Pronunciation Adaptive Self Speaking Agent Using WaveGrad

Tomohiro Tanaka,[1] **Ryota Komatsu,**[1] **Takuma Okamoto,** [2] **Takahiro Shinozaki** [1]

[1] Tokyo Institute of Technology, Japan
[2] National Institute of Information and Communications Technology, Japan
https://www.ts.ip.titech.ac.jp, https://www2.nict.go.jp

Abstract

The ability to automatically learn to speak through observation and dialogue without relying on labeled training data is essential for intelligent robots or agents to flexibly and expressively talk to humans on an equal footing. Previous methods have demonstrated that automatic spoken language acquisition becomes possible by combining unsupervised and reinforcement learnings with end-to-end neural networks. However, such utterances were a simple playback of segmented wave sounds, which lacked flexibility in pronunciation. This work introduces WaveGrad speech synthesizer as the agent's speech organ by embedding its optimization in the self-supervised learning framework. Experimental results show that WaveGrad gives the same speaking performance as the conventional method in a steady environment and outperforms it when the background noise changes, proving its ability to adjust its pronunciation for smoother communication.

Introduction

The performance of current speech recognition systems has reached the human level in some tasks (Xiong et al. 2017), and the quality of synthesized utterance has reached a level where we cannot easily distinguish an authentic utterance from a synthesized one (van den Oord et al. 2016; Shen et al. 2018). However, a machine's conversation ability is still significantly inferior to that of humans, and the machine lacks the power to handle open situations by extending its language knowledge on the fly (Taniguchi et al. 2016). Such an ability is fundamental for humanoid robots that coexist with humans to have flexible communication in individual circumstances, plant operating agents that collaborate with human operators to resolve unexpected problems, and others.

To fill the gap, speaking agents need a self-supervised learning mechanism to learn new words from speech conversation by creating a closed learning loop in human society. Learning a new word involves 1) identifying the speech segments corresponding to the word, 2) associating it with its meaning, 3) modeling the effect of using it, and 4) synthesizing its pronunciation. Realizing and improving such learning algorithms will also greatly contribute to revealing the mechanism enabling human beings to acquire spoken language from scratch, which is one of the fundamental issues in cognitive science as well as in artificial intelligence (Dupoux 2018; Kuhl 2004). Skinner was the first researcher who considered the question from an engineering perspective in the 1950s (Skinner 1957). He explained that children acquire language based on behaviorist reinforcement principles by associating words with meanings. However, it remained at the conceptual level at that time.

Recently, Gao *et al.* proposed an end-to-end deep neural network-based agent, which could learn to pronounce speech commands to move around in a simulated 3D space (Gao et al. 2020) [1]. For the speaking agent, executing an action was to pronounce an utterance. Theoretically thinking, a self-learning speaking agent is made possible by applying reinforcement learning to a neural network that has a speech synthesizer as the output and pattern recognizers as the inputs. However, the difficulty is how to handle the high dimensional continuous action space of the utterance waveform obtaining convergence in realistic time. Assuming 8kHz sampling frequency, the dimension of the action space of 1.0 second short utterances is already 8,000. The approach in (Gao et al. 2020) was to perform unsupervised word learning to produce a sound dictionary and use it as a discrete action space for the speaking agent. Zhang *et al.* extended the system by introducing a vision-based focusing mechanism based on unsupervised cross-modal representation learning, improving the learning efficiency (Zhang et al. 2020). To the best of our knowledge, these were the first systems jointly supporting the four aspects of the self-supervised new word learning of the segmentation, meaning association, effect modeling, and the pronunciation without relying on any pre-trained supervised models.

While Gao's and Zhang's systems work in an end-to-end manner, one limitation was that the agent's utterance was only based on selecting and replaying an element in the sound dictionary; thus, it lacked the mechanism to modify and improve the pronunciations. To overcome the problem, we introduce a neural vocoder (Tamamori et al. 2017; Kalchbrenner et al. 2018; Ping, Peng, and Chen 2019; Prenger, Valle, and Catanzaro 2019; Valin and Skoglund 2019; Ku-

[1] https://github.com/tttslab/spolacq.git

mar et al. 2019; Wang, Takaki, and Yamagishi 2020; Yamamoto, Song, and Kim 2020; Yang et al. 2020; Kong, Kim, and Bae 2020; Yang et al. 2021; Chen et al. 2021a; Kong et al. 2021; Chen et al. 2021b; Cong et al. 2021; Jang et al. 2021; Okamoto et al. 2021) that can synthesize high-fidelity speech waveforms as the agent's speech organ, substituting the sound dictionary. In the self-supervised scenario, the agent needs to try variations of pronunciations on the fly during the dialogue to investigate whether they are successful or not. Therefore, we use a generative vocoder that produces variations of waveforms based on random samplings in the inference.

