Pronunciation Adaptive Self Speaking Agent Using WaveGrad

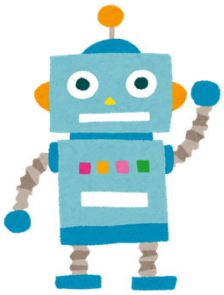
AAAI SAS 2022

Tomohiro Tanaka (Tokyo Institute of Technology, Japan)
Ryota Komatsu (Tokyo Institute of Technology, Japan)
Takuma Okamoto (National Institute of Information and Communications Technology, Japan)
Takahiro Shinozaki (Tokyo Institute of Technology, Japan)

Automatic Spoken Language Acquisition

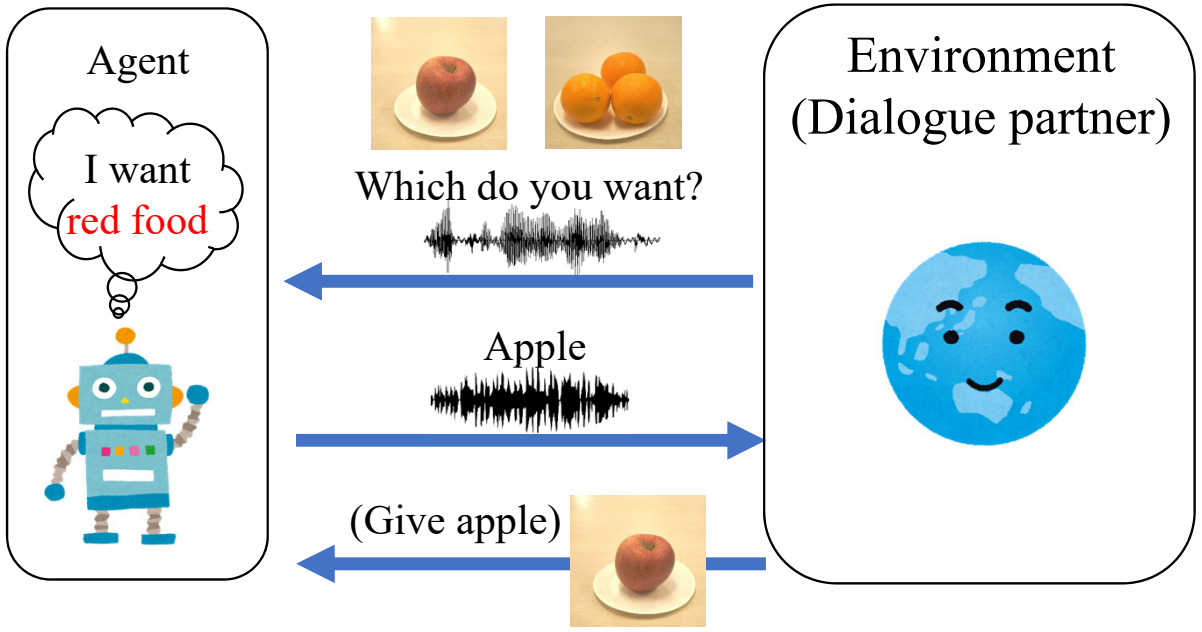
- Human babies learn language making a closed learning loop in human society
- They learn new words as well as their pronunciation without relying on labeled data

No initial knowledge of any language



Agent
(Human or AI)

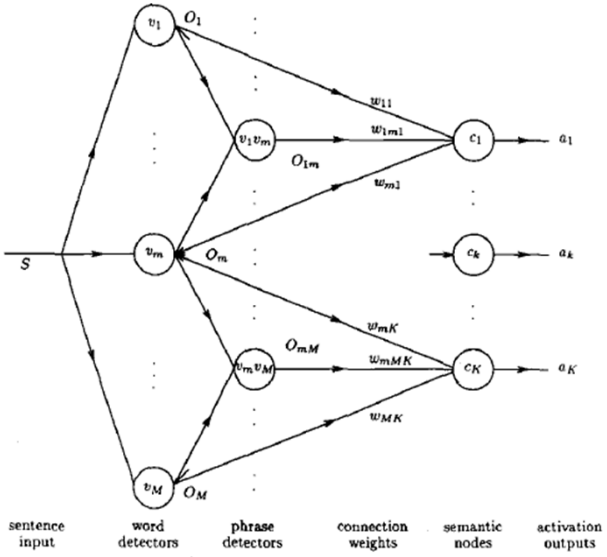
Learns spoken utterances as a means to interact with the environment



