出しているような声に加工する：存在しない声へ

本発表の流れ

- **声質変換技術概観**
  - パラレルからノンパラレルへ
  - 大規模データ、深層学習による高品質化
- **自然音声の人工感を連続的に制御する技術とその評価**
  - 自然音声をエージェントに合うように加工する技術
- **まとめ**

疑問がある場合は

- **Slido を用意しました**
  - 匿名で書き込めます
  - 質問を書き込んでもらえれば適宜対応します
  - 疑問点があれば気軽に質問してください
  - slido.com にアクセスし、#7338480 を入力

声質変換概観

声質変換とは

- Aさんの音声を入力として，Bさんの音声を出力する

声質変換

Aさん

入力 (Source)

Bさん

出力 (Target)

学習データにみる声質変換の分類

- **パラレル学習**
  - 入出力話者が同じ文章を読んだコーパスを用いて学習
- **ノンパラレル学習**
  - 入力話者、出力話者の学習データに制約がない
  - 単に入力1名、出力1名ではなく多くの人数の話者データを用いた多人数話者でのノンパラレル学習が近年の主流

近年の高品質化

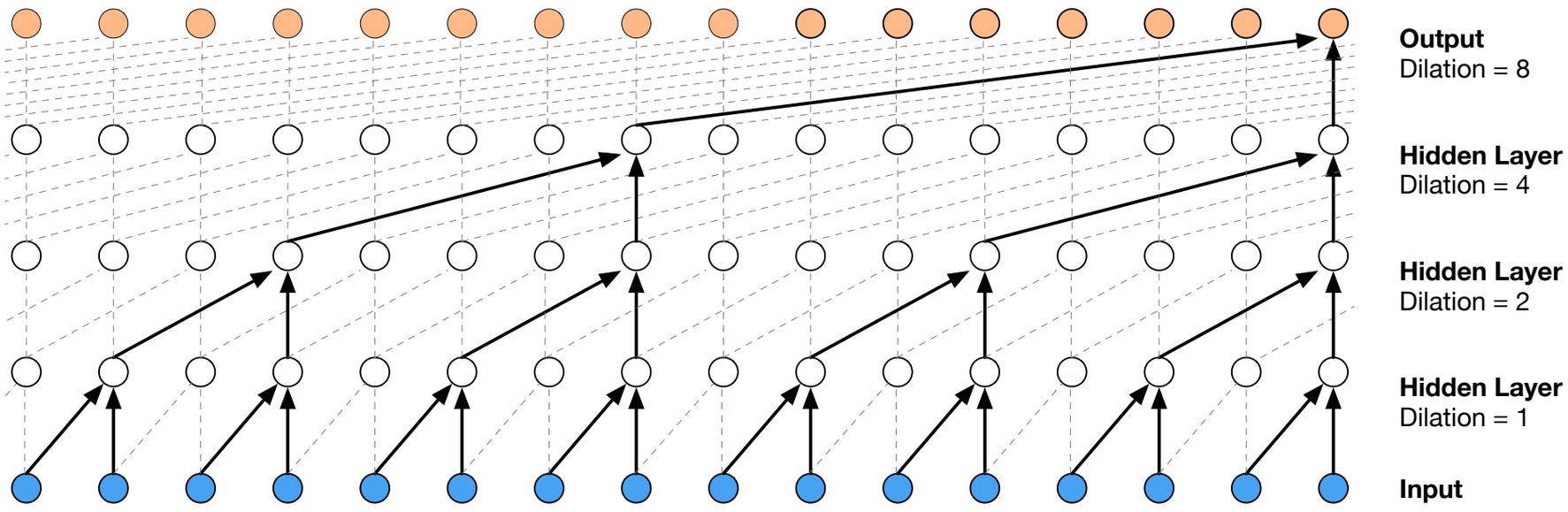
- **ASR-TTS方式による大規模データの活用**
  - 基本的なコンセプトは入力音声から言語特徴量を取り出す
    （音素事後確率：Phonetic PostrioGram / PPG）
  - PPGから出力話者の音声を個別に対応づけ、または
    話者埋め込みによる制御　　**独立した大規模データで音声認識部と音声生成部を学習**
- **高品質な音声波形生成**
  - ニューラルネットワークによる音声波形生成

ニューラル波形生成

- **音声波形サンプル点をニューラルネットワークで直接予測**
- **自己回帰構造を持つモデル**
  - 音声波形の過去のサンプルから次のサンプルをある程度説明可能（予測符号の考え方）
  - 条件付特徴量（メルスペクトログラムなど）と過去のサンプルのフィードバックから次のサンプルの生成を行う
  - 代表的な手法: WaveNet, SampleRNN, WaveRNN など

WaveNet

- **2016年登場したニューラル波形生成モデル**
  - Dilated causal convolution: 長期の過去のサンプルを考慮
  - Residual & Skip connection: 多層化
  - μ-lawで量子化された256クラスの系列識別問題として捉える



https://arxiv.org/pdf/1609.03499.pdf

ニューラル波形生成のカテゴリ

- **自己回帰モデル**
  - 自身の生成したサンプルを使い逐次的に生成する
  - WaveNet, SampleRNN, WaveRNN …
- **変数変換型**
  - 可逆な変数変換を用いてガウス分布から複雑なサンプルを生成
  - 複数サンプルを同時に生成することができる
  - Parallel WaveNet, ClariNet, WaveGlow …
- **信号処理との組み合わせ**
  - 信号処理と組み合わせて学習しやすい波形を学習する
  - LPCNet, GlotNet, Subband WaveNet …
- **学習基準の工夫**
  - GAN, 周波数領域の特徴量での誤差基準など…

声質変換とロボットの声

- **声質変換は基本的に存在する話者への変換**
  - モーフィング等で混ぜることは可能だがあくまで人間間の内挿
- **ロボットの見た目が“ロボット”である場合**
  - その声も“ロボット”である必要がある

ロボットに合う声としての指標が必要

人工感制御

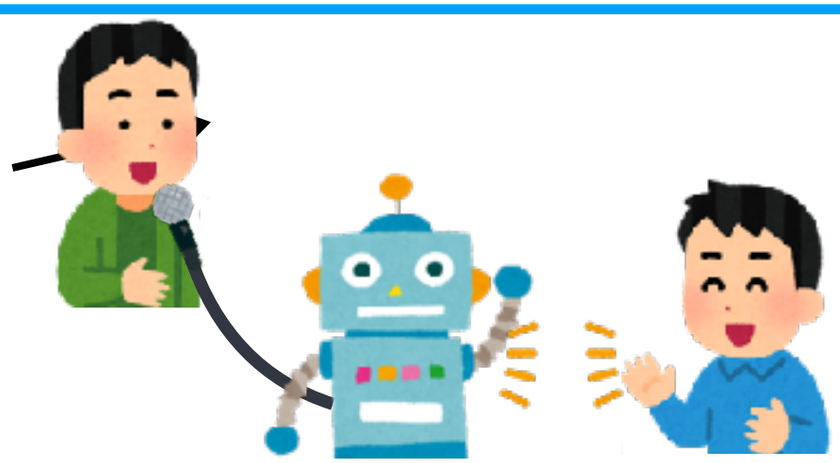
対話ロボットの印象と音声の印象

- **対話ロボットのような音声エージェント**
  - 音声の自然性 と同様に **音声の印象との合致が重要**
  - しかし、実際のロボットでは合致せず [McGinn+, 2019]
- **“合成された”音声という印象＝人工感**
  - 音楽や映像のロボットキャラクターの音声は、エフェクタが使用され人工感が付与

↓

適切な音声デザインのための人工感の制御

エフェクタ
コムフィルタ
ピッチ補正



本研究の目的

- **各手法の与える音声の人工感の分析**
  - 人工感を与えた音声とロボットの印象の主観的合致度を評価

    →**音声に人工感に与える手法を提案**
    - 基本周波数を離散化する手法
    - 音声の金属感を上昇させる手法
- **ロボットの音声デザインを支援**
  - 簡易的なアプリケーションの作成・評価

目次

- **従来手法**
  - コムフィルタ：反響しているような音声に変換する手法
  - Robotization Effect：基本周波数を一定に固定する手法
- **提案手法**
  - Robopitch：基本周波数を離散化する手法
  - Inharmonic Warping：音声の金属感を上昇させる手法
- **実験**
  - 音声とロボット画像の主観的合致度実験
  - ロボットの音声デザインを支援するアプリケーションの評価実験（追加実験）
- **まとめ**

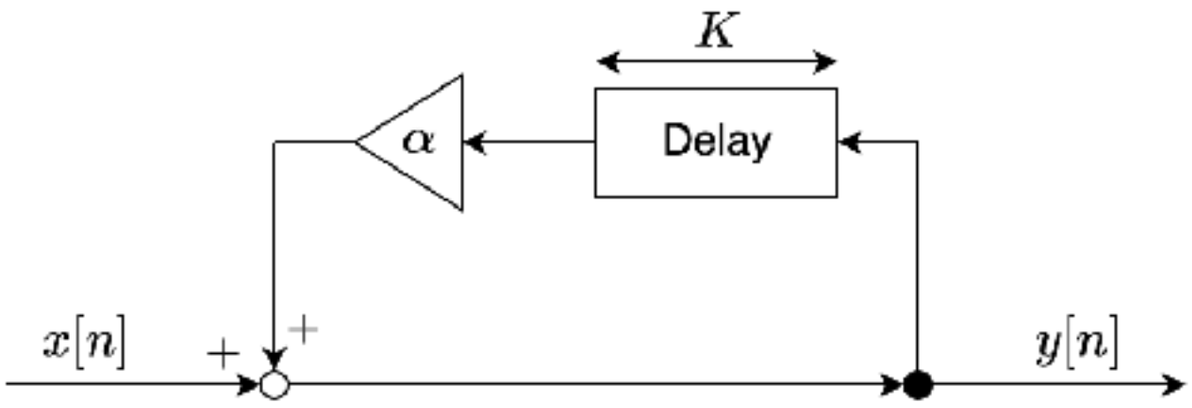
目次

- **従来手法**
  - コムフィルタ：反響しているような音声に変換する手法
  - Robotization Effect：基本周波数を一定に固定する手法
- **提案手法**
  - Robopitch：基本周波数を離散化する手法
  - Inharmonic Warping：音声の金属感を上昇させる手法
- **実験**
  - 音声とロボット画像の主観的合致度実験
  - ロボットの音声デザインを支援するアプリケーションの評価実験（追加実験）
- **まとめ**

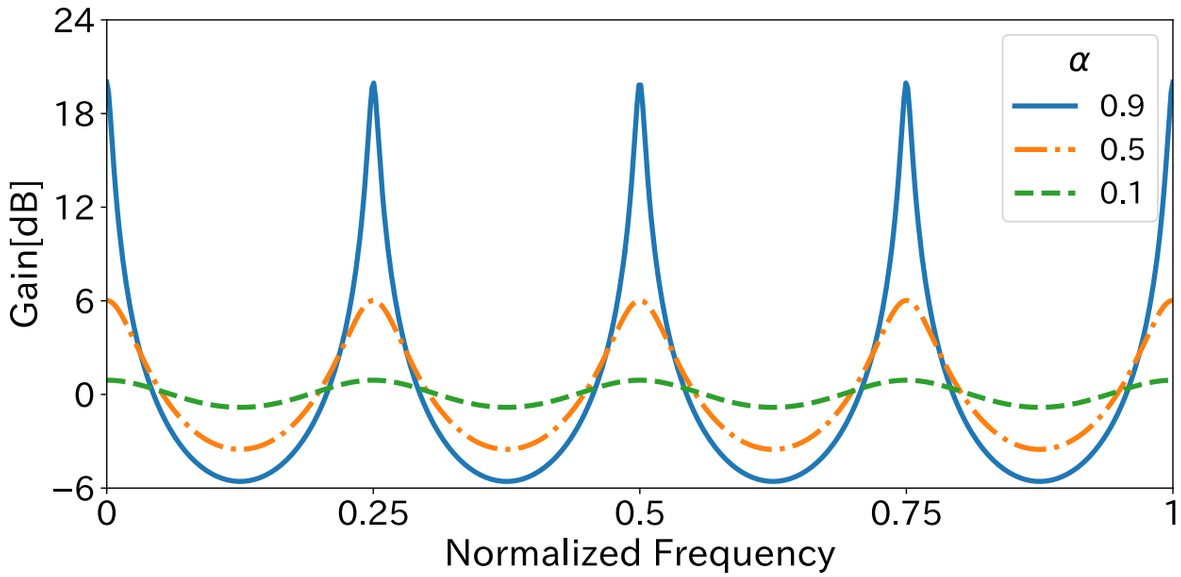
従来手法1｜コムフィルタ

コムフィルタ

- 自身の信号に遅延させた信号を加算することで干渉させるフィルタ
- 反響音が加えられたような音声に変換
- 機械的な音声への変換に利用 [McGinn+, 2019]