Although acoustic features, such as mel-spectrograms, are used in neural text-to-speech as the input, we control the neural vocoder using a fixed dimensional action vector inspired by the unconditional training and synthesis in DiffWave (Kong et al. 2021). While a simple WaveNet vocoder can synthesize high-fidelity speech waveforms, the synthesis speed is slow due to the auto-regressive structure (Tamamori et al. 2017). To accelerate the agent's training and inference speed, we adopt WaveGrad (Chen et al. 2021a) as a real-time neural vocoder based on preliminary experiments comparing Parallel WaveGAN (Yamamoto, Song, and Kim 2020), WaveGrad, and DiffWave vocoders.

Related works

As a pioneering constructive engineering work with spoken language acquisition, Gorin *et al.* developed an automatic call-routing system that detected new words in speech utterance and added them in the recognition vocabulary (Gorin, Levinson, and Sankar 1994). It applied an out-of-vocabulary detection and worked with speech utterance and call routing pairs without using text labels. Iwahashi implemented an advanced physical robot system that used a hidden Markov model (HMM) and stochastic context-free grammar (SCFG) (Iwahashi 2000). The system learned vocabulary from isolated word pronunciations and an association between words and images based on mutual information. The robot acted by moving an arm based on a heuristic program, combining the learned results. Roy *et al.* proposed a system to learn vocabulary from a continuously spoken utterance (Roy and Pentland 2002). It could also learn associations between words and image objects. However, it needed a pre-trained phonetic speech-recognizer. The system aimed to acquire image-grounded language knowledge and did not include action generation. Taniguchi *et al.* proposed a system based on a hierarchical Bayes model, which realized an integrated representation of language knowledge (Taniguchi et al. 2020). It needed a pre-trained phoneme recognizer as the starting point of vocabulary learning.

Per the research on learning the pronunciations of new words, Taguchi *et al.* proposed a system to recognize and utter new place names by assuming a trained phoneme model (Taguchi et al. 2011). It learned the place names from utterance and location coordinates, where the coordinate is estimated based on a laser range-finder. Zuo *et al.* proposed a method to interactively update pronunciation (Zuo et al. 2013). They oriented rudimentary investigation, and the method needed a pre-trained phoneme recognizer and a

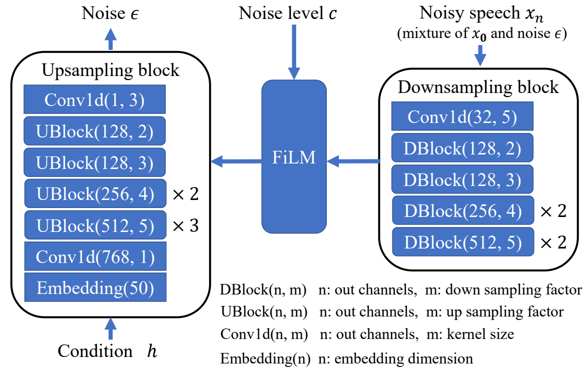


Figure 1: WaveGrad network structure used in our speaking agent. It consists of up and down sampling blocks and a Feature-wise Linear Modulation (FiLM) module.

pre-defined grammar that modeled the dialogue story. These studies adapt the pronunciation at a transcribed phone level and do not adapt the speech synthesizer.

Learning these systems, except for (Gorin, Levinson, and Sankar 1994) and (Zuo et al. 2013), were based on co-occurrence modelings of speech sounds, images, and actions, where the former was based on the error feedback (Gorin et al. 1991). For the language acquisition agents, actions are to interact with the environment. Pronouncing an utterance corresponds to moving a stone or pressing a controller button for game-playing agents (Silver and Huang 2016; Mnih et al. 2015). However, the co-occurrence modeling-based systems lack a mechanism to learn the effect of executing an action as the means to satisfy its goal through trial and error. Therefore, the agent's actions are limited to reproducing those directly thought by teachers. As a reinforcement learning-based language acquisition system, Hatori *et al.* proposed a robot arm system that learns to follow spoken instructions to move the arm (Hatori et al. 2018). While the system could learn the instructions, it assumed a pre-trained speech recognition system to transcribe speech utterances.