Related Works

[Gorin+ IEEE Trans. Speech and Audio Processing 1994]

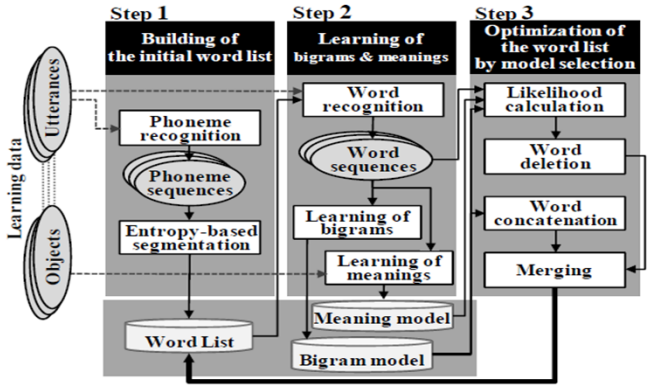
① Word discovery
② Semantic grounding
③ Action learning[1]
④ Pronunciation learning



[1] Action space is pre-defined

[Taguchi+ Interspeech 2011]

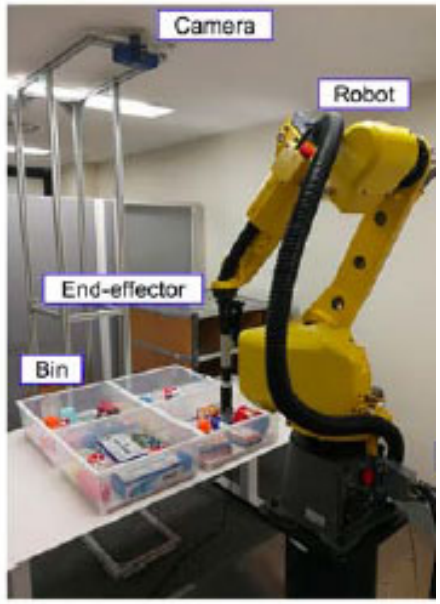
① Word discovery[2]
② Semantic grounding
③ Action learning
④ Pronunciation learning