コムフィルタの構造



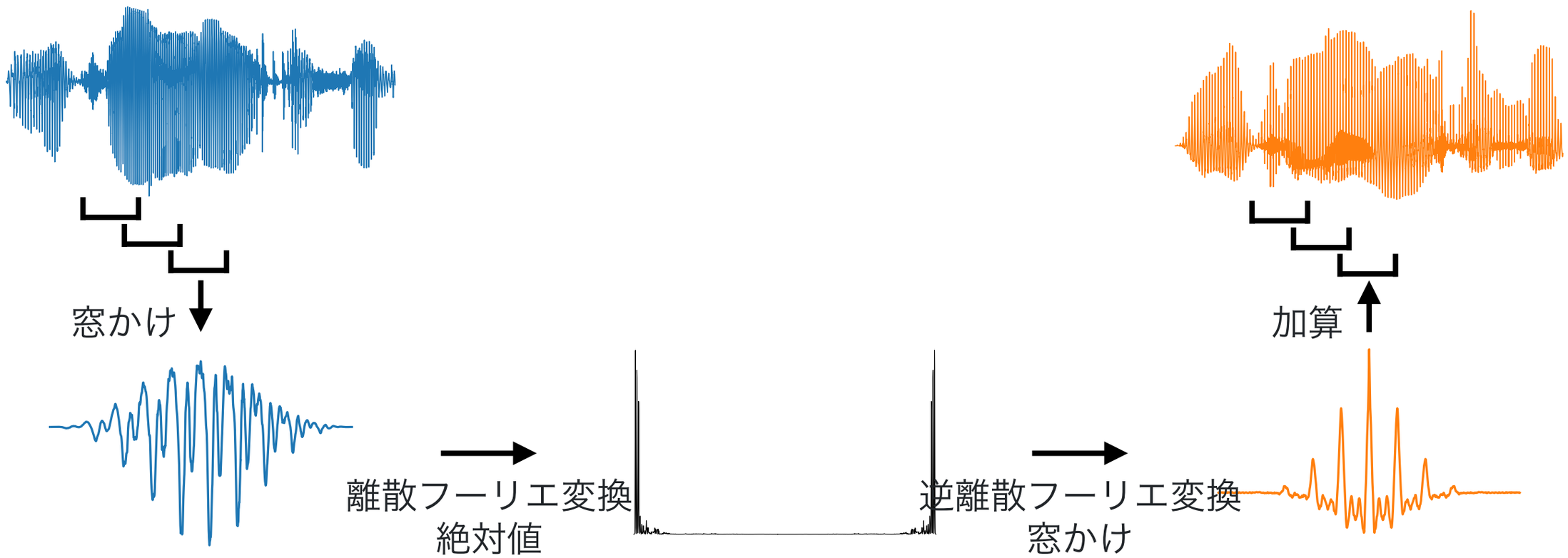
周波数応答

従来手法2｜Robotization Effect

Robotization Effect [Zölzer, 2002]

- フェーズボコーダの一種
- 短時間フーリエ変換（STFT）を行い、位相を0にする
- 基本周波数がF_s/Sに固定

（サンプリング周波数：F_s フレームシフト：S）



目次

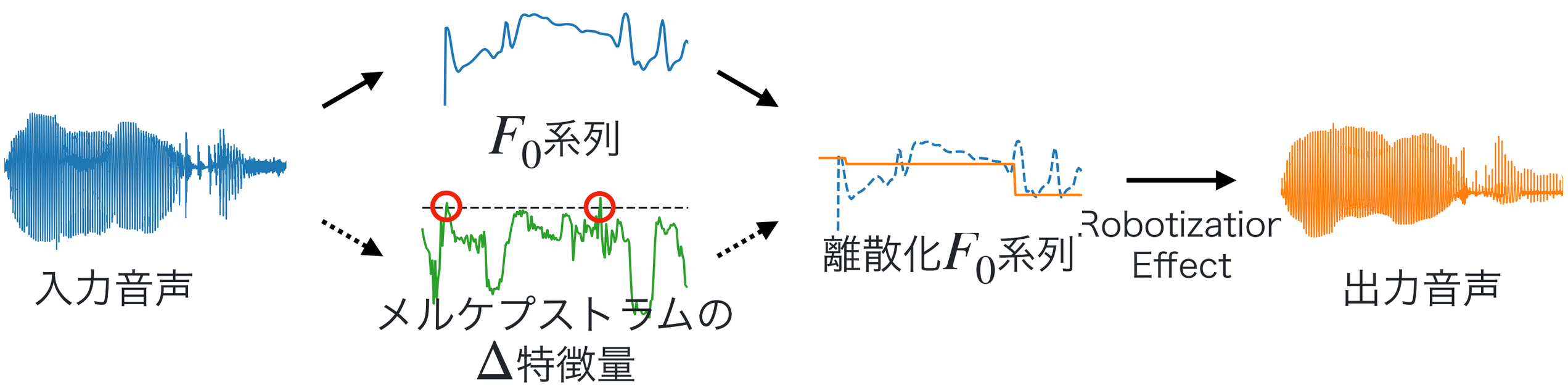
- **従来手法**
  - コムフィルタ：反響しているような音声に変換する手法
  - Robotization Effect：基本周波数を一定に固定する手法
- **提案手法**
  - Robopitch：基本周波数を離散化する手法
  - Inharmonic Warping：音声の金属感を上昇させる手法
- **実験**
  - 音声とロボット画像の主観的合致度実験
  - ロボットの音声デザインを支援するアプリケーションの評価実験（追加実験）
- **まとめ**

提案手法1｜Robopitch

基本周波数を適度に離散化するRobopitchを提案

- メルケプストラムのΔ特徴量は音声の移り変わりを表す

  →Δ特徴量のノルムが閾値を超えたとき区切りを入れる
- それぞれの区間に対して基本周波数の平均を、12平均律で近似したものに置き換える
- 周波数領域：12平均律で離散化、時間領域：音素ごとに離散化
- パラメータ：閾値、robotization effectの窓長

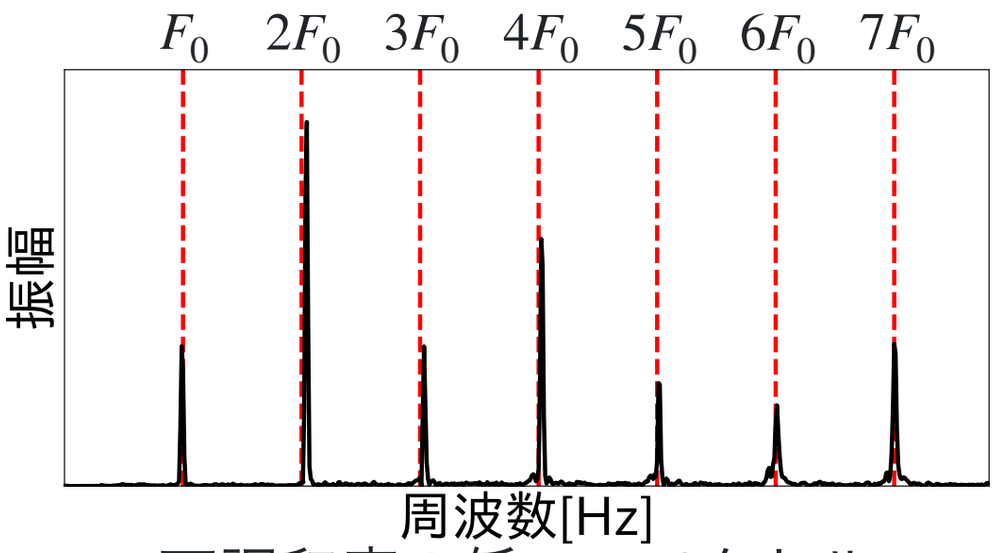


提案手法2｜Inharmonic Warping

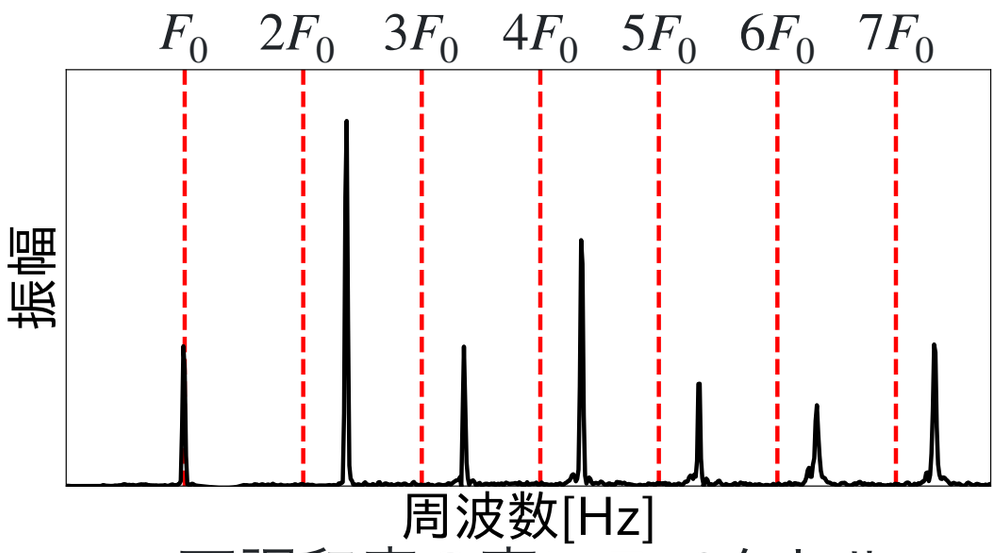
不調和度を上昇させる手法: Inharmonic Warping (IW)

- **不調和度（inharmonicity）**
  - 音の調波構造からの崩れを表す指標で、**音の金属感**を表す
  - 不調和度 I は次式で定義される

$$I = \frac{2}{F_0} \frac{\sum |F_h - h * F_0| A_h^2}{\sum A_h^2}$$



不調和度の低いスペクトル



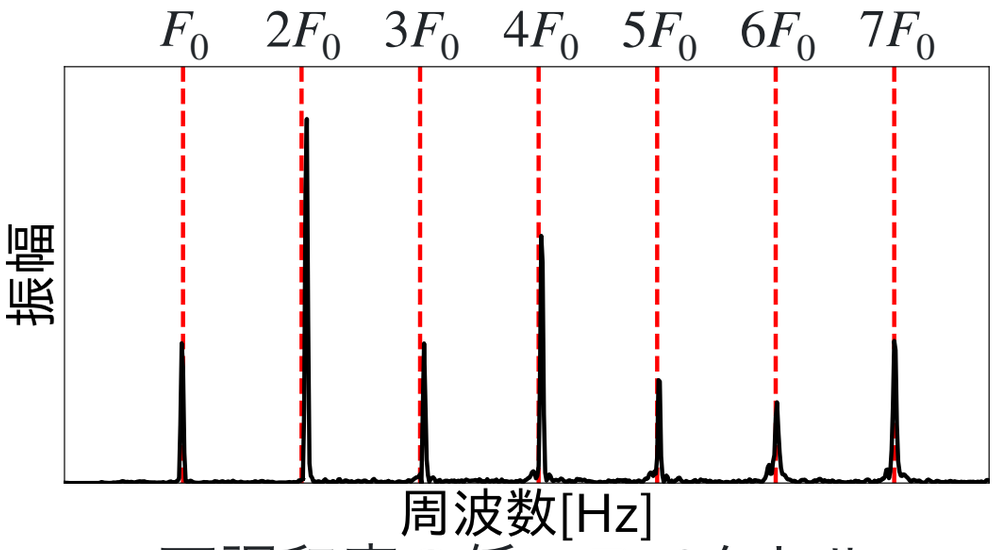
不調和度の高いスペクトル

提案手法2｜Inharmonic Warping

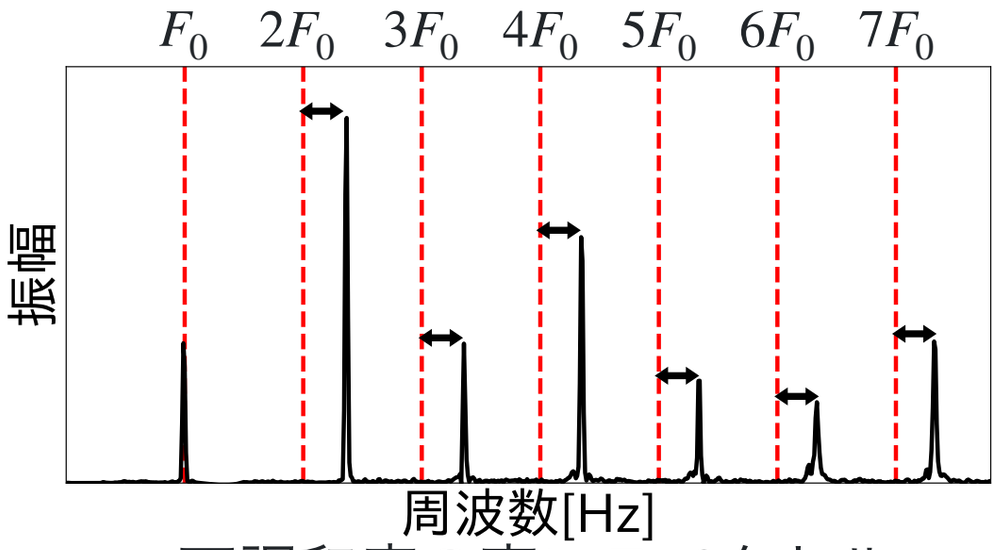
不調和度を上昇させる手法: Inharmonic Warping (IW)

