Spoken Language Acquisition Based on Reinforcement Learning and Word Unit Segmentation

Shengzhou Gao, Wenxin Hou, Tomohiro Tanaka, Takahiro Shinozaki

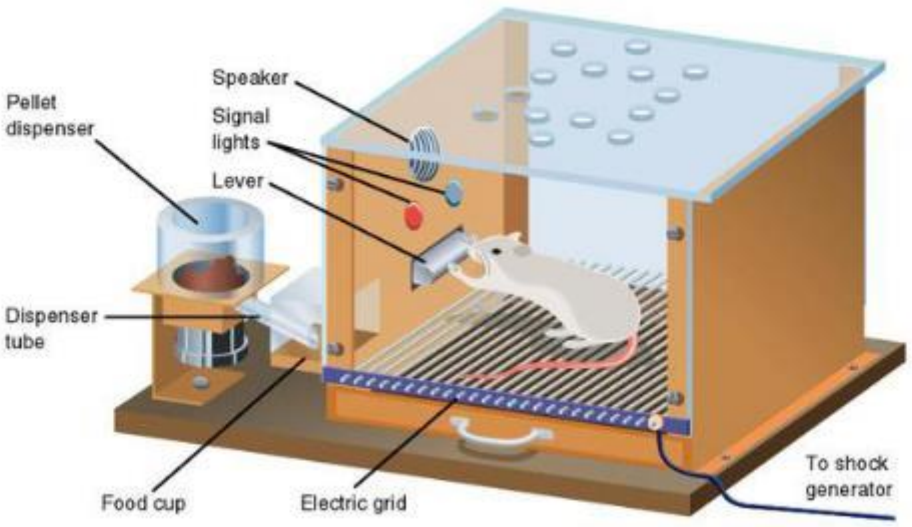
Tokyo Institute of Technology

Presenter：Wenxin Hou

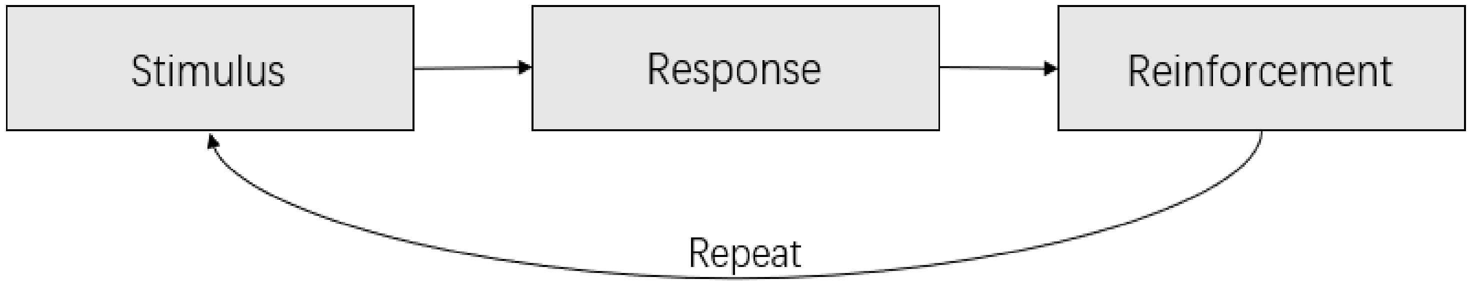
2020.04.17

- Language acquisition (LA) is the process of how human babies acquire languages which has been researched for decades. However, its mechanism remains as a mystery

- Skinner[1] gave a widely-accepted explanation: children learn the language based on behaviorist reinforcement principles by associating words with meanings



Skinner box [2]

