Continuous Action Space-Based Spoken Language Acquisition Agent Using Residual Sentence Embedding and Transformer Decoder

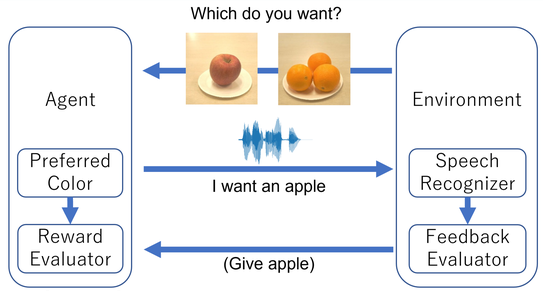
Paper: 6307

Ryota Komatsu[1] Yusuke Kimura[1] Takuma Okamoto[2] Takahiro Shinozaki[1]
[1]Tokyo Institute of Technology [2]National Institute of Information and Communications Technology

Introduction

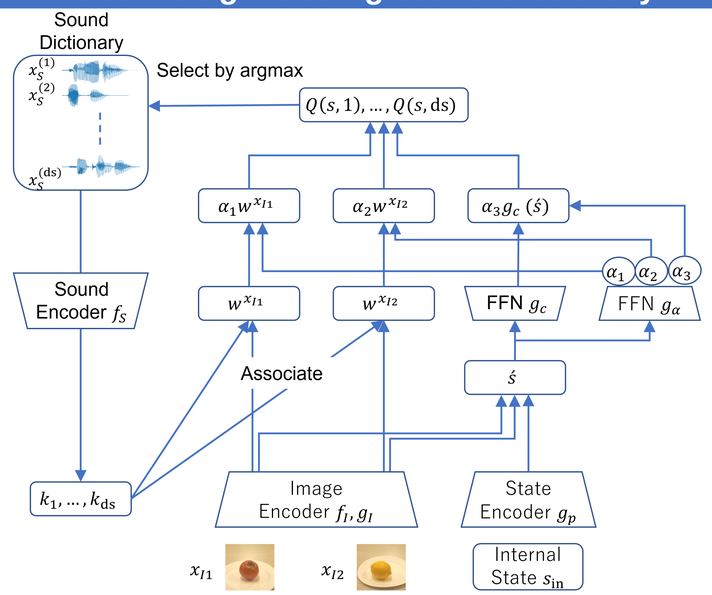
- This study aims to realize the mechanism of human language learning on computers
- Skinner explained the mechanism by behaviorist reinforcement learning principles, while Chomsky considered children learn verbal behavior by observation of adults and other children
- Currently, the true answer is an open question needing a mathematical model

Spoken language acquisition task [1]



- Observation phase
  Audio explanations are given while foods are shown
- Dialogue phase
  An agent has a favorite color as an internal state and is rewarded for answering “I want a <preferred food>”