- **不調和度（inharmonicity）**
  - 音の調波構造からの崩れを表す指標で、**音の金属感**を表す
  - 不調和度 I は次式で定義される

$$I = \frac{2}{F_0} \frac{\sum |F_h - h * F_0| A_h^2}{\sum A_h^2}$$



不調和度の低いスペクトル

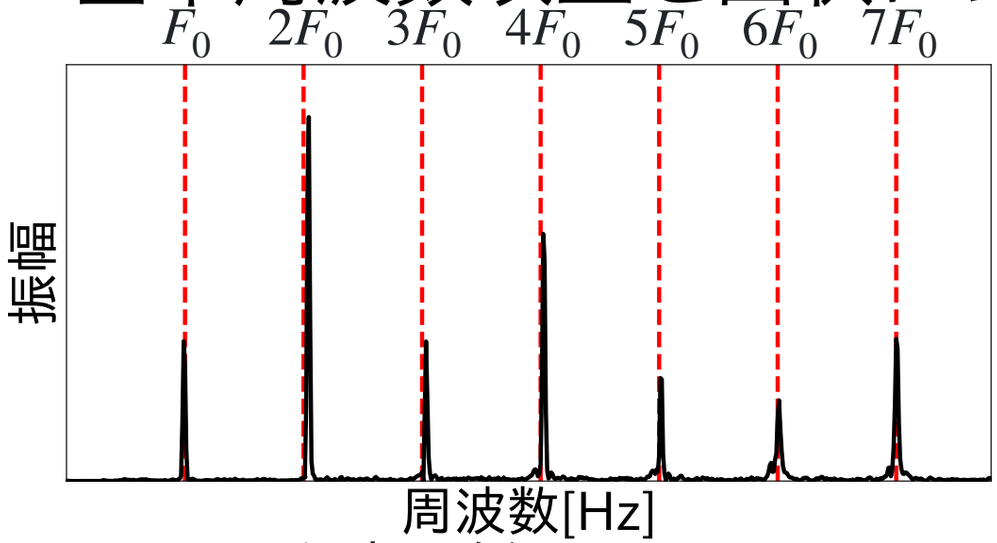


不調和度の高いスペクトル

提案手法2｜Inharmonic Warping

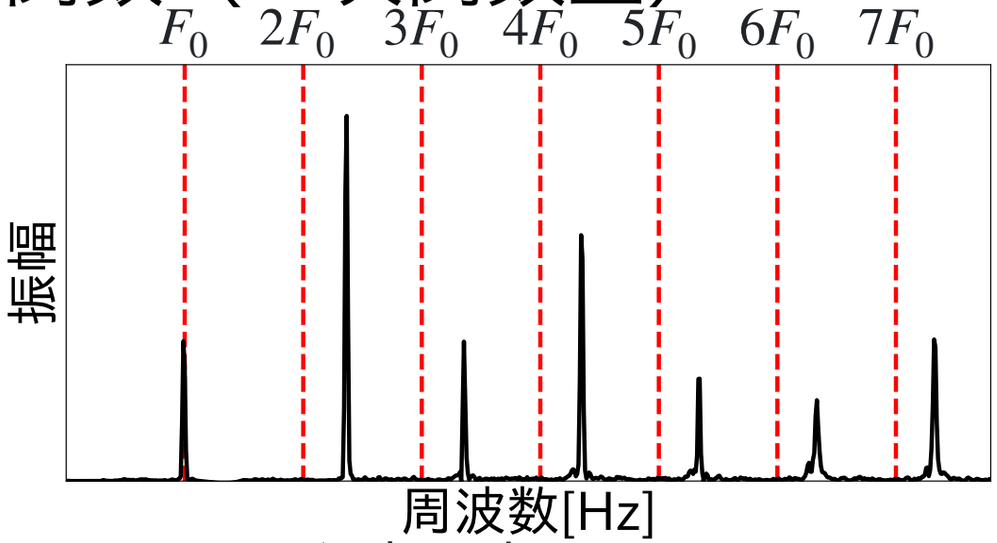
不調和度を上昇させる手法: Inharmonic Warping (IW)

- **スペクトログラムを非線形変換後、位相復元して出力**
- **変換するスペクトル**
  - 振動成分のみ、全ての2種類
- **変換に用いる関数**
  - 基本周波数以上を一定にシフトする関数（一次関数型）
  - 基本周波数以上を凸状にシフトする関数（二次関数型）



不調和度の低いスペクトル

変換

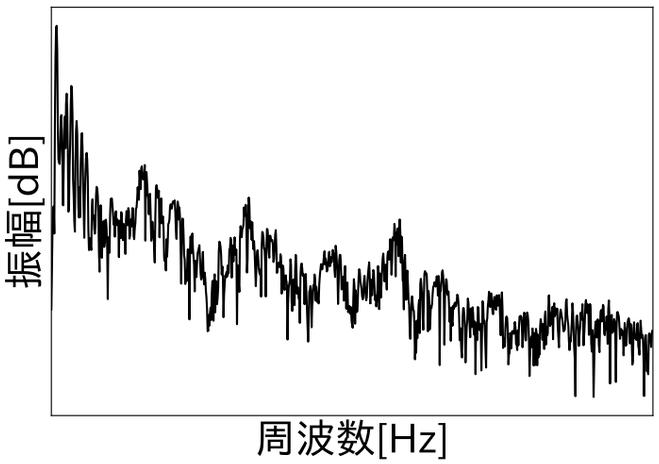


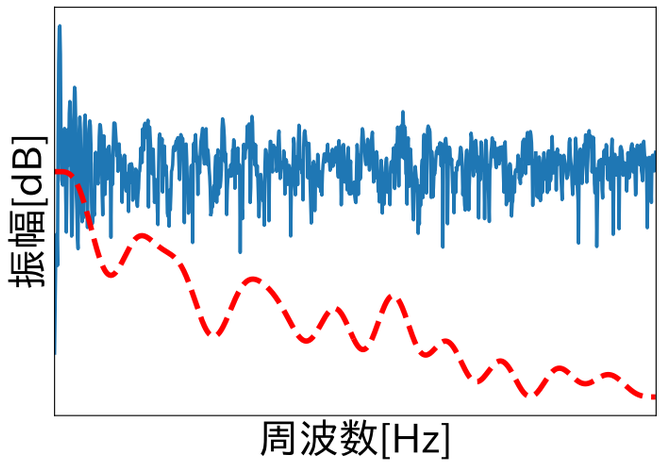
不調和度の高いスペクトル

提案手法2｜Inharmonic Warping

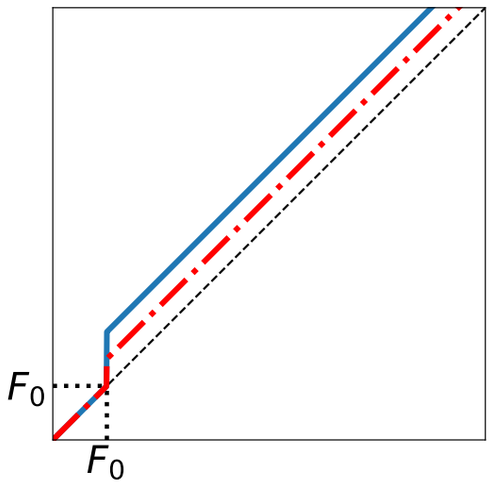
不調和度を上昇させる手法: Inharmonic Warping (IW)

- **スペクトログラムを非線形変換後、位相復元して出力**
- **変換するスペクトル**
  - 振動成分のみ、全ての2種類
- **変換に用いる関数**
  - 基本周波数以上を一定にシフトする関数（一次関数型）
  - 基本周波数以上を凸状にシフトする関数（二次関数型）

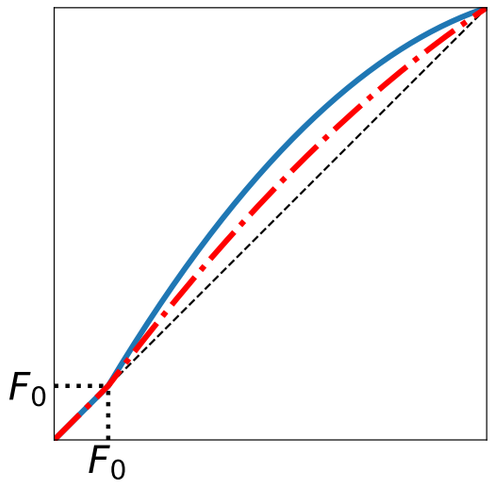




振動成分のみを抽出したスペクトル



一次関数型



二次関数型

目次

- **従来手法**
  - コムフィルタ：反響しているような音声に変換する手法
  - Robotization Effect：基本周波数を一定に固定する手法
- **提案手法**
  - Robopitch：基本周波数を離散化する手法
  - Inharmonic Warping：音声の金属感を上昇させる手法
- **実験**
  - 音声とロボット画像の合致度実験
  - ロボットの音声デザインを支援するアプリケーションの評価実験（追加実験）
- **まとめ**

実験概要

実験1：音声とロボットの主観的合致度の実験

- 人工感を与えた音声とロボットの印象がどの程度合致するか評価
- ロボットの主観的人間らしさの評価

実験2：ロボットの音声デザインを支援するアプリケーションの評価実験

- ロボットの音声デザインを支援するアプリケーションを作成、評価
- 各手法を適用した音声が自然音声と比較してどのように変化したか評価

実験概要｜結果目次

- 各手法の有効性
- 手法のパラメータの変化と主観的合致度の関係
- ロボットの主観的人間らしさの結果
- 手法とロボットの主観的人間らしさの関係
  - 手法が与える音声変化の印象
- アプリケーション評価

実験1｜音声とロボットの主観的合致度実験

音声コーパス	日英・日中バイリンガル独話音声データベースから男性話者5人、女性話者5人を選択 日本語の合文法無意味文を使用 サンプリング周波数48 kHz
被験者	音声とロボットの各ペアに対して50人が参加、クラウドソーシングサービスで募集
ロボット画像	以下の画像を使用、主観評価実験から主観的人間らしさを評価
評価方法	音声とロボット画像を同時に提示し、印象がどの程度合致するかをリッカート尺度の5段階で評価
手法	Robopitch：閾値8種類、窓長3種類の計24種類 IW：スペクトル2種類、関数2種類、パラメータ4種類の計16種類 コムフィルタ：frequency4種類、decay4種類の計16種類 複数手法の組み合わせ：robopitch、IW、コムフィルタを組み合わせた20種類 自然音声とGriffin–Lim法のみ適用した音声と合わせて58種類