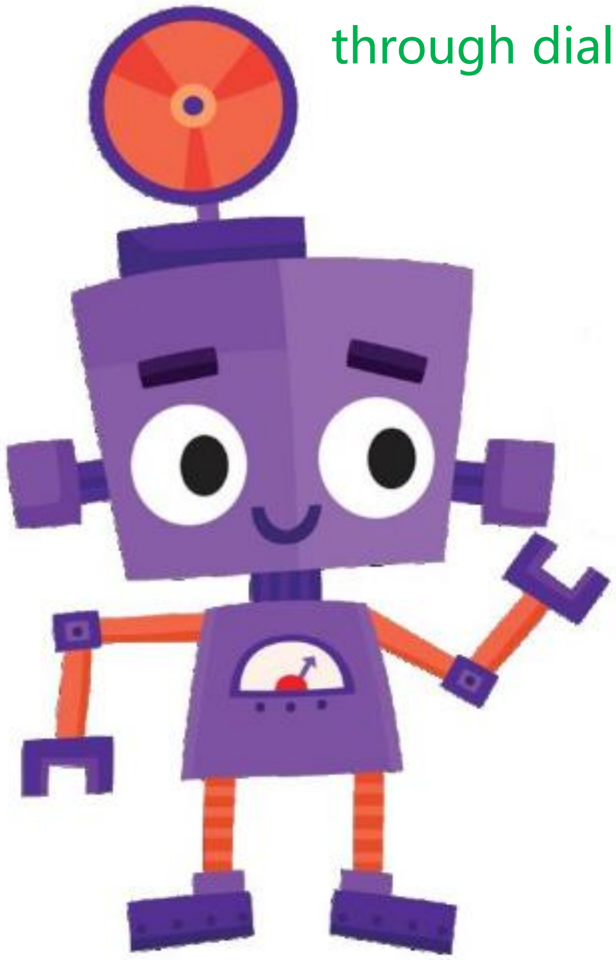


Skinner's language acquisition theory

[1] Burrhus Frederic Skinner. Verbal behavior. New York: Appleton-Century-Crofts, 1957
[2] McLeod, S. A. (2007). Skinner - Operant Conditioning. Retrieved from http://www.simplypsychology.org/operant-conditioning.html

- Assume a “baby” robot with no prior language knowledge

want to learn spoken language
through dialogue



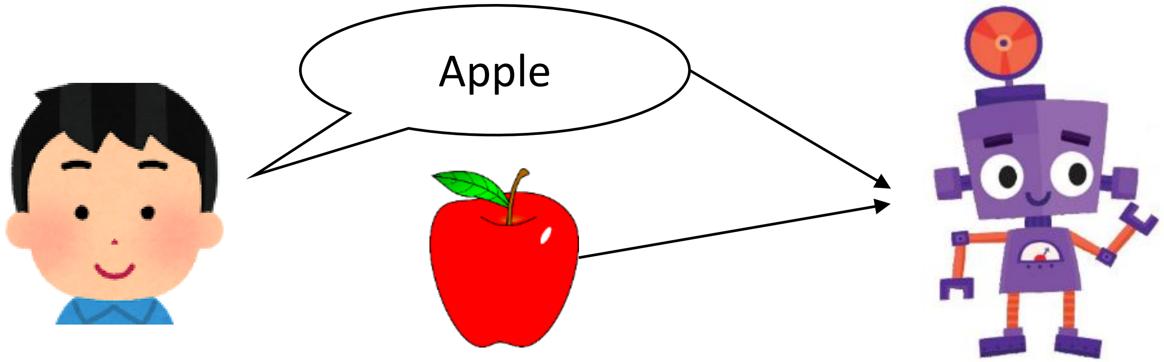
I have to acquire language directly from continuous speech by myself, but how?

How about following Skinner’s theory?

How can I evaluate my performance in language acquisition?

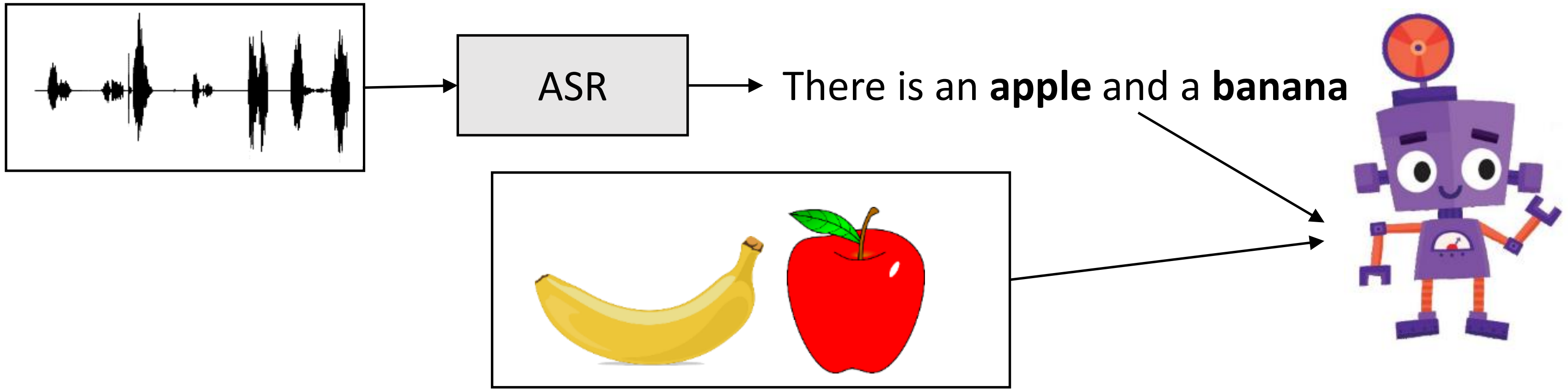
01 **Introduction** | Related works

- Chauhan et al. introduced a spoken language acquisition model by human-robot interaction, where humans teach and correct the robot to name visual objects word by word [3]



Sounds need to be manually segmented beforehand

- Yu et al. presented a co-occurrence-based language grounding model for object categorization, which firstly convert spoken language to text using automatic speech recognition (ASR) system [4]