[2] Phoneme recognizer is pre-trained

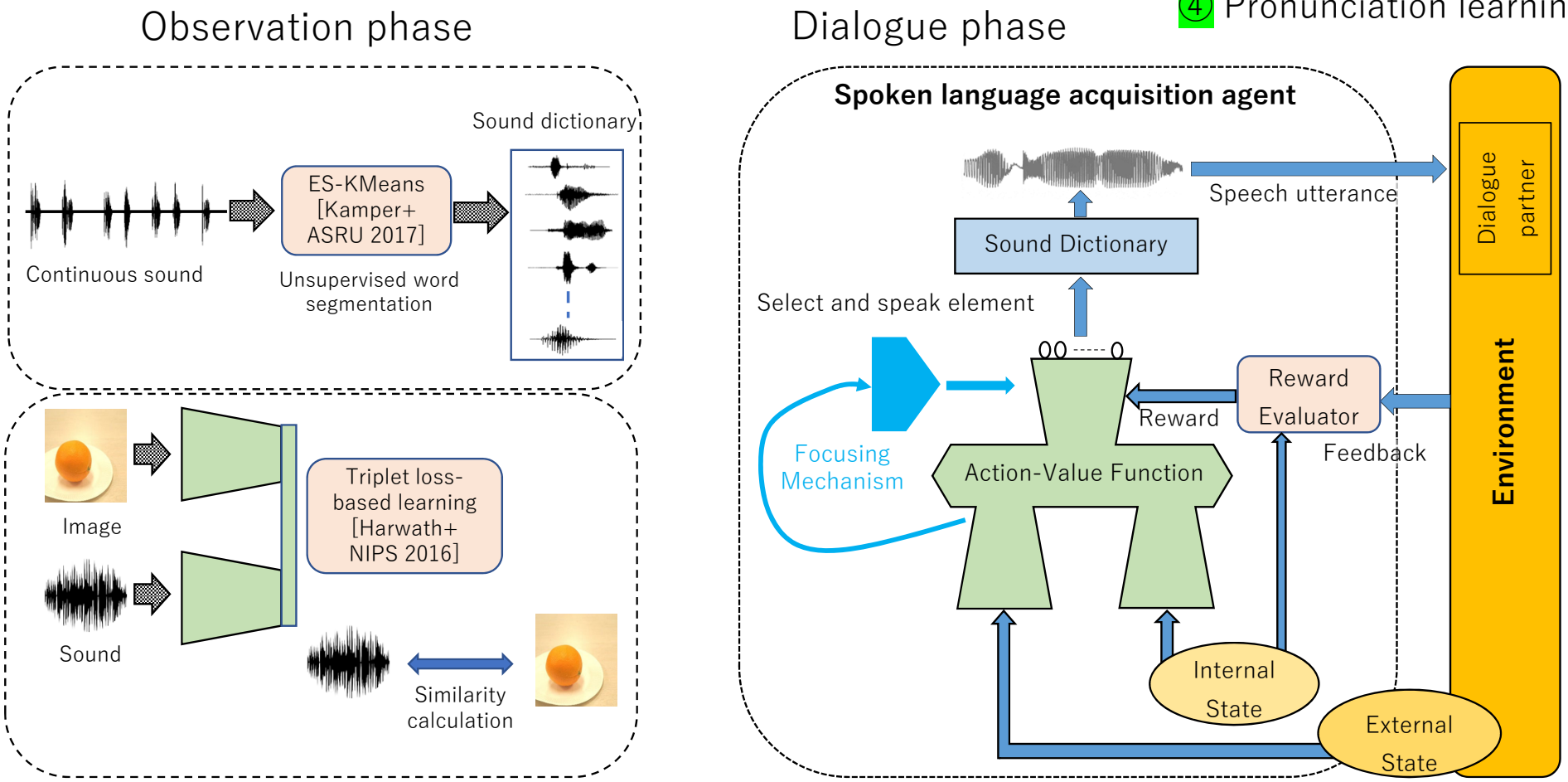
[Hatori+ IEEE ICRA 2018]

① Word discovery
② Grounding
③ Action learning
④ Pronunciation learning



Zhang's Agent [Interspeech 2020]

① Word discovery
② Semantic grounding
③ Action learning
④ Pronunciation learning



Pros and Cons of Using Sound Dictionary

- Pros:
  - Discretizes the action space of utterance pronunciation, and makes the reinforcement learning efficient
- Cons:
  - Can not adapt pronunciation other than changing segment selection

Our idea: Replace the sound dictionary with a generative neural vocoder

	Sound Dictionary	Neural Vocoder
Discrete/Compact Action Space	✓	✓
Adaptability		✓

Neural Vocoder

- Unconditional Training:
  - Generate waveform by random sampling without conditioning
  - Sounds like human baby's babbling