Poli


PR2


G5


Flash


Stevie


HUBO-SCIPRR


Pepper

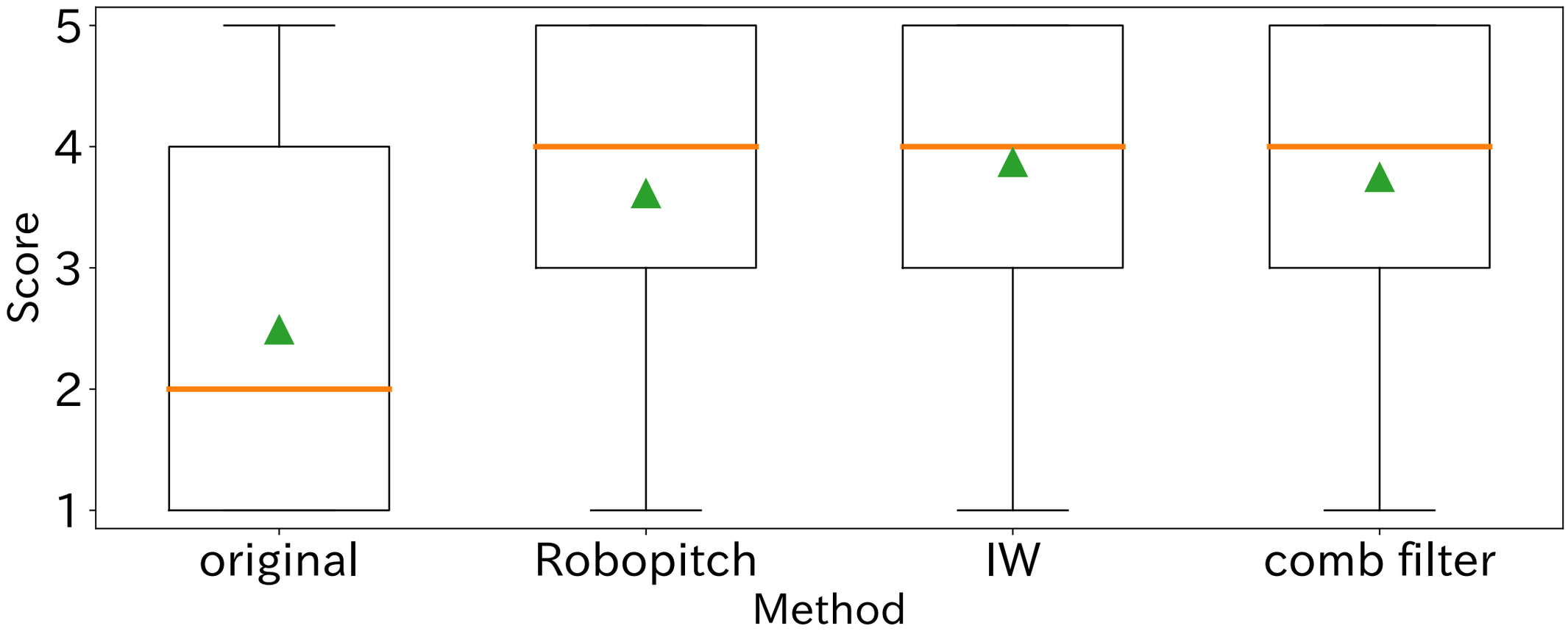

iCub

各ロボットの画像、主観的人間らしさの昇順に配置した

結果｜各手法の有効性

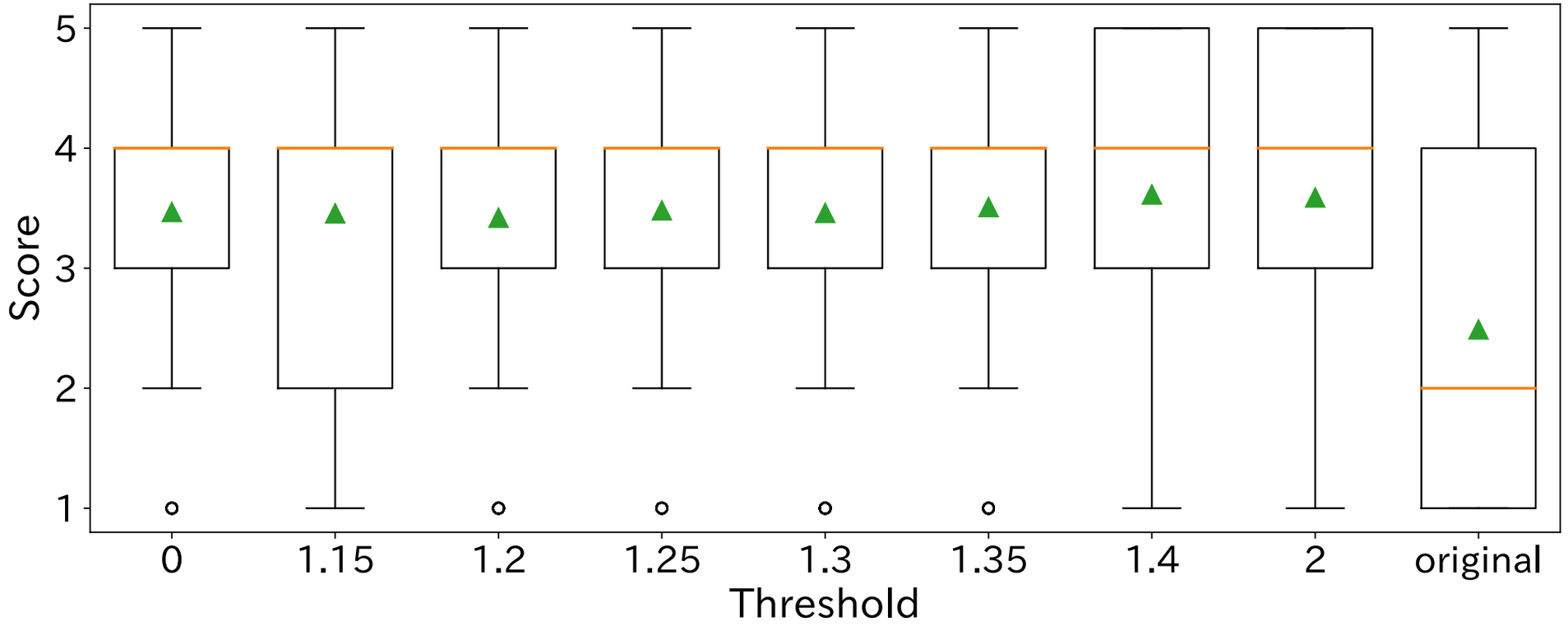
- **Robopitch、IW、コムフィルタは自然音声より有意に主観的合致度が高かった**

  →三手法ともロボットの印象と合致する印象を音声に与える



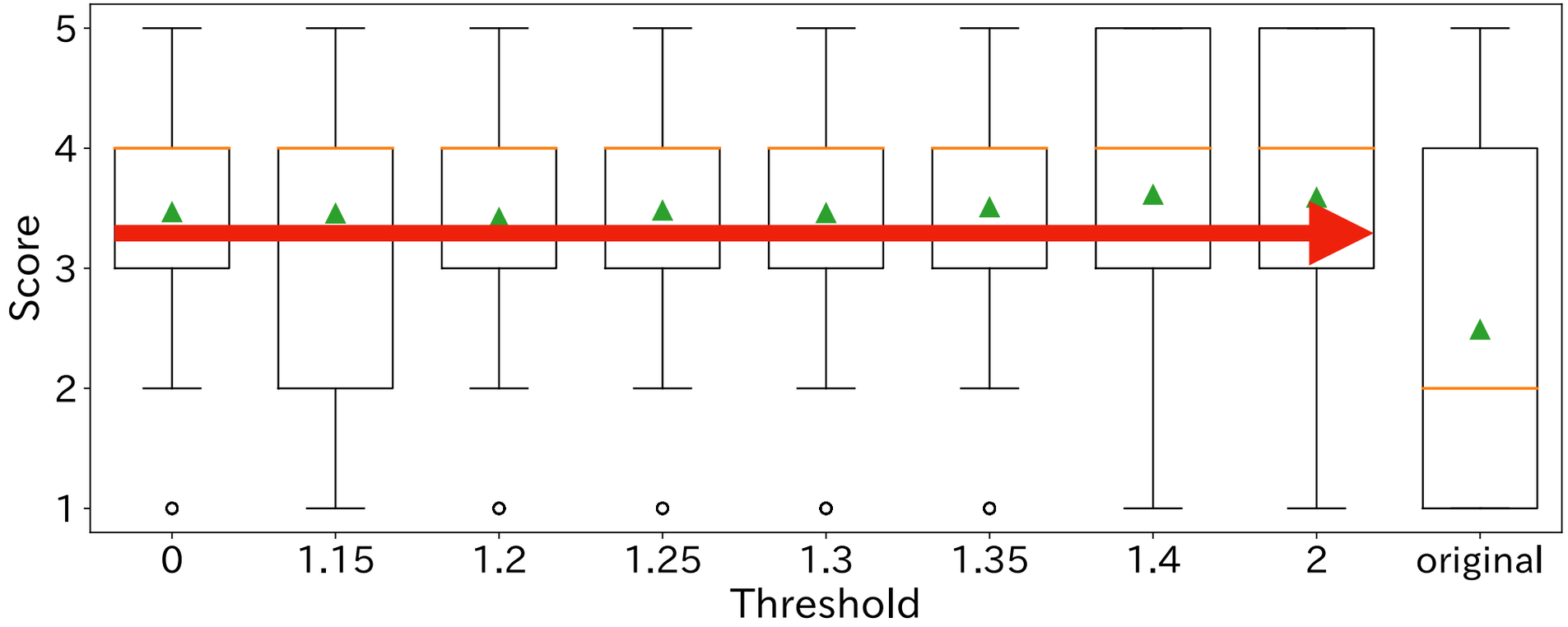
結果｜Robopitch

- **閾値間で有意差なし**
  - 閾値：基本周波数離散化の時間領域の細かさを変化させる
  - 了解性が高くなるように閾値を設定すると良いと示唆



結果｜Robopitch

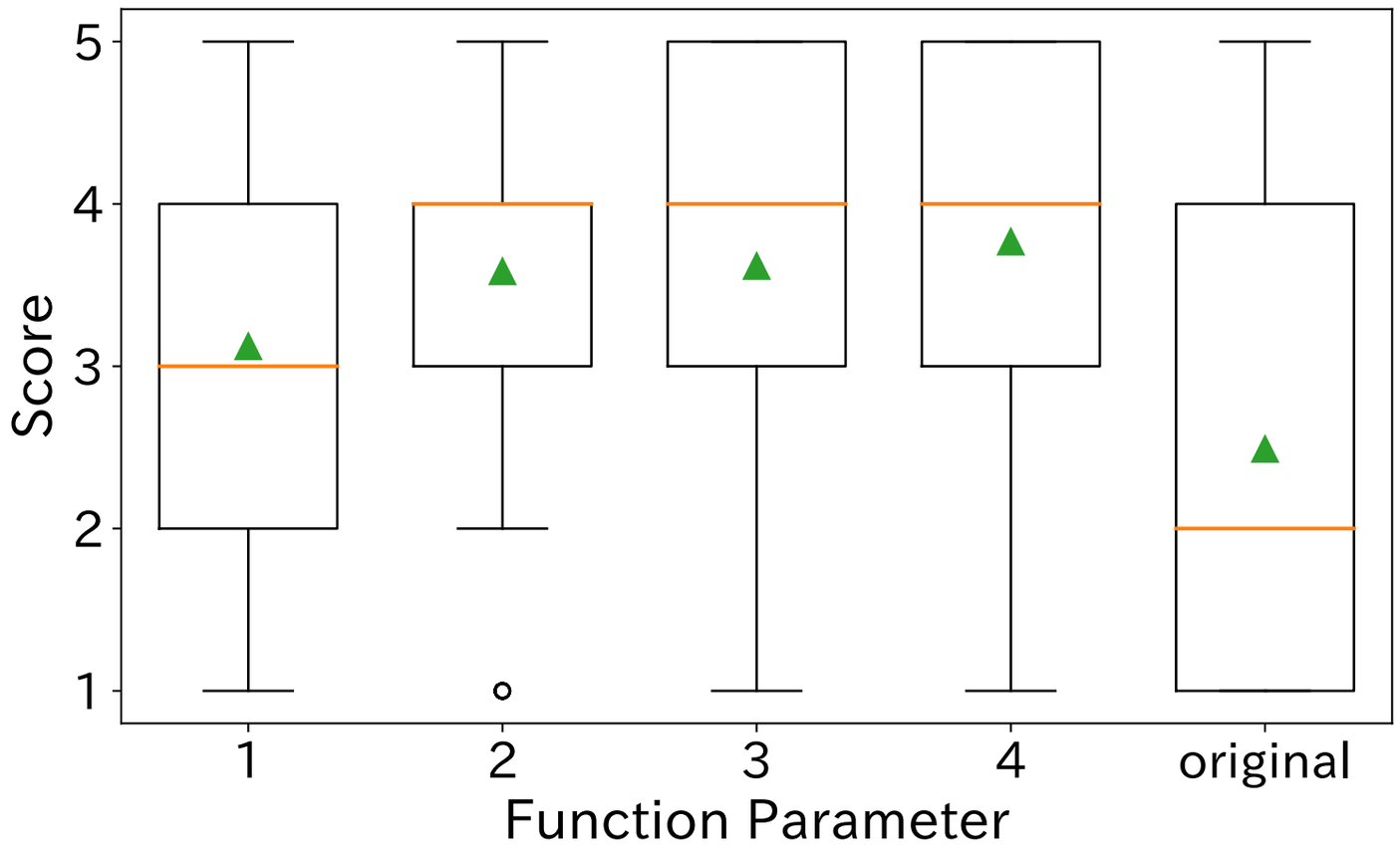
- **閾値間で有意差なし**
  - 閾値：基本周波数離散化の時間領域の細かさを変化させる
  - 了解性が高くなるように閾値を設定すると良いと示唆