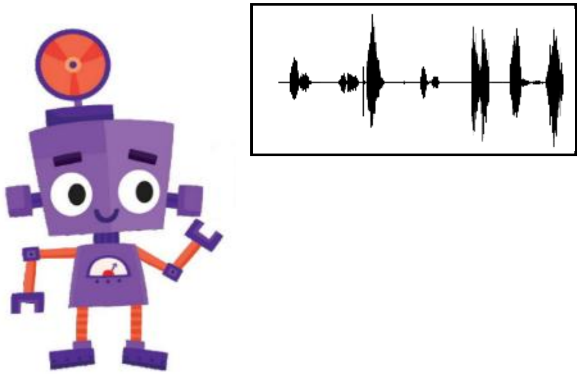
Needs pre-trained ASR inside the robot

[3] Chauhan, Aneesh & Seabra Lopes, Luís. (2011). Using spoken words to guide open-ended category formation. Cognitive processing. 12. 341-54. 10.1007/s10339-011-0407-y
[4] Chen Yu and Dana H Ballard, "On the integration of grounding language and learning objects," in AAAI, 2004, vol. 4, pp. 488–493

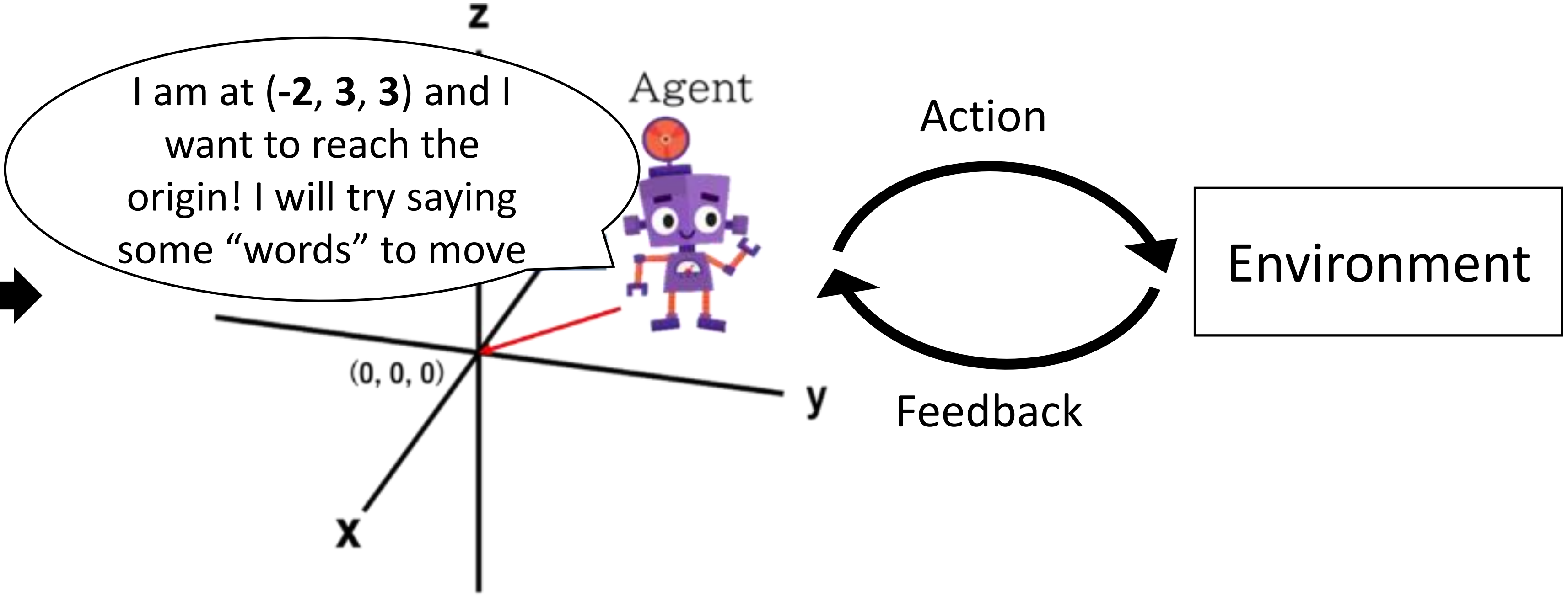
02 **Methodology** | Task design

1. Observation phase
The robot is given unsegmented spoken sound examples

2. Dialogue phase
The robot is placed in an environment where it gets reward according to its utterance

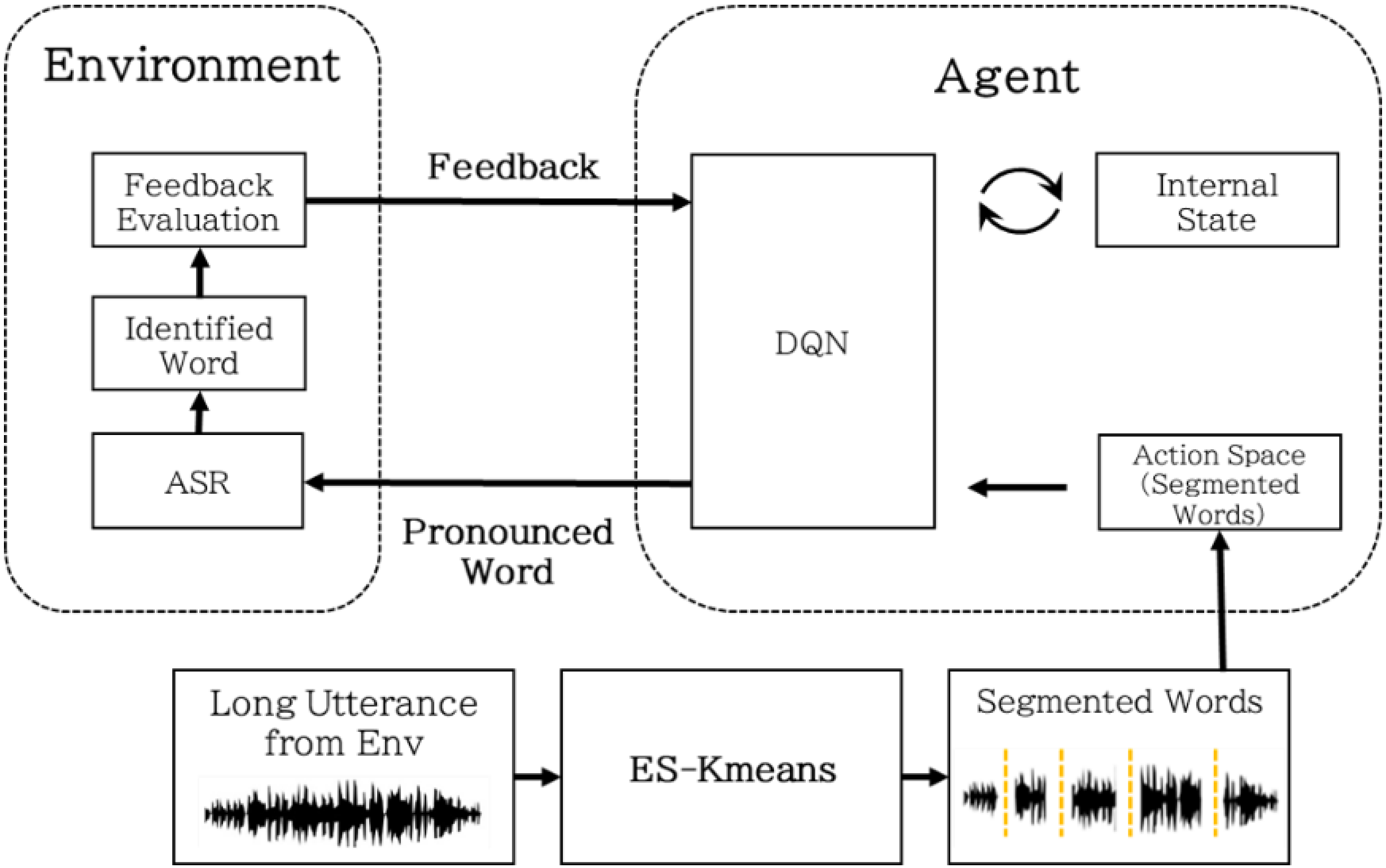


Unsegmented sound sample



The robot is randomly initialized at a 3-D point (x, y, z) and wants to reach the origin by pronouncing correct words

Methodology | System architecture



02 **Methodology** | ES-Kmeans algorithm

- **Input**: utterance sequence $y_{\{1:M\}} = y_1, y_2, \dots, y_M$
- **Goal**: cut this sequence into different sub-segments of meaningful words
- **Algorithm**:
  1. Cut the whole long utterances randomly into segments q
  2. Map arbitrary-length segments (e.g., $y_{\{t_1:t_2\}}$) into embeddings x_i, $x \in R^D$, where D is the embedding dimension