WaveGrad vocoder

WaveGrad is one of the diffusion probabilistic vocoders based on denoising score matching (Vincent 2011) and diffusion probabilistic model (Ho, Jain, and Abbeel 2020). It models the diffusion process from a speech sound to random noise and generates speech sounds as its inverse process (Chen et al. 2021a; Kong et al. 2021). Compared with conventional non-autoregressive neural vocoders, the diffusion probabilistic vocoders can be trained with a simple loss function in the time domain. It does not use the generative adversarial network (GAN) (Goodfellow et al. 2014) training as in (Kumar et al. 2019; Yamamoto, Song, and Kim 2020; Yang et al. 2020; Kong, Kim, and Bae 2020; Yang et al. 2021; Cong et al. 2021; Jang et al. 2021; Bińkowski et al. 2020).

It trains a neural network ϵ_θ that predicts Gaussian white noise ϵ from the mixture of speech waveform x_0 and noise ϵ. Figure 1 shows an implementation example of ϵ_θ, which we use in our experiment. It consists of up and down sampling blocks used in GAN-TTS (Bińkowski et al. 2020) and a Feature-wise Linear Modulation (FiLM) (Perez et al. 2018) module. To control the sequential diffusion process, it uses a gradually increasing noise schedule $\beta_1, \beta_2, \cdots, \beta_N$, where N is the number of diffusion steps (Ho, Jain, and Abbeel 2020). The network ϵ_θ has inputs h to control the speech waveform content and c to indicate the noise level at each step. The time domain loss function for the training is defined as shown in Equation (1).

$$\mathbb{E}_{\epsilon,c}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}}\epsilon, h, c\right)\right\|_1\right], \qquad (1)$$

where $\sqrt{\bar{\alpha}}$ is a random sample drawn uniformly between $\sqrt{\bar{\alpha}_n}$ and $\sqrt{\bar{\alpha}_{n-1}}$, $c = \sqrt{\bar{\alpha}}$, $\bar{\alpha}_n = \prod_{s=1}^{n} \alpha_s$, and $\alpha_n = 1 - \beta_n$.

In the inference, input Gaussian white noise $x_N \sim \mathcal{N}(0, I)$ is iteratively converted into a speech waveform by the denoising process based on Langevin dynamics (Song and Ermon 2019) with $n = N, N-1, \cdots, 1$ as follows:

$$x_{n-1} = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{1-\alpha_n}{\sqrt{1-\bar{\alpha}_n}}\epsilon_\theta\left(x_n, h, c\right)\right) + \sigma_n z, \quad (2)$$

where $\sigma_n = \sqrt{\beta_n(1-\bar{\alpha}_{n-1})/(1-\bar{\alpha}_n)}$, $z \sim \mathcal{N}(0, I)$ for $n > 1$, and $z = 0$ for $n = 1$.

Typically, h is an acoustic feature sequence (Chen et al. 2021a) or phoneme sequence (Chen et al. 2021b) that represents a sentence for text-to-speech. It can also be a one-hot vector as in DiffWave (Kong et al. 2021) to specify a word for unconditional training and synthesis. We adopt the latter strategy for our speaking agent.

Spoken language acquisition systems

Sound segment dictionary based baseline system

We use Zhang's sound segment dictionary-based spoken language acquisition agent (Zhang et al. 2020) as a baseline. The agent does not know any specific language initially; it learns a language in a self-supervised manner. The learning process consists of an observation phase and a dialogue phase. In the observation phase, the agent observes untranscribed speech utterances and utterance-image pairs. In the dialogue phase, the agent interacts with the environment and learns to speak through trial and error. The environment here means the outside world for the agent, including a dialogue partner who recognizes spoken utterances. The agent has an internal desire that depends on an agent's internal state, and the goal of the dialogue for the agent is to satisfy its desire. The internal state represents the agent's preference, mood, nutritional condition, etc., depending on the design by the agent's creator. The environment can get the information about the agent's internal state only through spoken dialogue. An intuition behind their system is that human babies first observe speech sounds uttered by their parents and people around them. After obtaining some initial knowledge, babies try to use it to communicate with others, driven by internal motivation.