結果 | Inharmonic Warping

- **元の周波数からの変化が大きいほど、主観的合致度が高かった**

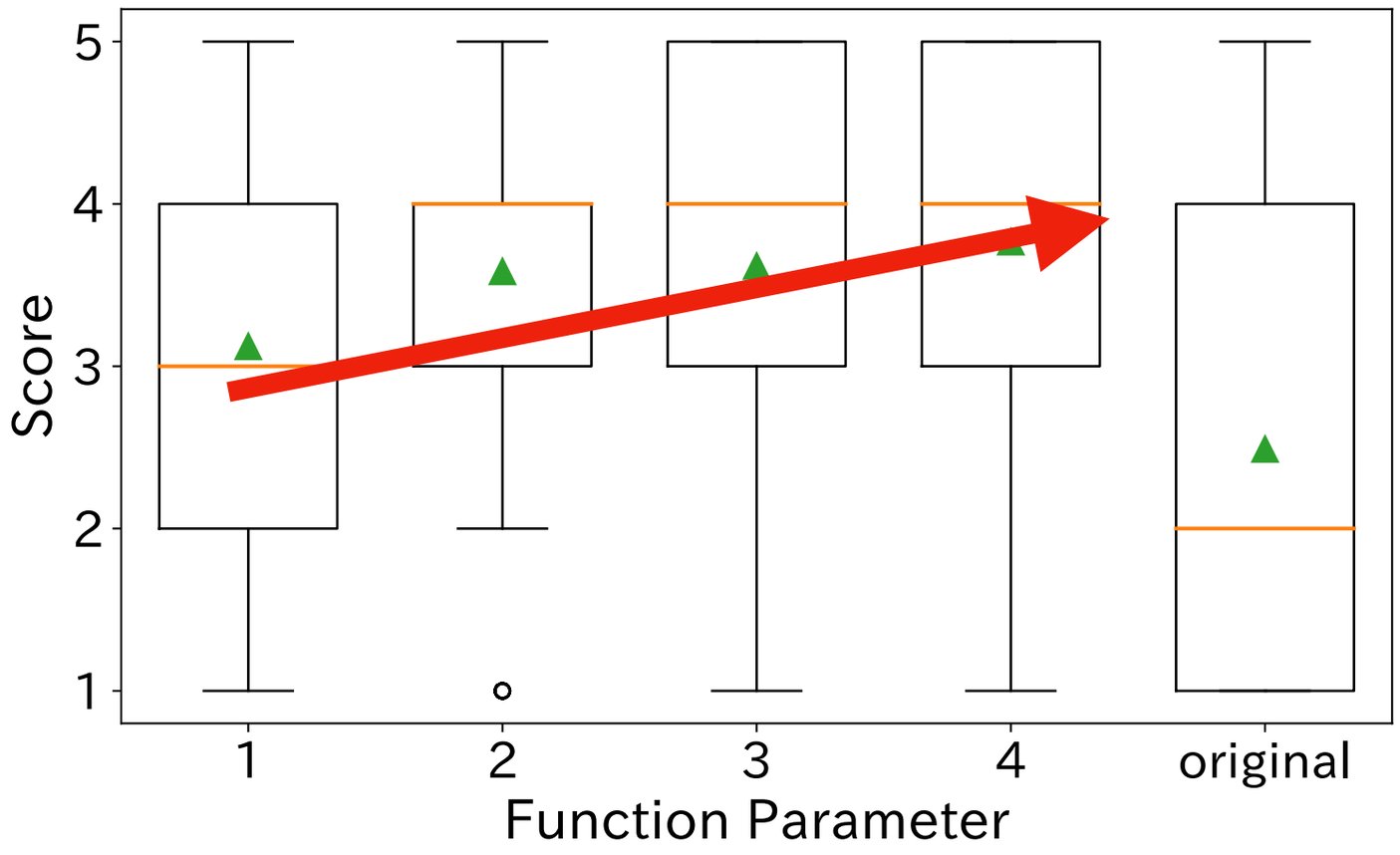
  →手法の目的に合致



結果｜Inharmonic Warping

- **元の周波数からの変化が大きいほど、主観的合致度が高かった**

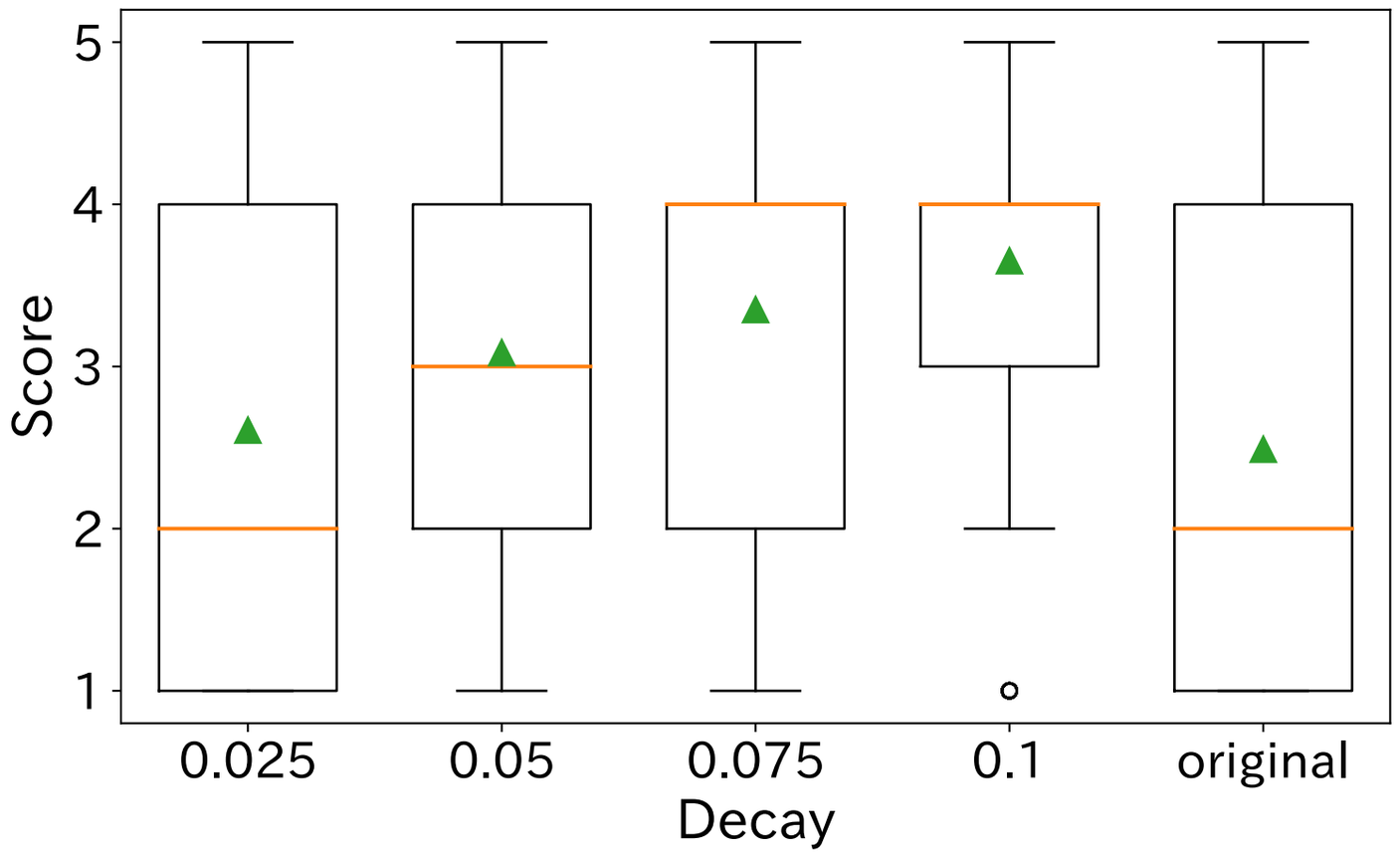
  →手法の目的に合致



結果｜コムフィルタ

- **Decayが大きいほど主観的合致度が高かった**

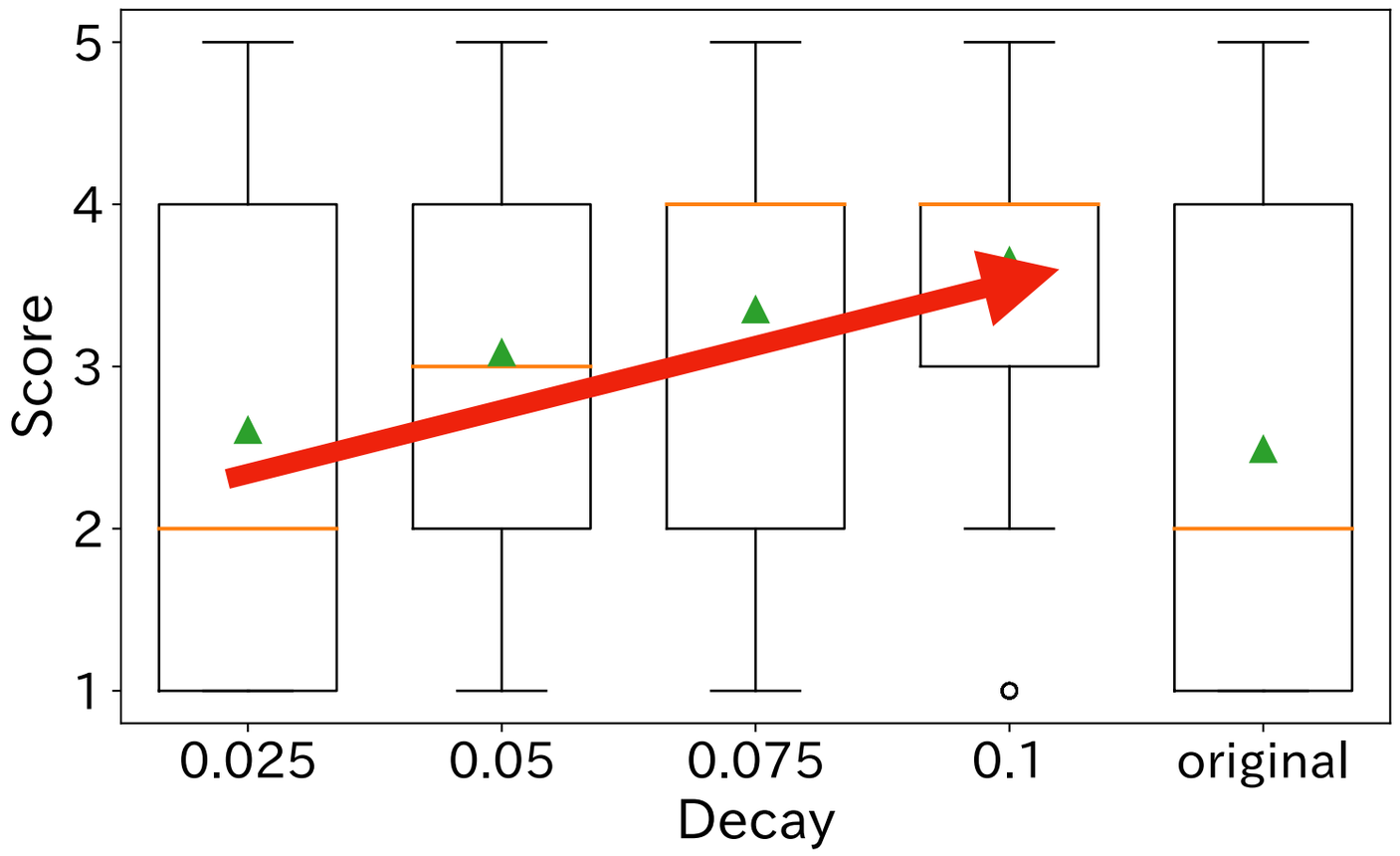
　→Decayが大きいほど周波数応答のピークが強くなり、コムフィルタの効果が大きい