  3. Cluster the embeddings by a K-means algorithm
  4. Keep the cluster assignments z and optimize segmentation q
  5. Keep the segmentation q and optimize cluster assignments z
  6. Repeat 4 and 5 until convergence of the target function:

$$\min_{z} \sum_{c=1}^{K} \sum_{x \in X_c} ||x - \mu_c||^2$$

where $\{\mu_{c=1}^{K}\}$ are cluster centers, X_c are vectors assigned to cluster c, element z_i in z indicates to which cluster x_i belongs

[Kamper et al., ASRU 2017]

Methodology | Deep Q-Network (DQN)

- **Action space**: speech segments output by segmentation algorithm
- **Policy network $\boldsymbol{Q}$**: select action from the action space
- **Target network $\widehat{\boldsymbol{Q}}$**: generate learning target, copying parameters from Q every 10 episodes
- **Reward function $\boldsymbol{r}$**: change in satisfaction level (minus Euclidean distance to the origin) between steps

$$SL(t) = -(x_t^2 + y_t^2 + z_t^2)$$

$$r_t = SL(t) - SL(t-1)$$

- **Loss function**:

$$y = r_t + \gamma \max_{a_{t+1}} \hat{Q}(S_{t+1}, a_{t+1}; \theta^-)$$

$$L(\theta) = \left(y - Q(S_t, a_t; \theta)\right)^2$$

where r_t is the reward of the current action a_t, γ is the discounting factor

[Mnih et al., Nature 2015]