Figure 2 shows the system structure. The right-hand side (b) is the main body of the agent that interacts with the environment in the dialogue phase. It consists of an action-value function, an internal reward evaluator, and a sound segment dictionary. The reward evaluator defines the agent's inner desire, where the agent is rewarded when the desire is satisfied. The left-hand side (a) performs unsupervised learning in the observation phase and initializes the agent.

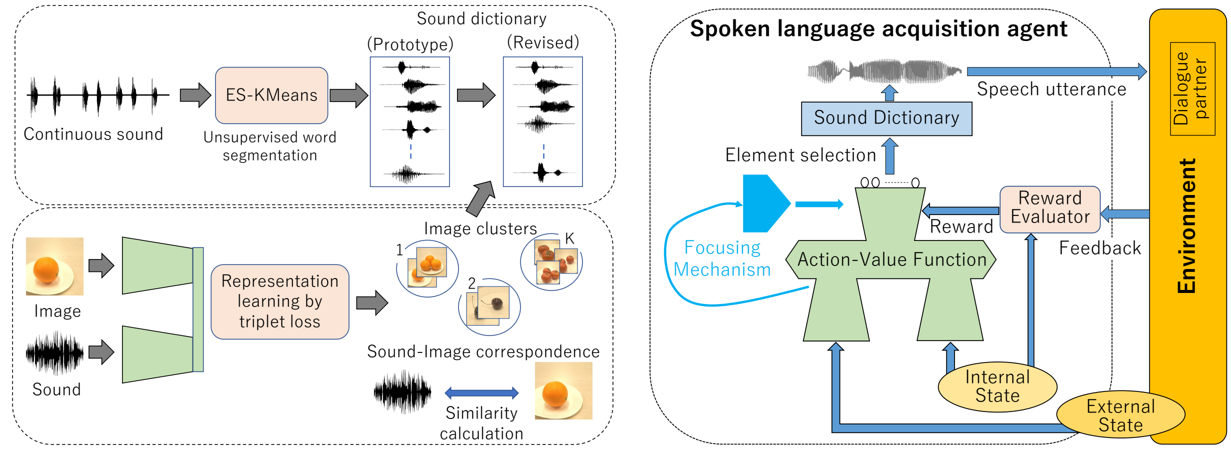
In the observation phase, the agent first makes candidates of possible word units from the observed sounds by an ES-KMeans (Kamper, Livescu, and Goldwater 2017)-based segmentation and makes a prototype sound dictionary. Then, it learns sound-image correspondence from the sound-image pairs using a triplet loss-based unsupervised cross-modal representation learning (Harwath, Torralba, and Glass 2016; Ilharco, Zhang, and Baldridge 2019). The representation learning makes image and sound front-end networks that map sound and image inputs to the same feature space. The agent uses the networks to initialize the input modules of the action-value function. The agent also classifies the image samples into K clusters by applying K-Means clustering in the feature space. Ideally, the K clusters correspond to image object types. For each cluster's centroid, L closest sound segments are selected from the prototype sound dictionary by mapping the waveforms to the features using the sound front-end. A revised sound dictionary is made from the selected $K \times L$ sound segments, which the agent uses as the action space. Besides, the agent uses the learned sound-image correspondence to initialize a focusing mechanism, which helps efficient learning in the dialogue phase by guiding the agent's attention to those word entries in the dictionary related to the current eyesight.

In the dialogue phase, the agent tries to interact with the environment by speaking. Initially, the agent has no idea how to pronounce words to gain the reward. Therefore, the agent randomly selects and replays a segment in the sound dictionary, with the help of the focusing mechanism. Many of the segments are broken fragments, but some are meaningful words. The environment or the dialogue partner listens to the agent's utterance. Based on the recognition result, the environment gives feedback, such as food, to the agent. The agent equips with a reward evaluator inside, and becomes happy if it successfully gets desired feedback. Based on Q-learning (Watkins and Dayan 1992), the agent gradually learns which segment to choose, understanding the meaning of pronouncing it for the current internal and external states.