結果｜コムフィルタ

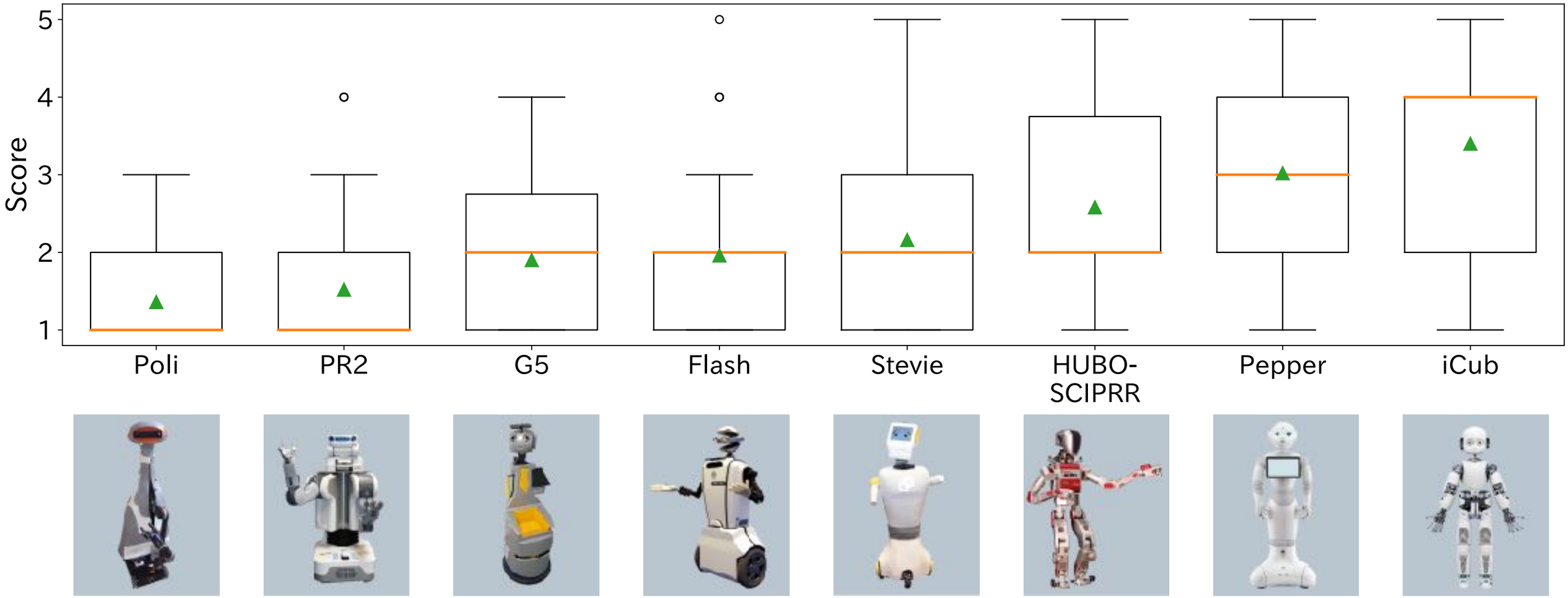
- **Decayが大きいほど主観的合致度が高かった**

　→Decayが大きいほど周波数応答のピークが強くなり、コムフィルタの効果が大きい



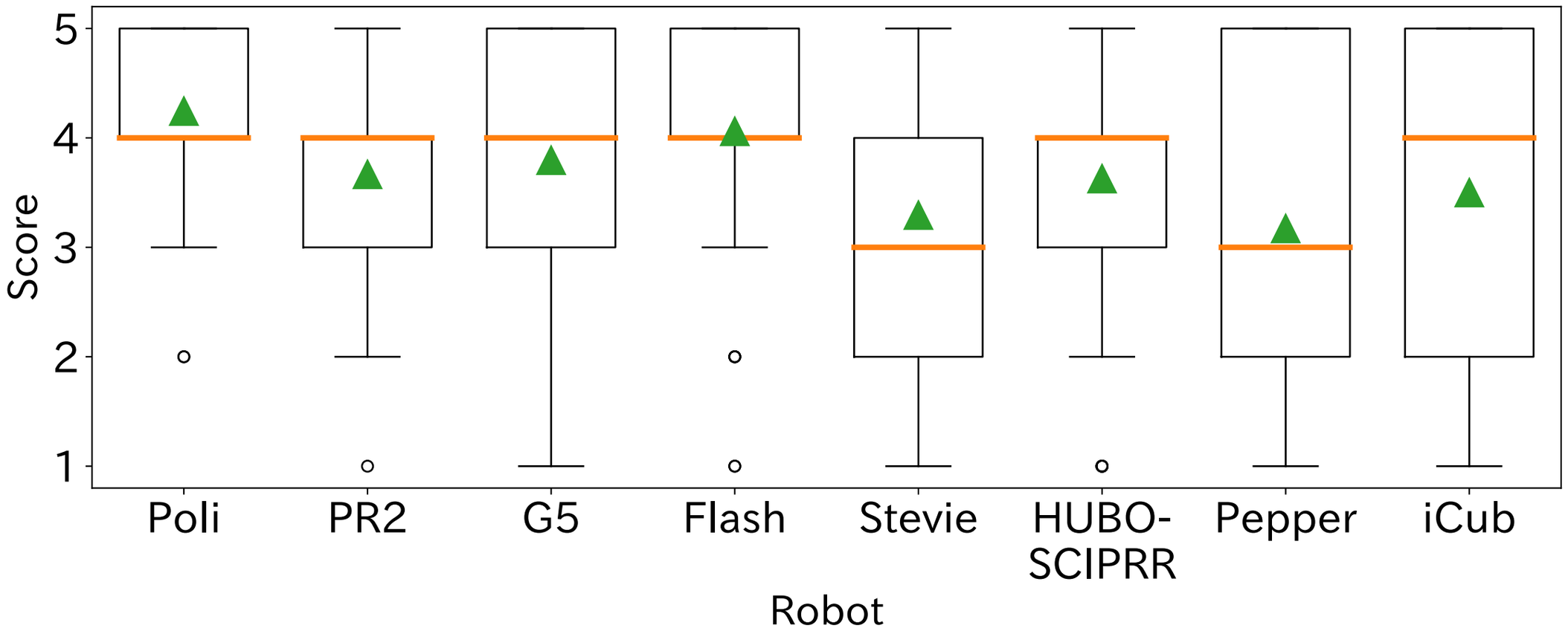
結果｜ロボットの人間らしさ

- 足のあるHUBO-SCIPRRやiCubの主観的人間らしさが高かった
- 認知度の高いPepperも主観的人間らしさが高かった



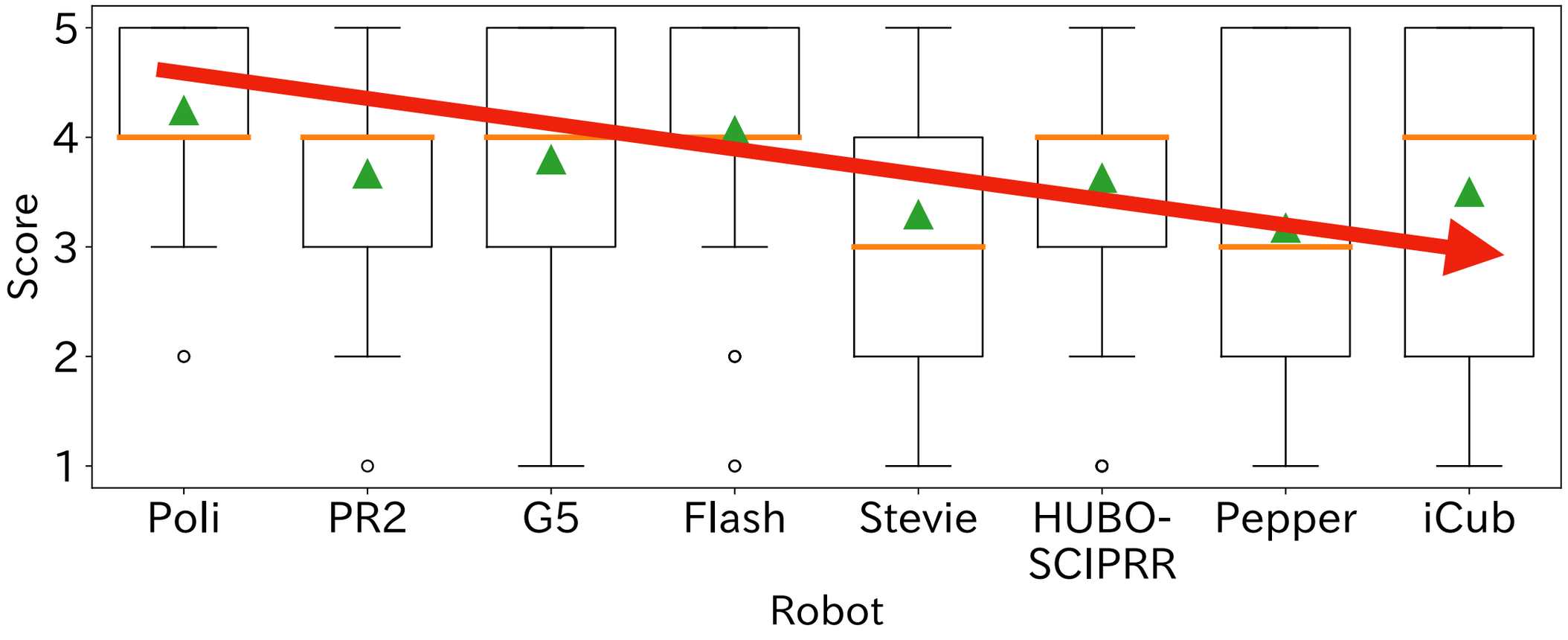
結果｜Robopitch

- 主観的人間らしさが低いロボットと相性が良い
- 「機械的な音声に変換」「閾値を変えると抑揚が変化」



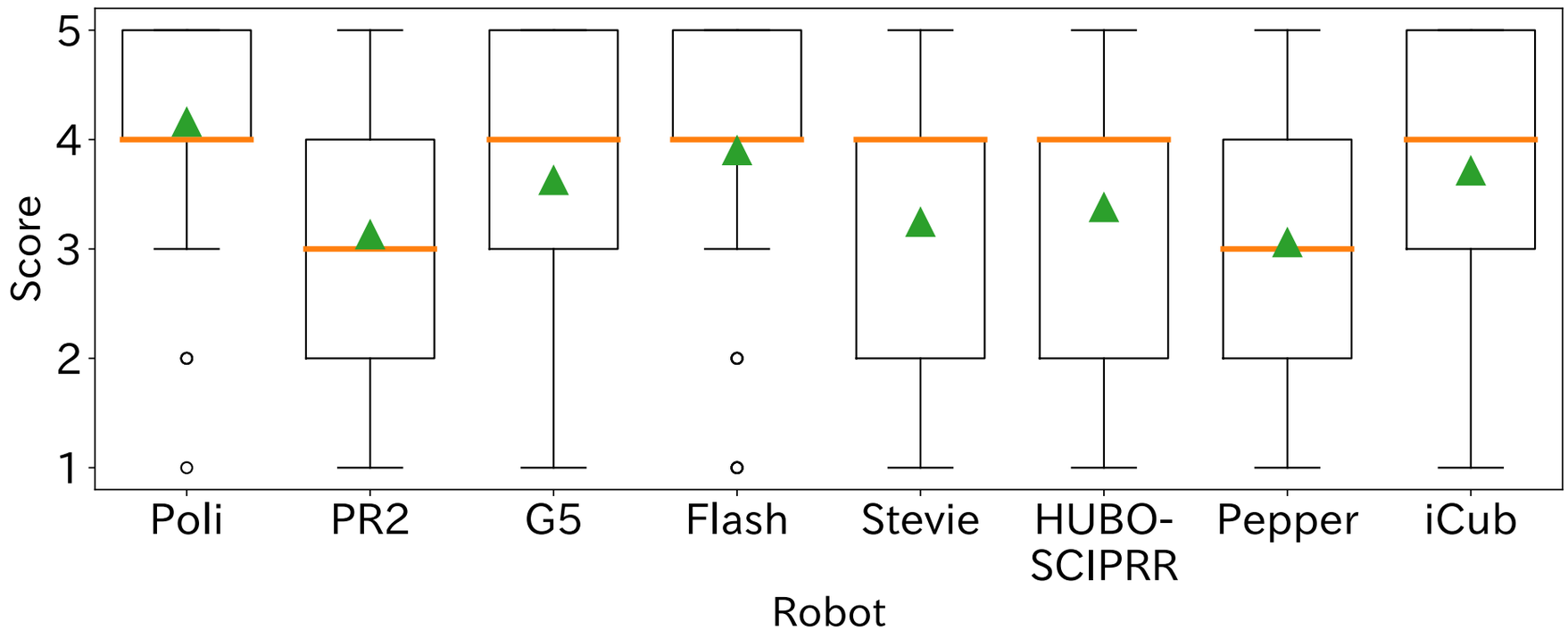
結果｜Robopitch

- 主観的人間らしさが低いロボットと相性が良い
- 「機械的な音声に変換」「閾値を変えると抑揚が変化」



結果｜Inharmonic Warping

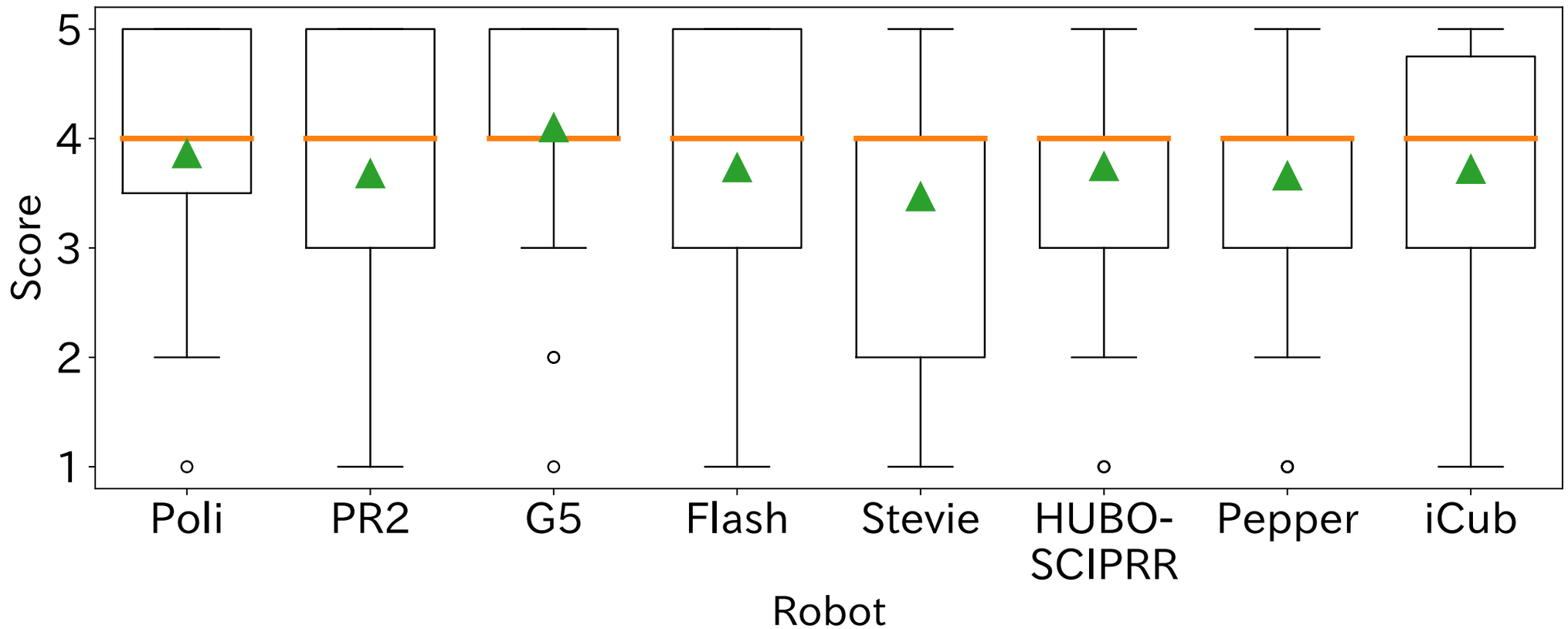
- 一貫した傾向は見出せず
- 「音声に厚みが出て、金属的な印象」
  →手法の目的に合致
- 「ピッチが高く変化した」
  →変換に用いる関数の更なる検討が必要



結果｜コムフィルタ

- ロボット間の相性の差があまりない
- “反響音を加えたような音声” ”人工的とは感じられない”

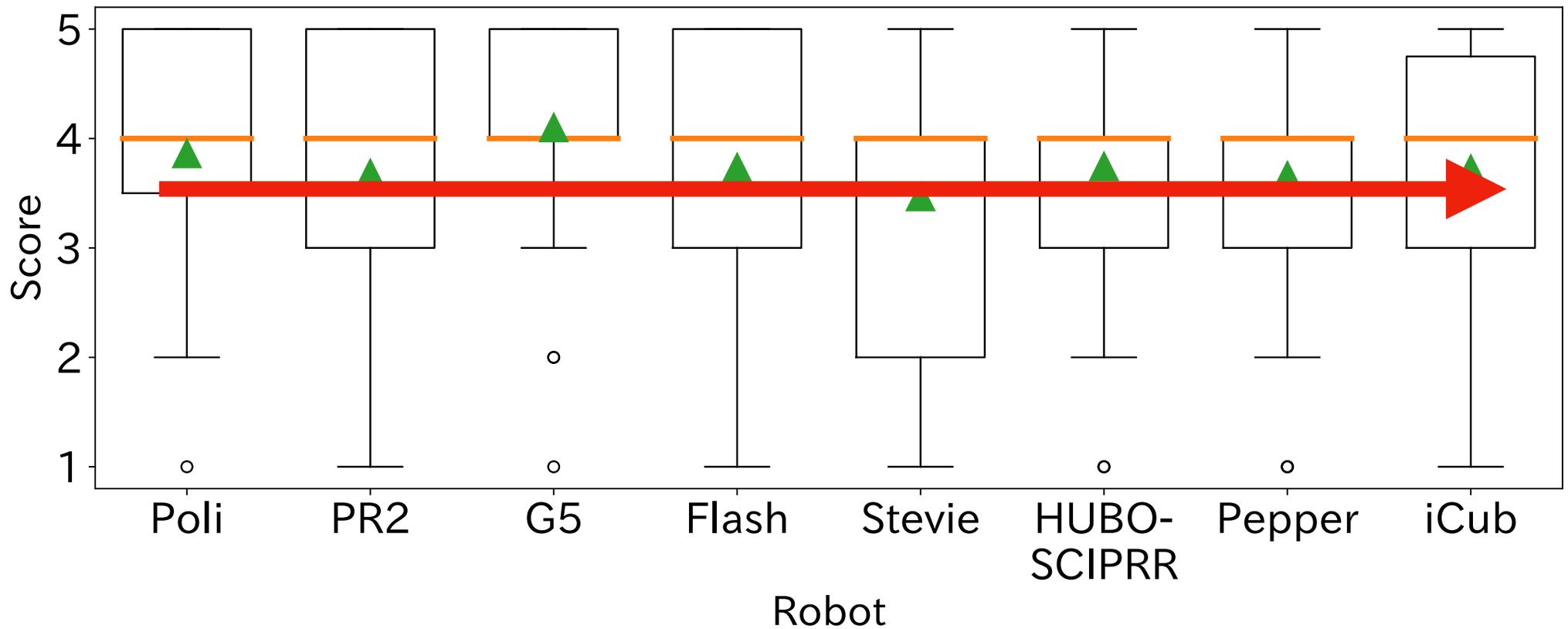
  →反響音と人工感の印象が合わない



結果｜コムフィルタ

- ロボット間の相性の差があまりない
- “反響音を加えたような音声” ”人工的とは感じられない”

  →反響音と人工感の印象が合わない



実験2｜ロボット音声デザイン支援の実験

ロボットの音声デザインを支援するアプリケーション

Artificializer

- **各手法のパラメータを、ユーザが自由に選択可能な簡易的なアプリケーション**
  - 実時間処理は行わず、音声はあらかじめ用意したものとした
- **表示されたロボット画像と音声の印象が合致するようにパラメータを操作**

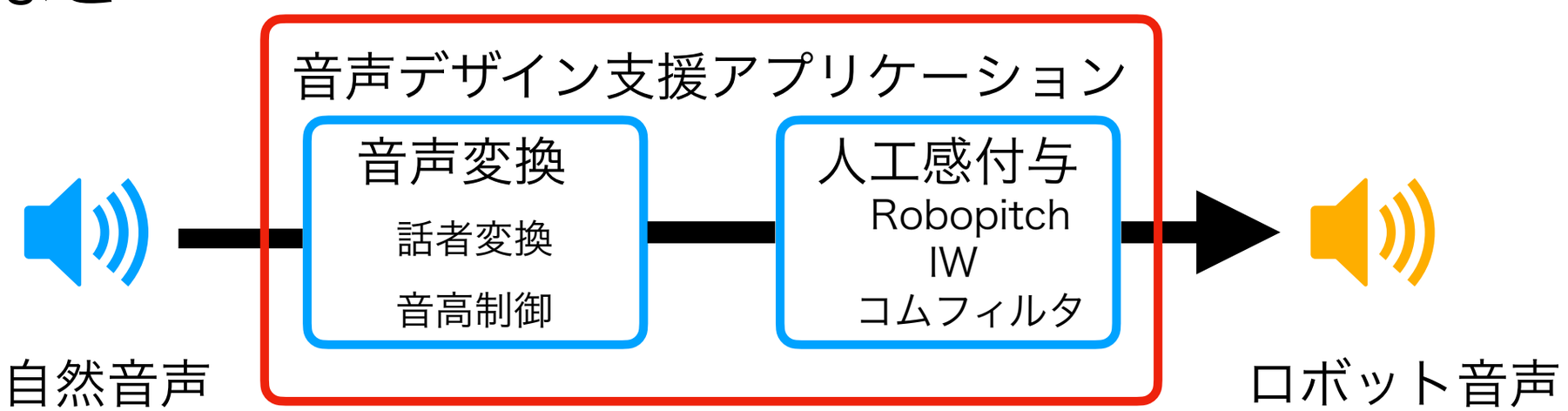
実験2｜ロボット音声デザイン支援の実験

音声コーパス	日英・日中バイリンガル独話音声データベースから男性話者5人、女性話者5人を選択 日本語の合文法無意味文を使用 サンプリング周波数48 kHz
被験者	50人が参加、クラウドソーシングサービスで募集