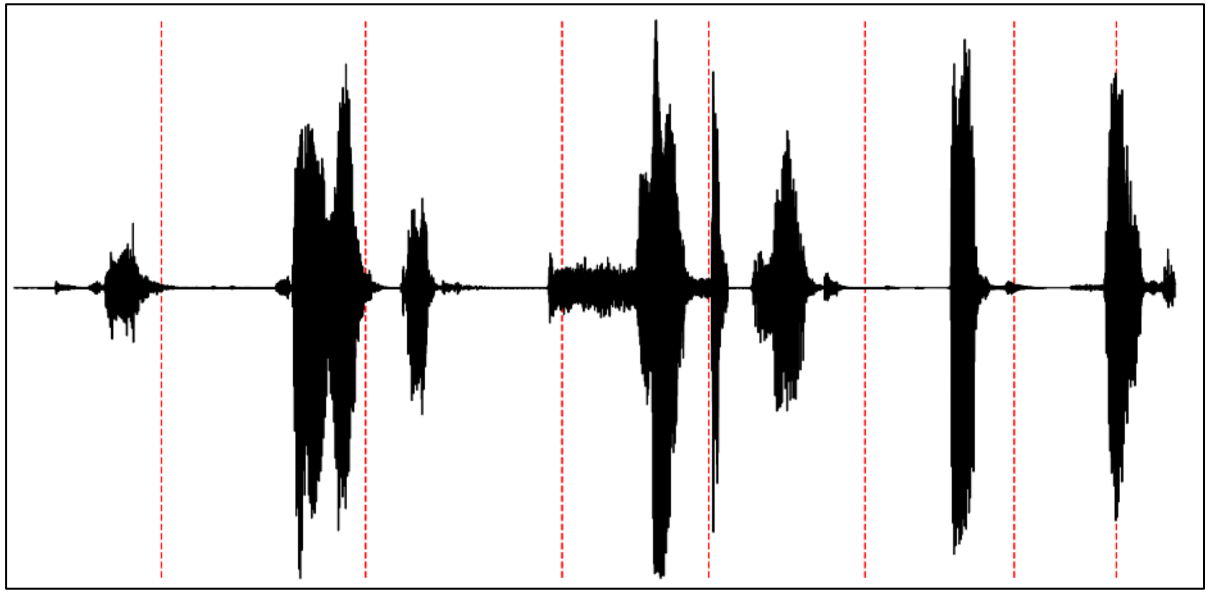
- **Dataset**: Google Speech Command

Total number of one-second utterances	65,000
Vocabulary size	35
Number of command word types	6 (up, down, left, right, forward, backward)

- **Preprocessing**: 200 samples are randomly picked from 6 types of command words plus a noise word "Marvin" and concatenated into a continuous speech

- **Environment module**: Google Speech-to-Text API [1], a **general-purpose** ASR system

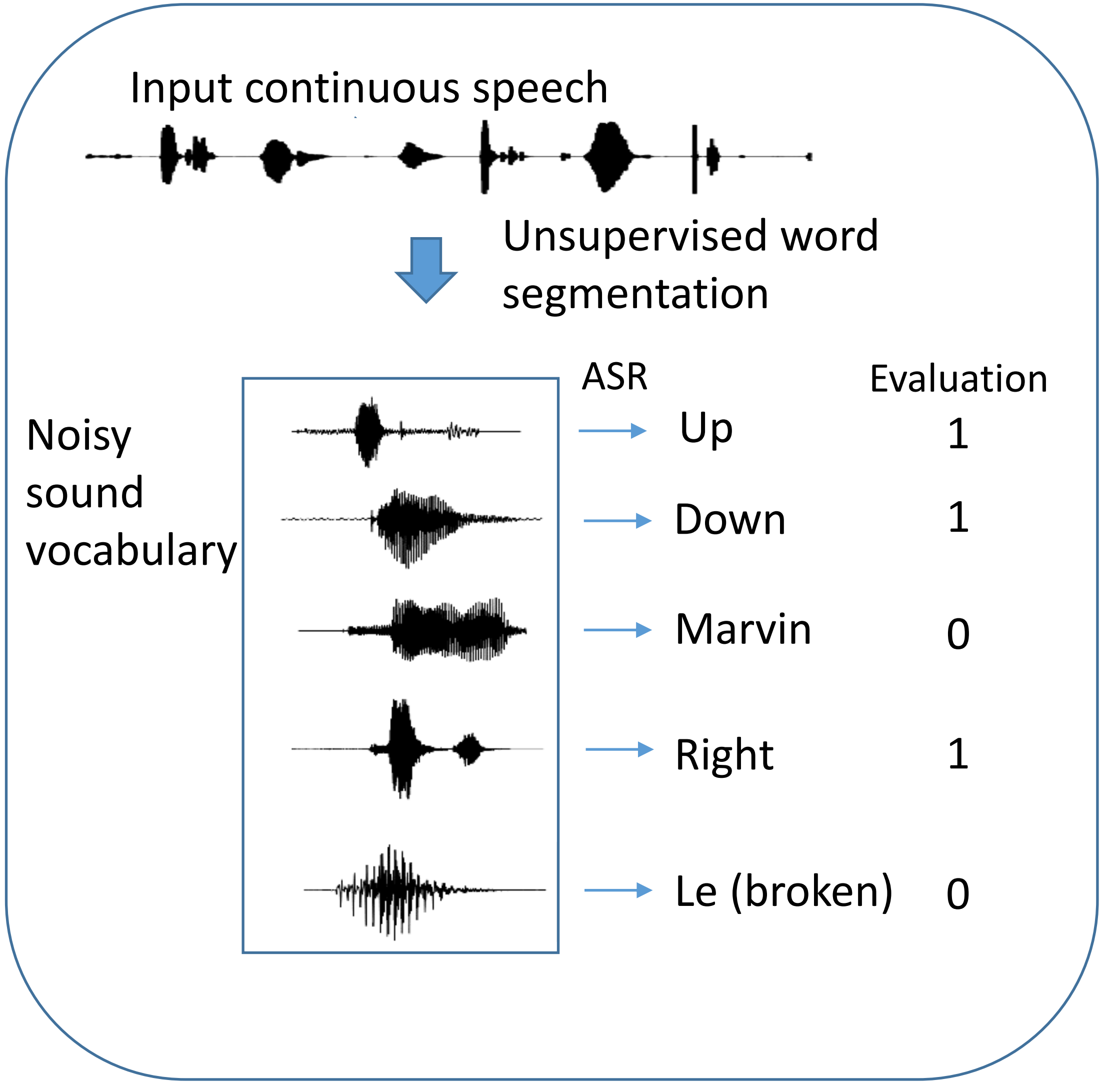
- **Baseline**: random-cut word unit segmentation method (Random Method), which cuts the utterance randomly with an average duration of approximately one word (e.g. 500-1,200ms)