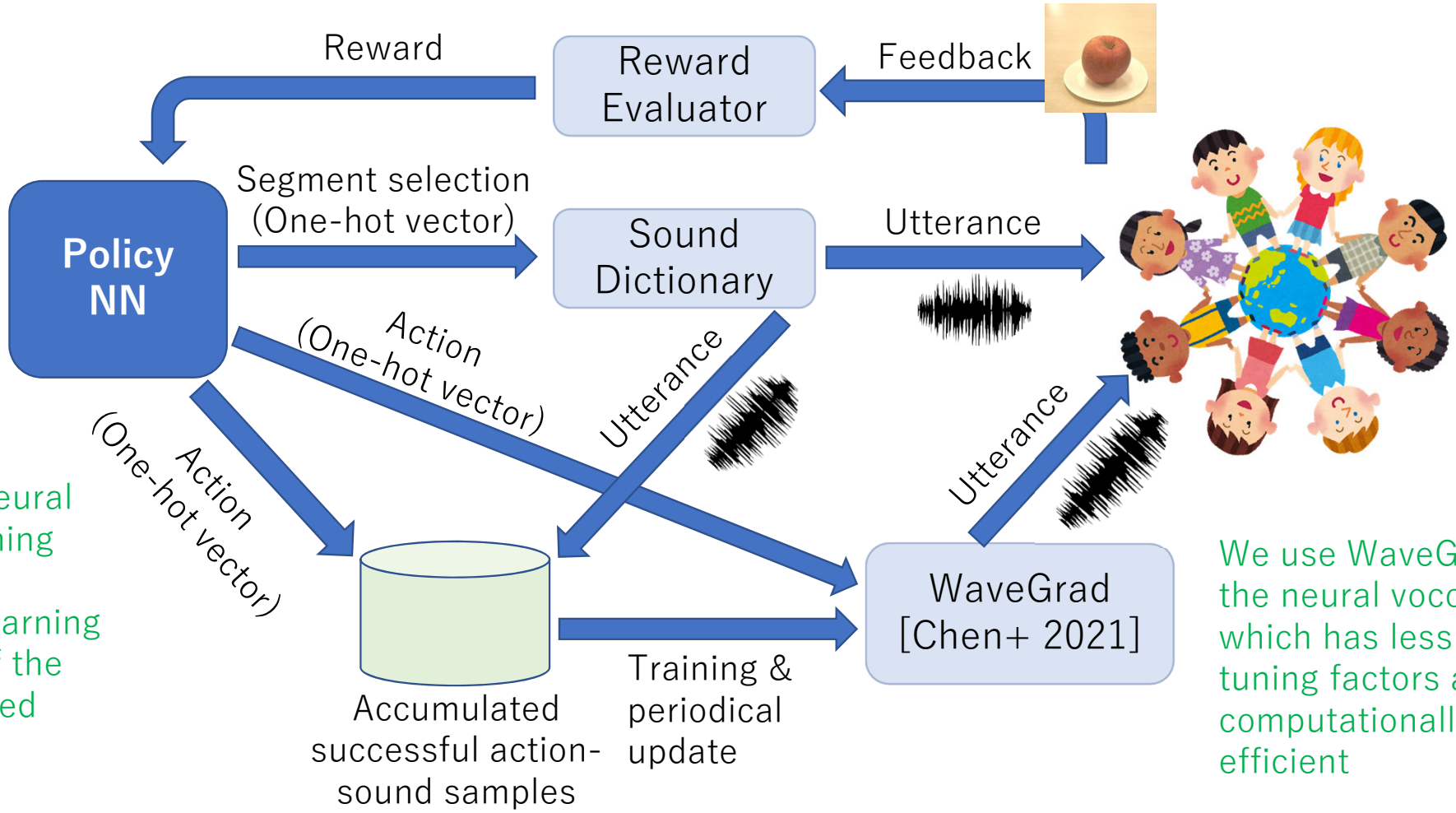
WaveNet [Oord+ 2016]
with unconditional training


original


generated

- Semi-conditional Training:
  - Conditioned on time invariant information
  - It can be word ID as in DiffWave [K. Kong+ ICLR 2021]