ロボット画像	以下の画像を使用、あらかじめ主観的人間らしさを評価
評価方法	アプリケーションの使いやすさをリッカート尺度で評価、使用した感想を自由記述 各手法を適用した音声の印象を自由記述
手法	Robopitch：閾値8種類、窓長3種類の計24種類 Inharmonic Warping：スペクトル2種類、関数2種類、パラメータ4種類の計16種類 コムフィルタ：frequency4種類、decay4種類の計16種類 自由にパラメータを選択可能

各ロボットの画像

実験2｜ロボット音声デザイン支援の実験

- “男性話者、女性話者の切り替えをしたい” “声の高さを変えたい”：音声変換との組み合わせを検討
- “音声を自由に変更したい” “パラメータをより細かく調整したい”：実時間処理が問題
- “パラメータの変化が見づらい”
  - 音声の変化の視覚化
    - Robopitchの基本周波数、IWの関数、コムフィルタの周波数応答など



目次

- **従来手法**
  - コムフィルタ：反響しているような音声に変換する手法
  - Robotization Effect：基本周波数を一定に固定する手法
- **提案手法**
  - Robopitch：基本周波数を離散化する手法
  - Inharmonic Warping：音声の金属感を上昇させる手法
- **実験**
  - 音声とロボット画像の主観的合致度実験
  - ロボットの音声デザインを支援するアプリケーションの評価実験（追加実験）
- **まとめ**

まとめ

- **各手法が音声に与える人工感とパラメータの関係を分析**
  - 基本周波数を離散化する手法、音声の金属感を上昇させる手法を提案
  - Robopitch： パラメータ間で大きな差なし
  - Inharmonic Warping、コムフィルタ：パラメータと主観的合致度に単調な関係あり
- **ロボットの音声デザイン支援**
  - 音声デザイン支援アプリケーションは実時間処理、視覚化が課題

今後の展望

- **ロボットの外見の特徴をより詳細に抽出**
  - ロボットの外見から適切な手法とパラメータを推定する

    自動音声デザインの実現

まとめ

根拠のある音声デザインのために

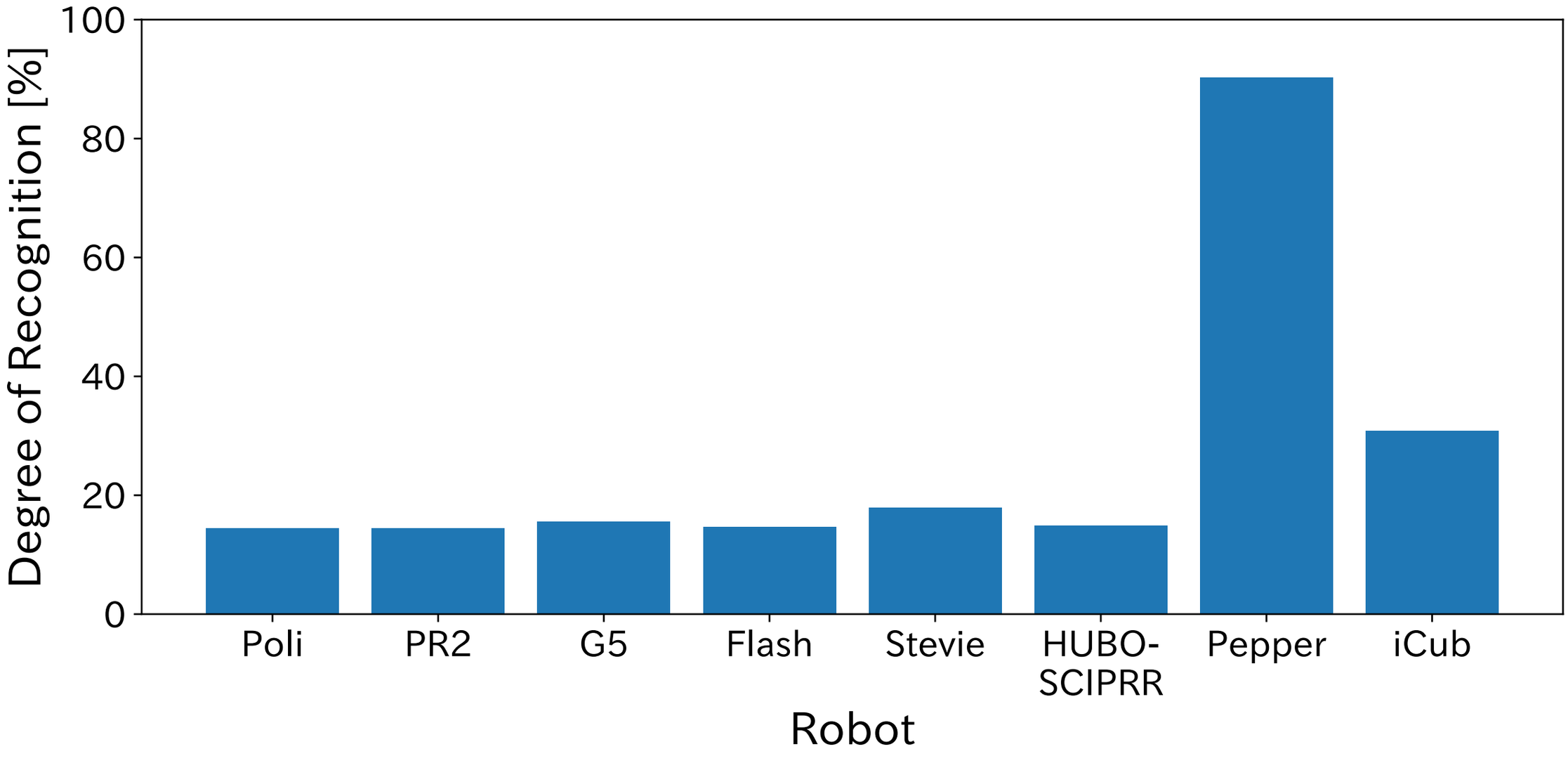
- **エージェントに調和した音声にむけて**
  - 吹き替える話者を選ぶのではなく調整して合わせる
- **声質変換技術による話者性の操作**
  - 声質変換によって話者情報を操作する：存在する話者へ
- **人工感制御技術による人工感の操作**
  - エージェントが出しているような声に加工する：存在しない声へ