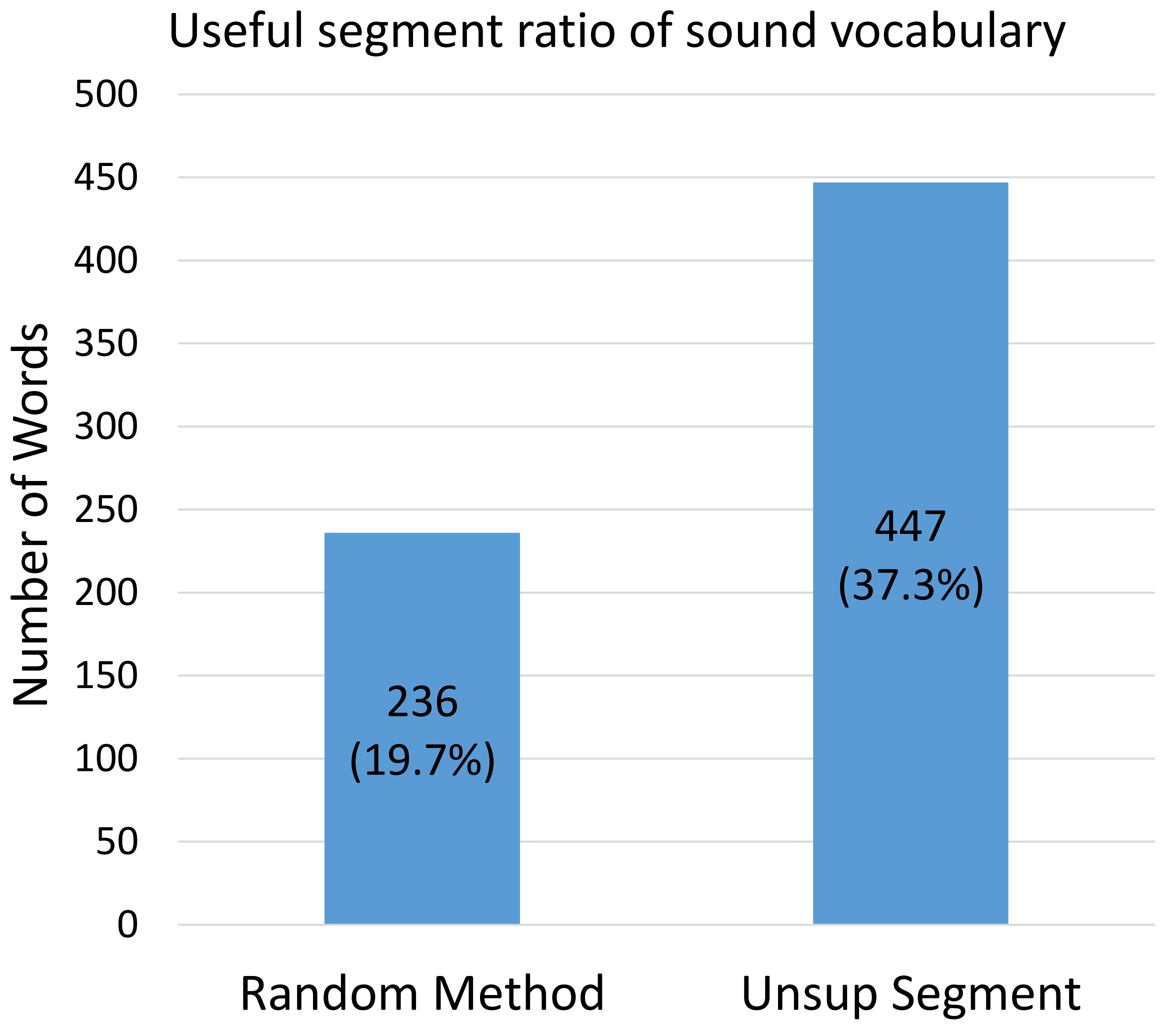


Example of random method

[1] https://cloud.google.com/speech-to-text/docs/reference/rest/

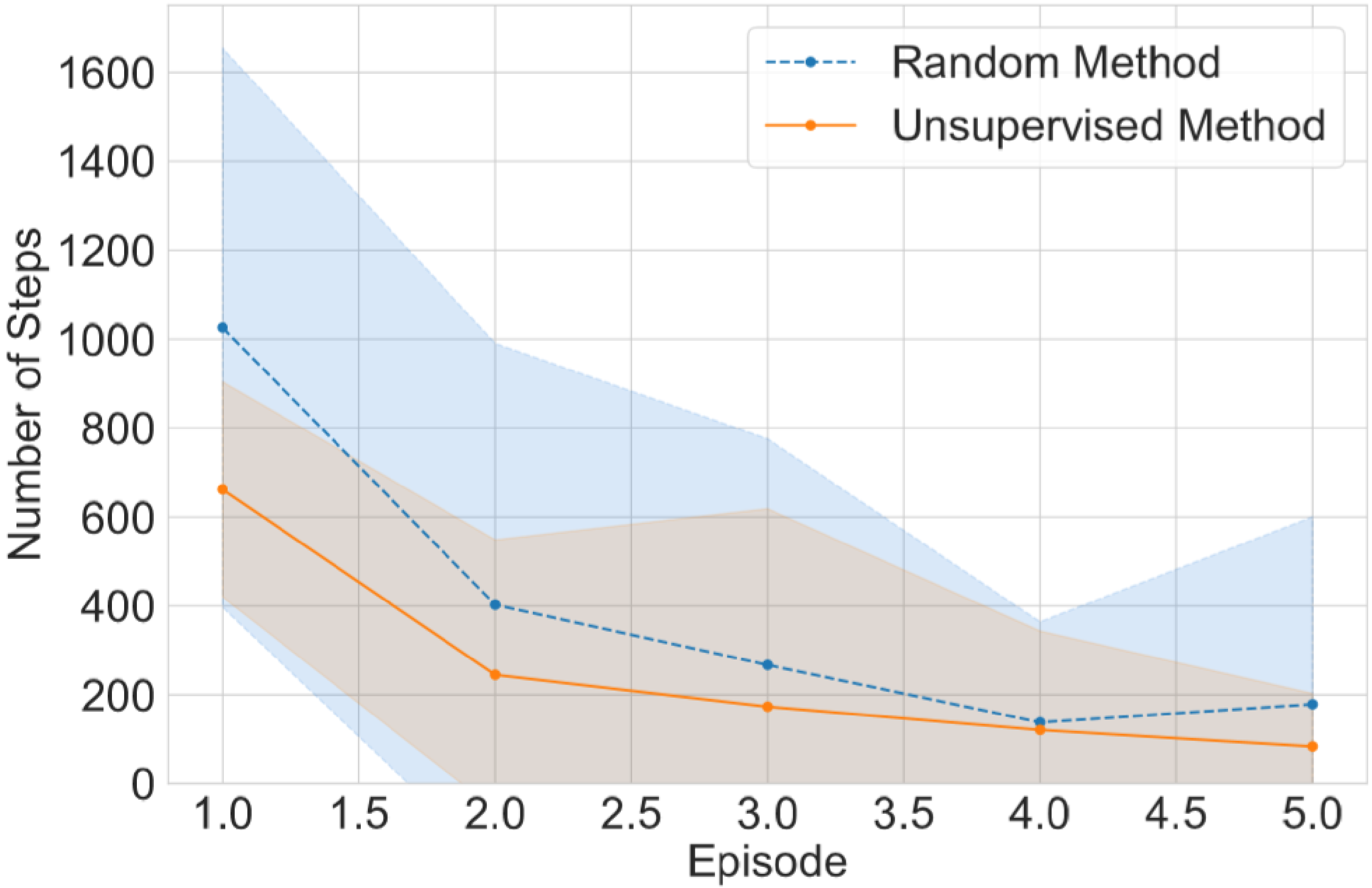
03 **Evaluation** | Word segmentation results