Baseline agent using sound dictionary

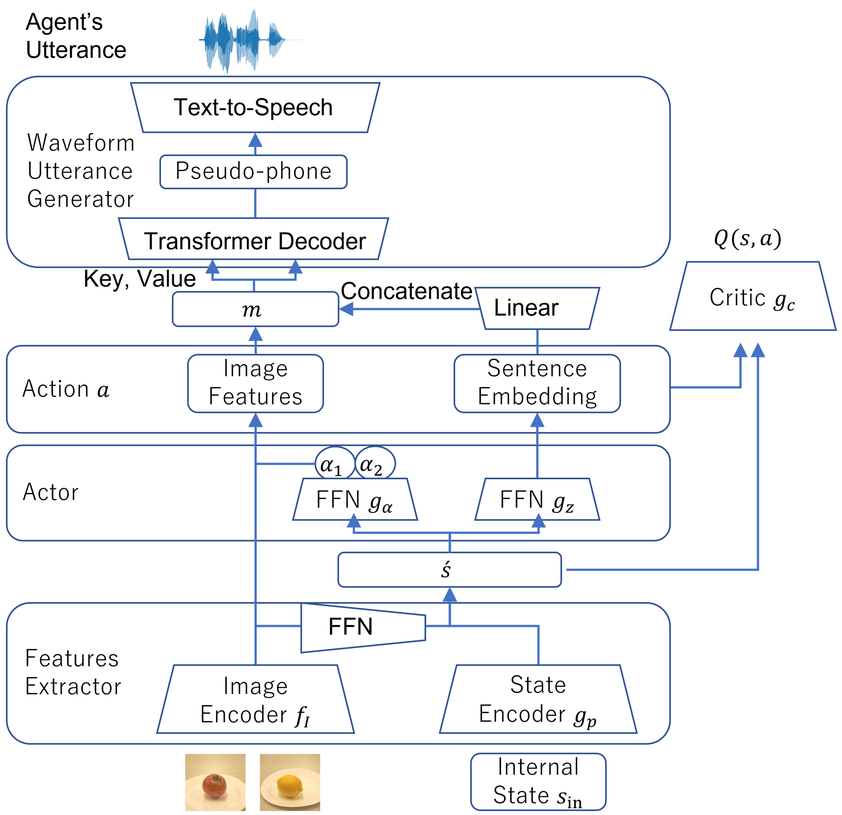


- Learn visually grounded words from scratch without relying on any labels
- Generate an utterance by selecting a word with a help of vision based focusing mechanism
- Can only pronounce single-word utterances

Proposed agent

- Propose an agent that can speak multi-word utterances
- Replace the sound dictionary of the baseline agent with a language model learned from raw speech
- Replace the focusing mechanism with a language model conditioning based on image features and residual sentence embedding

The image features provides a context to narrow the choices of utterances, and the residual sentence embedding identifies the utterance to speak



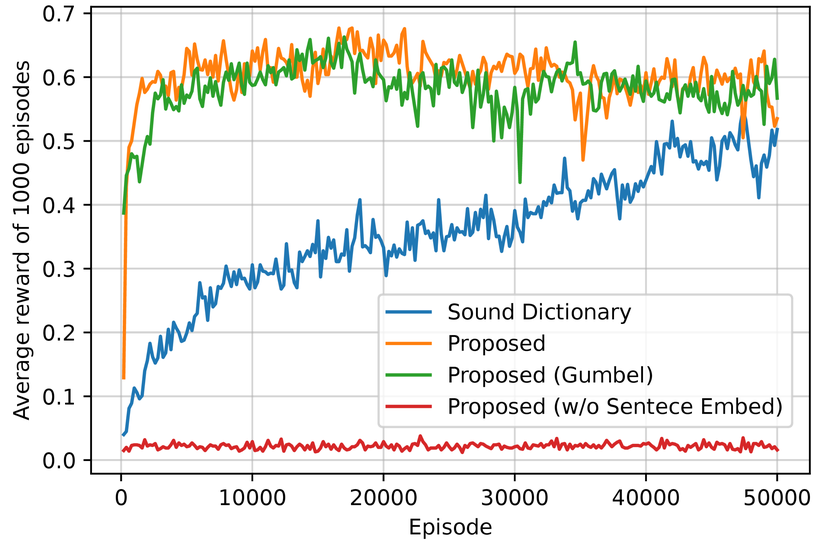
Training method

- Automatically transcribe speech utterances into pseudo-phone unit sequences
- Pre-train a language model and speech synthesizer based on the pseudo-phone units
- Perform reinforcement learning in a *continuous* action space of sentence embedding and image features

Results

Task setup

- Num. of food types: 20
- Description types: “<food>,” “A/An <food>,” “A/An <color> <food>,” “I want a/an <food>”



Average rewards on dev set in the dialogue phase

Average rewards on dev/test sets in the dialogue phase

	Dev (400th ep.)	Dev (2000th ep.)	Dev (best)	Test (best)
Sound Dict.	0.045	0.156	0.549	0.460
Proposed	0.429	0.581	0.677	0.646
Proposed (Gumbel)	0.446	0.507	0.663	0.604
Proposed (w/o Sentence emb.)	0.020	0.022	0.038	0.028

Acknowledgement

This work was supported by JSPS KAKENHI under Grant JP22K12069.

Reference and toolkit

[1] Komatsu+, *IEEE JSTSP*, 2022.

Toolkit
https://github.com/tttslab/spolacq