ロボット自体のデザインも含めた
総合的なエージェントデザインへ

謝辞

- 本発表中の研究の一部は科研費（21H04900）の助成を受けたものである.
- 日頃ディスカッションしてくれる研究室メンバー各位に感謝申し上げます.

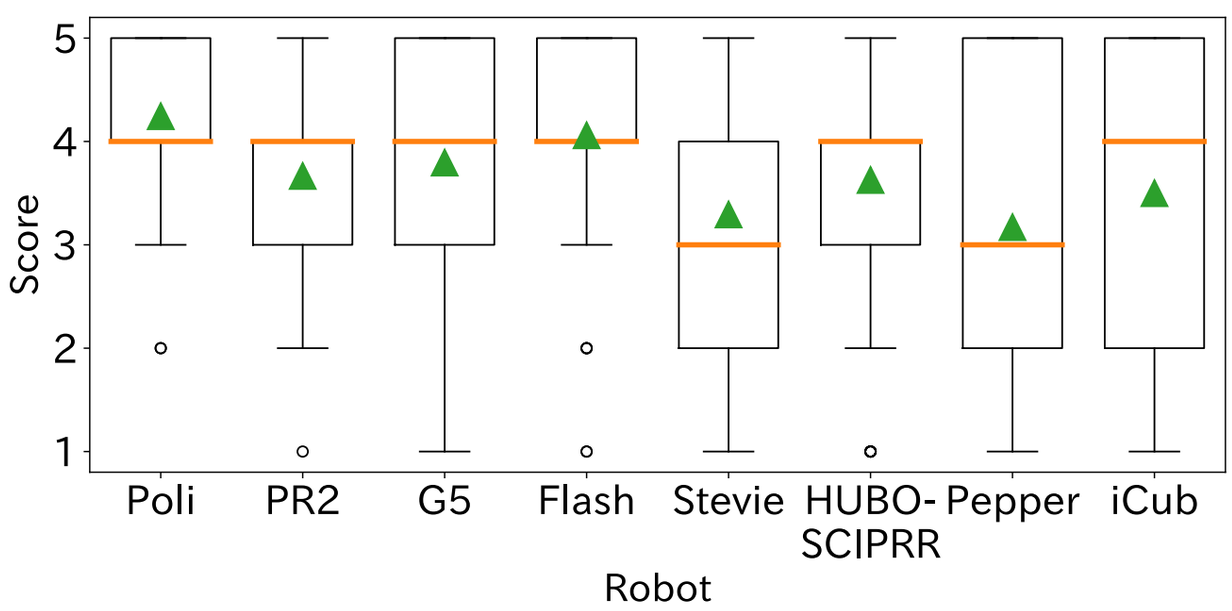
結果｜ロボットの認知度


Degree of Recognition [%]
100
80
60
40
20
0
Poli
PR2
G5
Flash
Stevie
HUBO-SCIPRR
Pepper
iCub
Robot

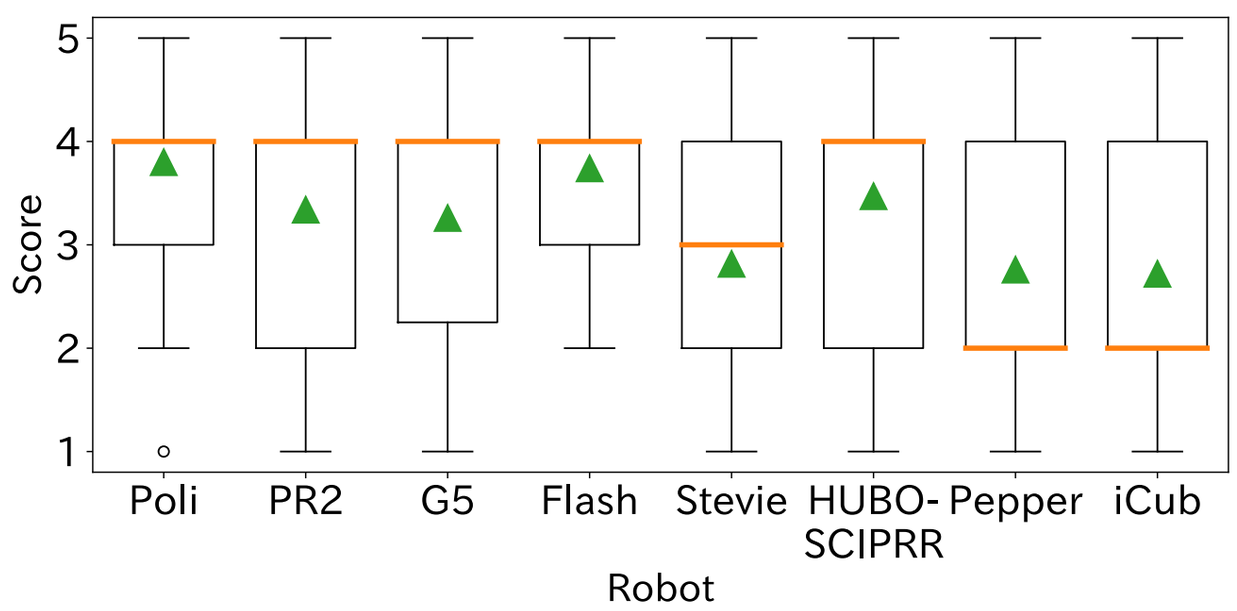
結果｜複数手法の組み合わせ

- **主観的合致度：単一手法 ≧ 複数手法**

  →複数手法の組み合わせは有効とはいえず
- **Robopitchの傾向が強く表れる**
- **相性の良い手法のみ適用する方が主観的合致度は高い**



Robopitchのみ

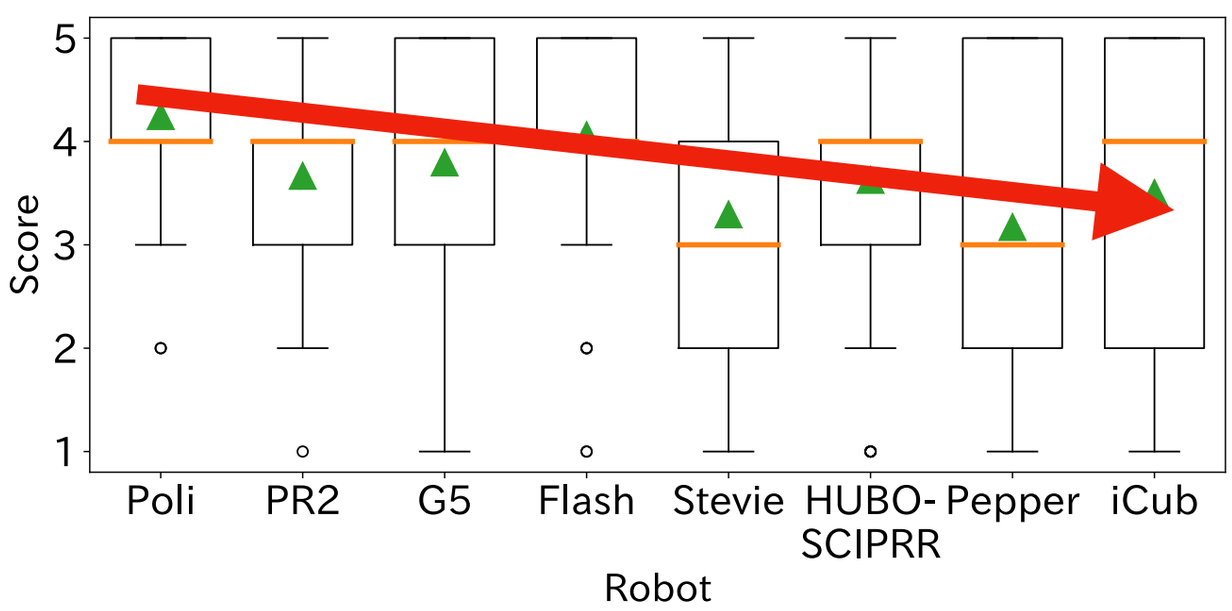


Robopitch、Inharmonic Warping、コムフ

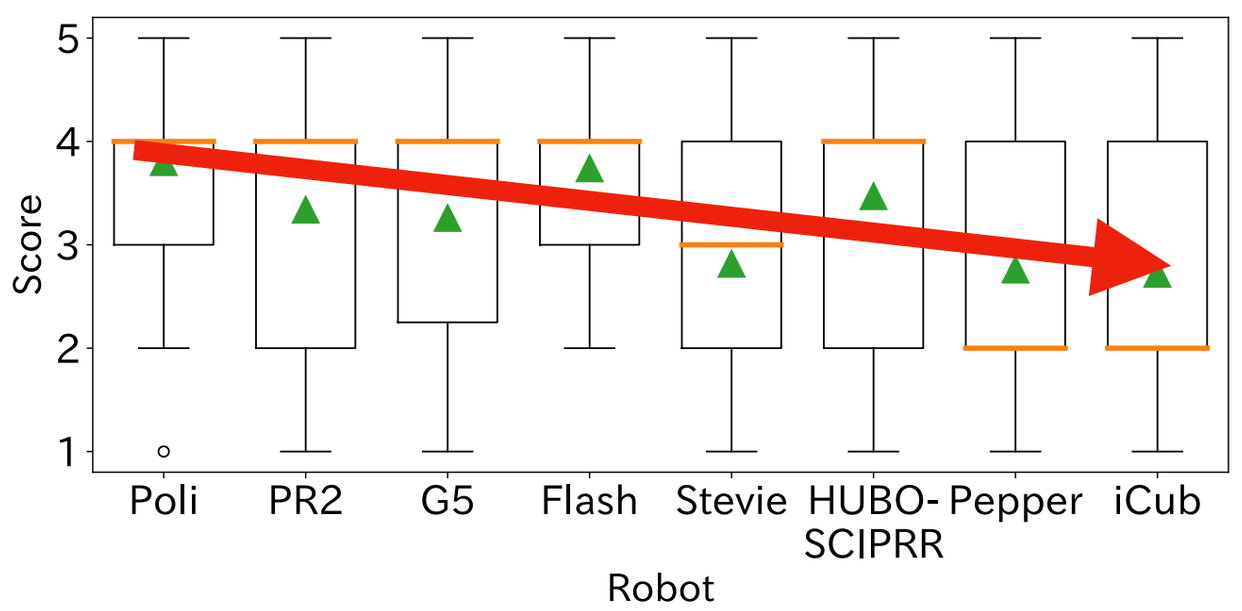
結果｜複数手法の組み合わせ

- **主観的合致度：単一手法 ≧ 複数手法**

  →複数手法の組み合わせは有効とはいえず
- **Robopitchの傾向が強く表れる**
- **相性の良い手法のみ適用する方が主観的合致度は高い**



Robopitchのみ



Robopitch、Inharmonic Warping、コムフ

補足｜メルケプストラム離散化

- **メルケプストラム離散化 [Kobayashi+, 2015]**
  - メルケプストラムを用いて混合ガウス分布（GMM）を学習
  - 事後確率最大のガウス分布の平均で置き換える
- **隠れマルコフモデル（HMM）に基づく音声合成を再現**