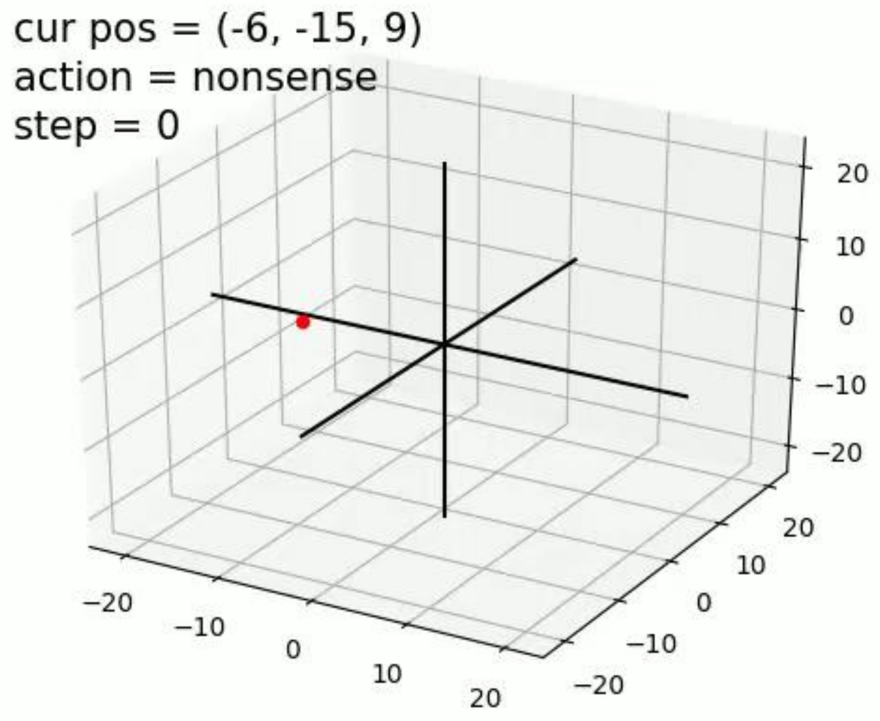
- Evaluation metric: proportion of segmented words recognized by the environment among 1200 command words
- Unsupervised word segmentation method achieves 18% higher ratio than the random-cut method

03 **Evaluation** | Task completion results

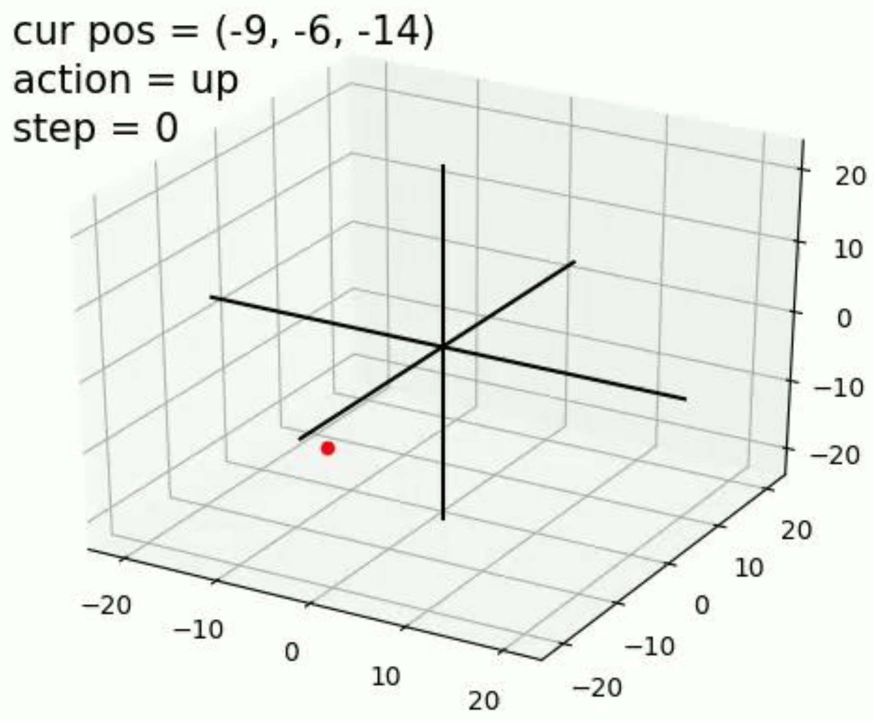


- Evaluation metric: number of steps taken by the robot to return to the origin for each episode
- An episode ends when the robot reaches the origin
- Unsupervised method does excel by a 35.45% reduction in the average number of steps taken for the first episode

- First five actions at episode 0
- First five actions at episode 5



Moving process at episode 0



Moving process at episode 5

Conclusions

1. We simulate the language acquisition process following Skinner's theory

2. Our experiment shows the effective increasing in the learning efficiency by utilizing auxiliary unsupervised segmentation method

3. Our future works will include extending the task to more complicated application, e.g. learns to organize language for image description, training of speech synthesizer for more flexible utterance pronunciation

Q & A

Thank you for your listening !

Presenter: Wenxin Hou