2-3Q-1　残差文埋め込みを用いた連続行動空間に基づく音声言語獲得エージェント

Continuous action space-based spoken language acquisition agent using residual sentence embedding

☆小松亮太（東工大），△木村友祐（東工大），岡本拓磨（NICT），篠崎隆宏（東工大）

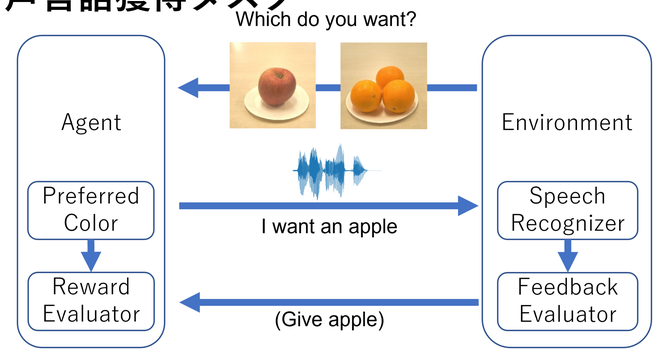
1. 研究目的

人間と同等の音声言語学習能力を備えたエージェントを計算機上で実現

2. 音声言語獲得の仕組みに関する議論

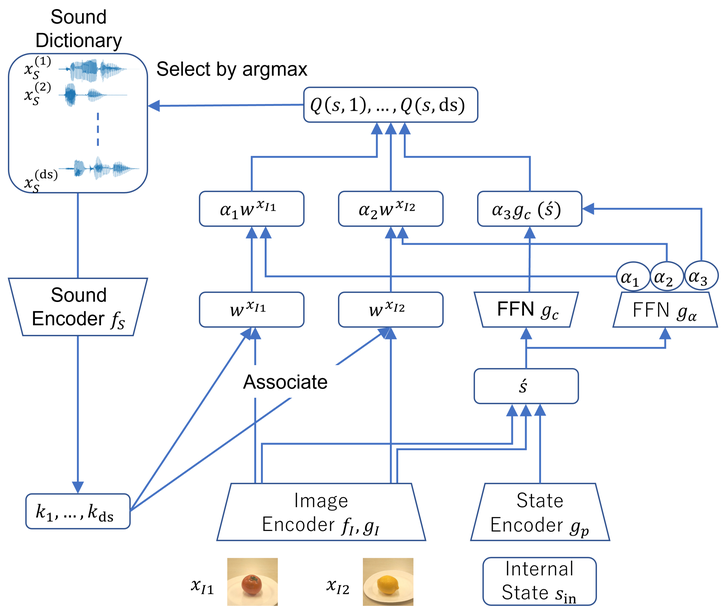
Skinnerは強化学習に基づいていると述べ，Chomskyはそれのみでなく周囲の人間の観察にもよると主張

3. 音声言語獲得タスク



- 観察学習
  画像を提示されながらその音声説明を聞き，**音声と画像の関連性を学習**
- 対話学習
  エージェントは好きな色を内部状態として持っており，**質問に対し適切な方を”I want a/an <food>”と発話**すると報酬1，それ以外の場合報酬0を得る