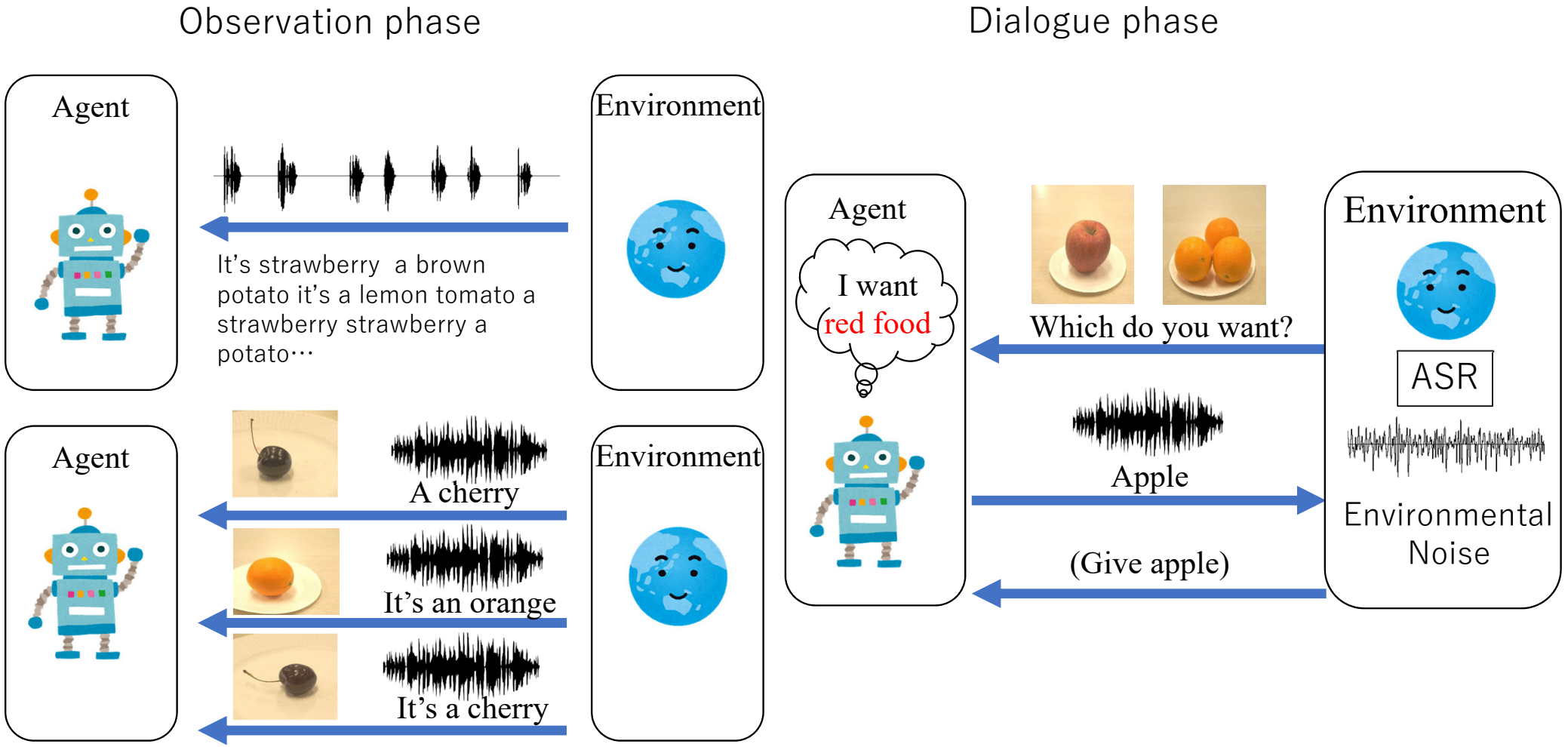
Proposed Self-Supervised Learning Method



Embed the neural vocoder learning into the self-supervised learning framework of the dialogue-based learning

We use WaveGrad as the neural vocoder which has less tuning factors and is computationally efficient