Proposed system

We replace the sound dictionary in the baseline system with a trainable speech synthesizer. The synthesizer should be able to produce variations of waveforms for the same utterance so that the agent can gradually adjust it. Additionally, it should have few tuning factors and be computationally efficient to fit with the self-supervised learning framework. For these reasons, we choose the WaveGrad method. As shown in Figure 3, the new agent has the same structure as the original one except for the WaveGrad-based speech organ. The action vector obtained from the action-value function is used as the condition h for the WaveGrad.

We integrate the training of WaveGrad into the self-



(a) Modules used for unsupervised learning in observation phase (b) Agent structure that interacts with environment in dialogue phase

Figure 2: Baseline agent system that utilizes sound dictionary.

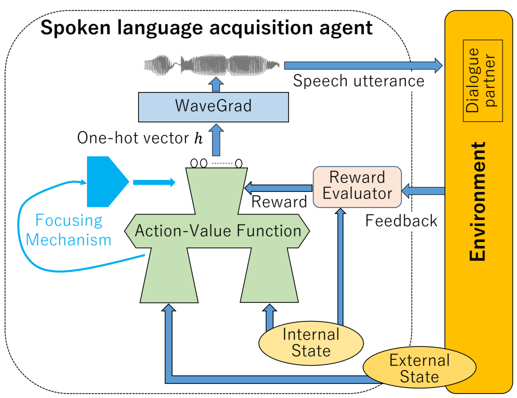


Figure 3: Structure of proposed agent having WaveGrad speech organ.

supervised learning framework as shown in Algorithm 1. In the algorithm, the agent first uses the sound dictionary approach with N_A dialogue episodes. During the dialogue-based learning process, it records its experience. Using the record, the agent makes a data set D_0 of pairs of an index of the sound dictionary element and the corresponding waveform segment obtained from a successful dialogue episode. Because the original sound dictionary contains many junk segments, the agent compresses it by removing those elements that have never resulted in successful episodes. Accordingly, the agent prunes action-value function's output units that correspond to the removed entries. The index IDs in D_0 are also updated for consistency. Then the agent uses the modified D_0 to train the WaveGrad synthesizer that takes the element index represented by a one-hot vector as the condition h, and generates the corresponding waveform as the output.

After obtaining the trained WaveGrad, the agent plugs in it on top of the action-value function's output and starts using it. In the succeeding dialogues, the agent makes a new data set D_l of pairs of an action vector and a generated sound waveform for every N_B episodes by choosing successful experiences and updates the WaveGrad module using it.

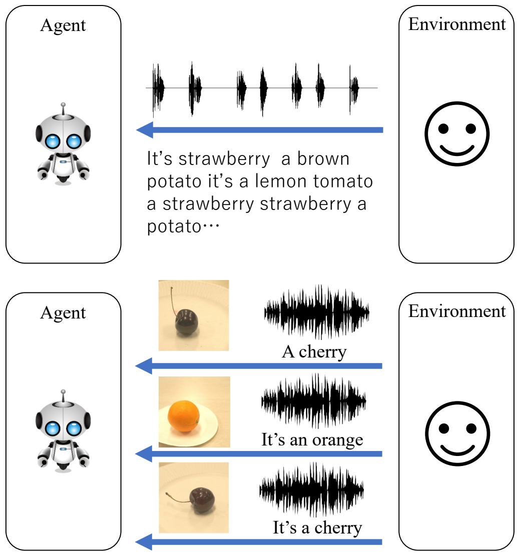
Algorithm 1: WaveGrad speech organ-based spoken language acquisition

Train the sound segment dictionary-based language acquisition agent with N_A dialogue episodes.
Make a set of paired data D_0 of an index of the sound dictionary element and the corresponding waveform segment that is used in a successful dialogue episode.
Train a WaveGrad speech synthesizer using D_0. Replace the sound dictionary of the agent with the trained WaveGrad synthesizer.
for $l = 1$ to L_T **do**
    Repeat the dialogue episodes for N_B times and train the action-value function of the agent.
    Make a set of paired data D_l of the action vector and the generated sound waveform recorded from the successful episodes.
    Update WaveGrad speech synthesizer using D_l.
end for

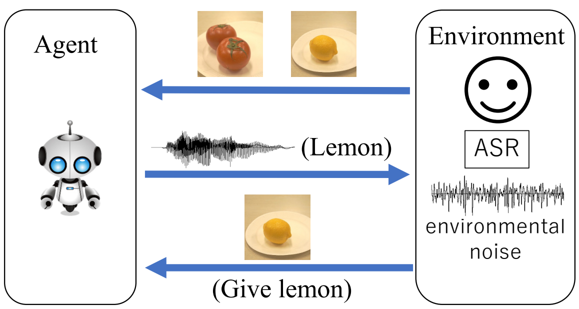
Language acquisition task

We evaluate the agents with a task of obtaining a favorite food. In the world, there are eight types of foods; cherry, green pepper, lemon, orange, potato, strawberry, sweet potato, and tomato.

