IEEE JSTSP paper

3-Q-11　実音声を用いた音声言語獲得エージェントの評価

Evaluation of Spoken Language Acquiring Agent Using Real Voices

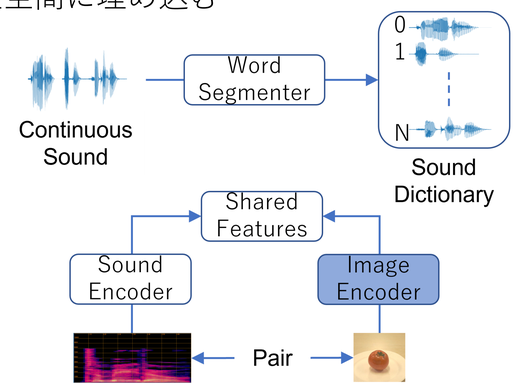
☆小松亮太（東工大），岡本拓磨（NICT），篠崎隆宏（東工大）

1. 研究目的

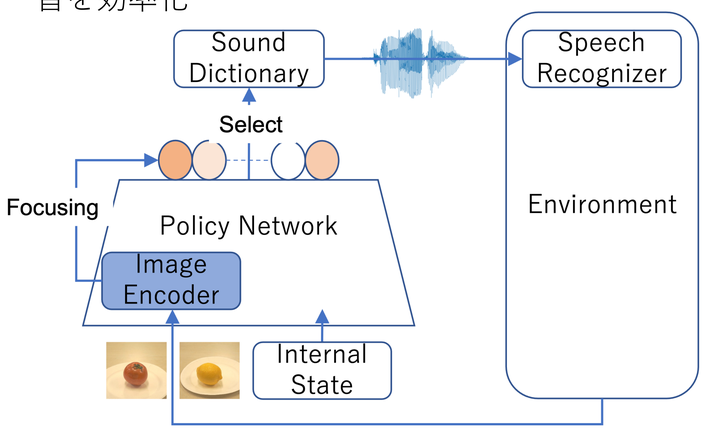
- 人間と共生するロボットには自律的な音声言語学習能力が必要不可欠
- 先行研究では教師ラベルを用いずに音声単語を学習するエージェントが提案された
- しかし，評価には合成音声を用いたため，話速や声の高さが異なる複数の話者との対話に対する一般性は不明
- 本研究では，収録した実音声を用いて学習アルゴリズムの有効性を検証する

2. エージェントの構造

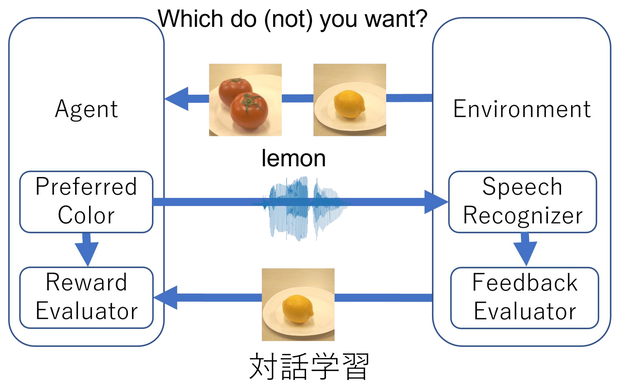
- 観察学習
  - 連続音声を教師なし単語分割して音声辞書を作成
  - 音声と画像を接地するために，音声と画像のペアを同一潜在空間に埋め込む



- 対話学習
  - 深層Qネットワークによって音声辞書からセグメントを選択して発話
  - 視覚に対応する単語に注意を向けることによって，学習を効率化



3. タスク設定



対話学習

- 単語発見では，連結された音声説明を聞く
- 音声説明は”<food>”，”A <food>”，”A <color> <food>”，”It’s a <food>”の4種類
- 対話学習では，エージェントは対話相手から2つの食べ物を提示され，Which do (not) you want?と尋ねられる
- 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の名前を発話すると報酬1，それ以外の場合報酬0を得る

4. データ収録

- 実験協力者 (31名) はクラウドソーシングによって募集
- Webブラウザで画像とともに表示される文章を読み上げるよう依頼
- 各話者は2種類の音声質問と20種類の食べ物それぞれにつき4種類の音声説明の計82発話を5セット収録

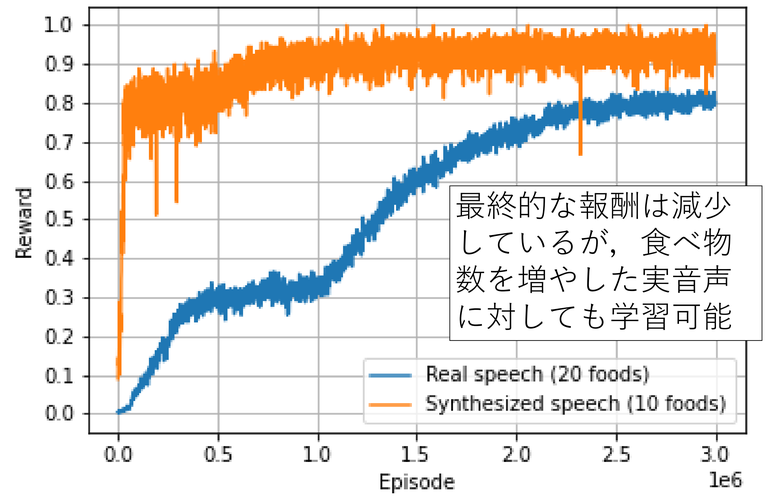
5. 実験

- 実験1：主観評価
  - 食べ物：16種類
  - 音声質問：実音声 (被験者と重複なし)
  - 音声説明：合成音声

Episode	MOS		Reward	
	Trial 1	Trial 2	Trial 1	Trial 2
50,000	1.67	1.67	0.233	0.233
1,000,000	4.33	4.33	0.800	0.800

訓練データにない話者による評価でMOSは4.33，正解率は0.8であり，エージェントの一般性を確認

- 実験2：実音声と合成音声での比較 (※実験条件は一部異なる)



対話学習における平均報酬

6. まとめ

- 実音声を用いて音声言語獲得エージェントの有効性を確認
- 訓練データにない話者との音声対話を行い，エージェントの頑健性を確認
- 今後の課題としては，さらに多くの教師なし学習による言語学習能力の拡張が挙げられる

謝辞

本研究はJSPS科研費 JP22K12069の助成を受けたものです