4. 先行研究とその課題



- 従来法では**離散的な音声辞書から要素選択**して発話生成
- しかし，発話方法は録音再生であるため，自ら単語を組み合わせて発話文を生成できない制約がある

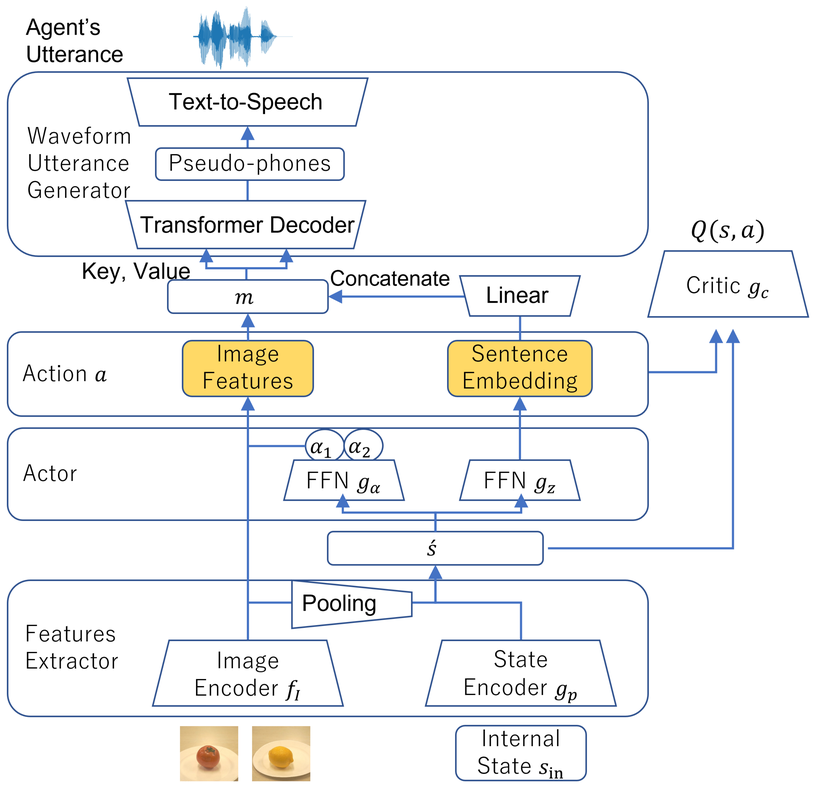
5. 提案法 [1]: 言語モデルを用いて複数単語発話生成

音声辞書を言語モデルで置き換える単純な方法として，画像特徴量で言語モデルを条件付けることが考えられる

しかし一般に画像と関連する文は複数あるため，それだけではエージェントの内部状態などのコンテキストに応じた文指定ができない

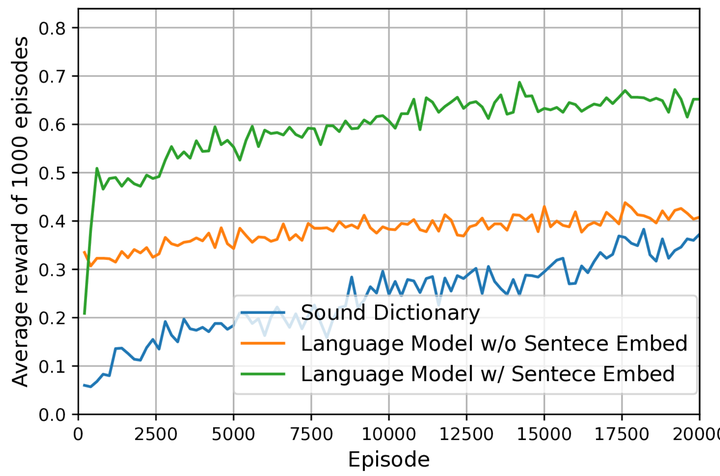
アイデア
文指定のための残差情報として
文埋め込みを画像特徴量に付加

- 観察学習
  画像特徴量抽出器や言語モデルと文埋込空間，および音声合成器はラベル無しデータで事前学習
- 対話学習
  連続行動空間を用いた強化学習でアクター・クリティック を学習



6. 実験

- 文クローズド条件
  - 食べ物20種類
  - 音声説明は”<food>”，”A/An <food>”，”A/An <color> <food>”，”I want a/an <food>”の4種類
  - 結果



文を特定するための情報を低次元の残差埋込ベクトルとして生成するこで，高速な学習を実現

- 文オープン条件
  - 言語モデル・強化学習モジュールの学習において特定の色タグと食べ物の組を取り除く

  - 結果

Removed pair	Average reward
red, onion	0.0024
green, sweet potato	0.24
blue, white radish	0.42

初めて見る組に対しても平均約20%の割合で単語を組み合わせて発話生成できる

7. まとめ

- 残差文埋め込みを用いて複数単語発話生成するエージェントを提案
- 今後の課題としては，言語モデルの事前学習による zero-shotでの発話生成の改善が挙げられる

参考文献

[1] Komatsu+, IEEE ICASSP, 2023 (accepted).

謝辞

本研究はJSPS科研費 JP22K12069の助成を受けたものです