Figure 4a shows the observation phase. In this phase, the agent first listens to a long sound signal that contains utterances mentioning food names in random order with random intervals of 1 to 3 seconds. The utterances are generated based on templates of "<food>," "A <food>," "A <color> <food>," and "It's a <food>", where <food> is the food name and <color> is a color name that explain the food. For example, if the food is cherry, they are "Cherry," "A cherry," "A red cherry," and "It's a cherry". Each of the eight food names appears 90 times. The agent utilizes this signal to make the prototype sound dictionary.



(a) Observation phase



(b) Dialogue phase

Figure 4: Spoken language acquisition task.

Then, the agent is shown a food photo with an utterance indicating the food. The indication repeats 2880 times for combinations of 90 photos of each food, four types of the utterances, and eight types of the foods. All the photos are different, and they have one to three food objects of the same type.

In the dialogue phase, the agent has a random color as its internal state at each dialogue episode. The agent wants a food object with an average color closer to the internal state, where the Euclidean distance is measured in the RGB space. As shown in Figure 4b, the environment shows two food photos to the agent in each episode, and the agent speaks an utterance. The environment chooses the two photos by random sampling with replacement from a pool that contains 30 photos for each food type. The environment listens the agent's utterance and recognizes it. If the recognition result is one of the shown food names, the environment gives the food to the agent. The agent gets a reward of $r = 1$ if the obtained food is what it desires between the two. In case the agent receives the opposite food or gets nothing, it gets a reward of $r = 0$. The agent regards a dialogue episode as successful if $r = 1$. The photo samples used in the dialogue phase have no overlap with those used in the observation phase.

We generated the utterances by using the Google Text-To-Speech library [2]. The food images are a subset of those gathered in (Zhang et al. 2020). To better simulate a realistic situation, we add 30 dB white noise to the agent's utterance as a background environmental noise. To evaluate the agent's adaptability to an environmental change, we further add a 10 dB, 300 Hz sine wave noise after repeating a certain number of the dialogue episodes. To implement the speech recognition function of the environment, we use ESPnet (Watanabe et al. 2018) with a super multilingual speech model (Hou et al. 2020) [3] instead of the Google's speech recognition API used in the original Zhang's system [4].

Evaluation measure

We evaluate the agent's learning performance by an average reward for the number of dialogue episodes. Because of the reward definition, the average reward becomes 1.0 when the agent always pronounces the right food name and the environment correctly recognizes it. In other words, it is the agent's side's responsibility to make a clear pronunciation so that the speech recognizer in the environment can accurately recognize it. When the agent uses the baseline sound dictionary approach, the unclear pronunciation's main cause is the blunted word due to inaccurate segmentation. When the agent uses the WaveGrad speech organ, the speech synthesis performance becomes another factor that affects it. As we add the environmental noises to the agent's utterance, the agent needs to somehow make its pronunciation clear under the noisy condition.

To evaluate the agent's pronunciation separately from the action selection, we additionally evaluate the valid word recognition rate (VWRR). We define it as a ratio that the recognition result is one of the eight food types among the total number of the agent's pronunciation. The VWRR becomes 0% if the environment never recognizes agent's pronunciation even if the agent internally decides the right food name. It becomes 100% if the environment always recognizes the agent's pronunciation as one of the eight food names.