Task Design

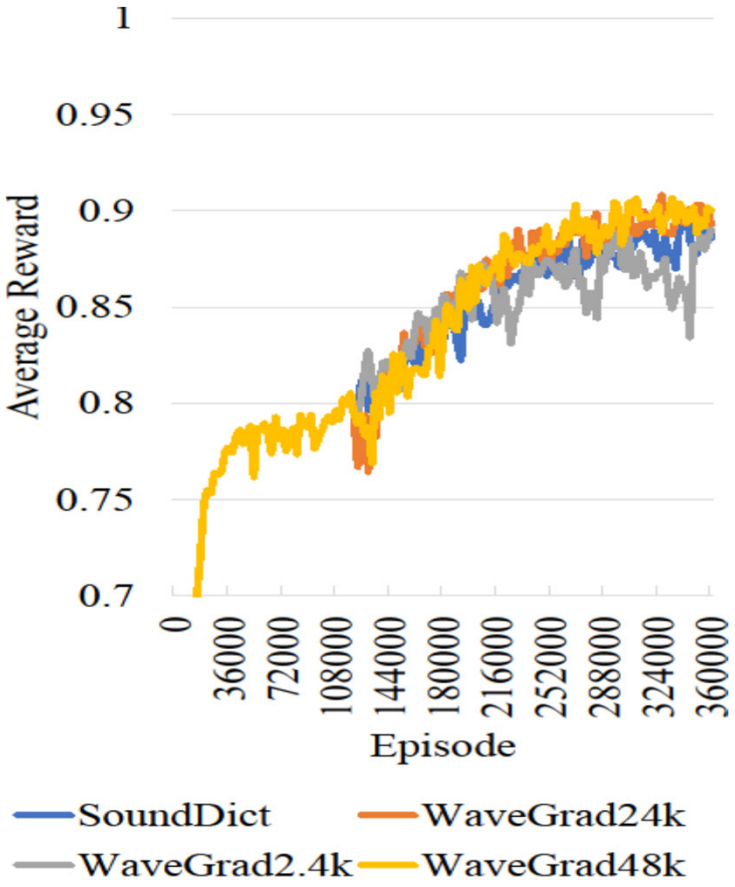

Observation phase
Dialogue phase
Agent
Environment
It's strawberry a brown potato it's a lemon tomato a strawberry strawberry a potato…
Agent
A cherry
It's an orange
It's a cherry
Environment
Agent
I want red food
Which do you want?
Apple
(Give apple)
Environment
ASR
Environmental Noise

Experimental Setup

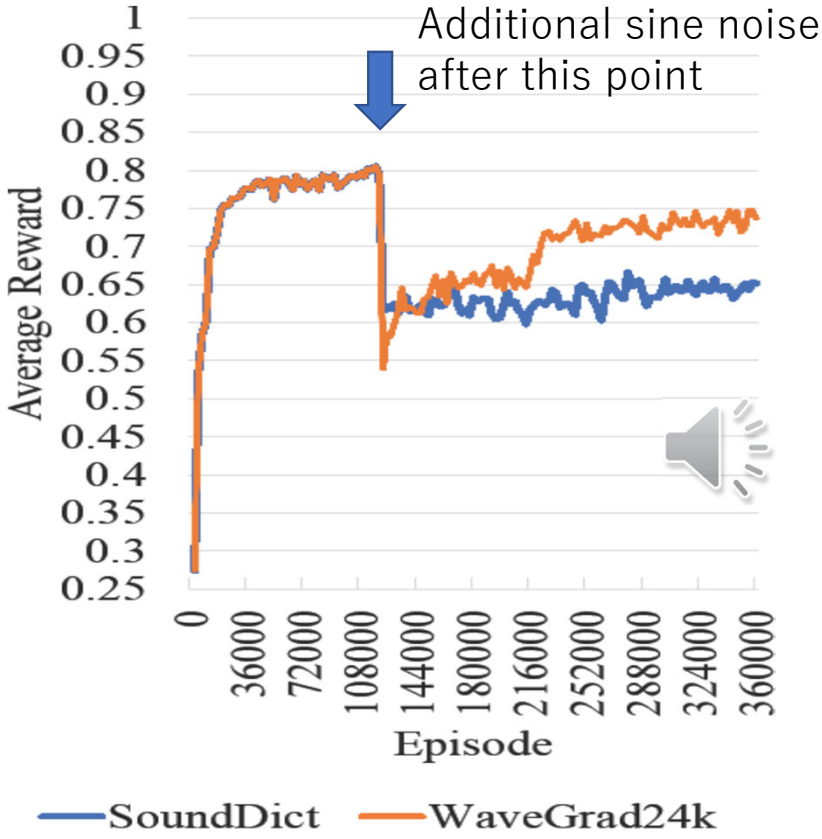
- Baseline:
  - Zhang's system
- Image dataset:
  - Food type: 8
  - Number of training images: 90 per each food
- Observation phase
  - Audio description: Generated using Google Text-To-Speech using templates
    - "<food>"
    - "A<food>"
    - "A<color><food>"
    - "It's a<food>"
- Dialogue phase
  - Sound dictionary size: 12,000
  - Noise condition: White noise + additional sine noise after 120,000 episodes

Results

Steady environment with white noise



Changing environment with additional sine noise after 120,000 episodes



Summary and Future Work

- **Summary**:
  - We proposed a pronunciation adaptive spoken language acquisition agent using WaveGrad
  - The WaveGrad learning is embedded in the self-supervised learning approach
  - The agent can automatically adapt its pronunciation to a changing environment
- **Future work:**
  - Improve the sample efficiency of the pronunciation adaptation by introducing model adaptation techniques
  - Support multi-word utterances by extending observation and dialogue learnings

Spolacq toolkit: https://github.com/tttslab/spolacq