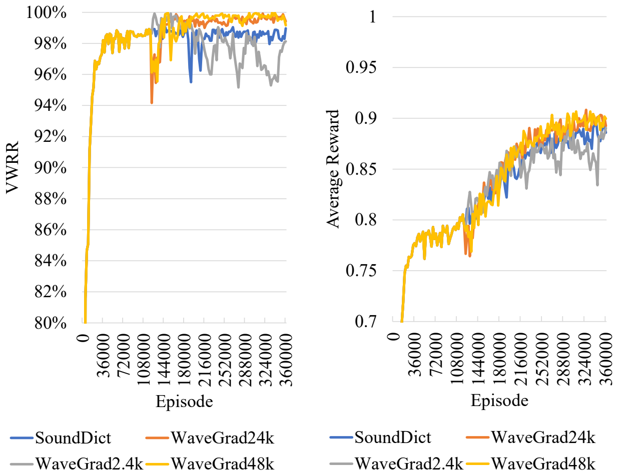
Experimental setup

For the sound dictionary based-agent, we mostly followed the network structure and the learning setup used in (Zhang et al. 2020). Some differences were that we set the cluster number K to 120 and per the cluster segment number L to 100 when making the sound dictionary, and the parameter $\lambda = 0.97$ used in the action filter of the focusing mechanism. The sound dictionary size was $KL = 12000$. We set

[2]https://pypi.org/project/gTTS/

[3]https://github.com/Porridge144/sup-mlt-demo

[4]Due to the limit of the number of API calls, we could not use Google's speech recognition in this work. Zhang et al. first recognized all the sound segments in the sound dictionary and cached the results because the pronunciations did not change in their experiment. However, we have to recognize the agent's utterance every time since the agent's pronunciation changes dynamically. Related to this change, we shrank the number of the food categories to eight because some of the food names did not exist in the environment's recognition dictionary.



(a) VWRR (b) Average reward

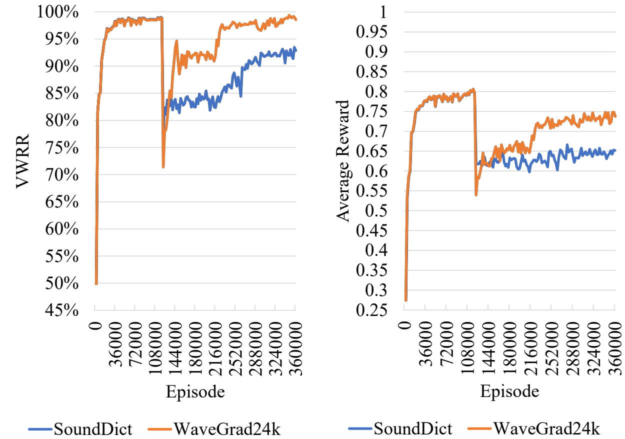
Figure 5: Valid word recognition rate (VWRR) and reward when the environmental noise is steady. The environment always hears the agent's utterance with a 30 dB background white noise. We obtain the plot by calculating an average score at every 2400 episodes.

the number of the initial dialogue episodes $N_A = 120000$. For the number of the succeeding dialogue episodes N_B, we investigated 2400, 24000 and 48000. We refer to them as WaveGrad2.4k, WaveGrad24k, and WaveGrad48k systems, respectively. For the WaveGrad noise schedule, we used the arithmetic progression with $N = 50$, $\beta_1 = 0.0001$ and $\beta_N = 0.05$. The sampling rate of the waveforms was 8000 Hz.

Results

Figure 5 shows the results when the environmental condition does not change, where the environment always hears the agent utterance with 30 dB background white noise. In the figure, we denote the baseline system as SoundDict. It kept using the same sound dictionary through the dialogue episodes. The proposed systems switched to the WaveGrad speech organ after 120000 episodes. The action vector size was 161 after the dictionary entry pruning.

We observe in Figure 5(a) that VWRR of WaveGrad2.4k after 120000 episodes first increases but soon begins to decrease. This was probably because of an unstable parameter update of WaveGrad due to the small number of samples. When we used the update interval of 24k or 48k, we observed mostly a monotonic increase of VWRR after the switching, and they gave slightly better performance than the baseline. The baseline system showed a relatively flat VWRR for the dialogue episodes larger than 120000 because there was no mechanism to adjust the pronunciation. Once the agent identifies useful entries in the sound dictionary, it keeps using them. Figure 5(b) shows the plot of the averaged reward. The VWRRs of WaveGrad24k and WaveGrad48k systems are slightly better than the baseline and WaveGrad2.4k was slightly worse. However, the scores of the four systems had overall similar trends. This was because even the lowest VWRR by WaveGrad2.4k after 120000 episodes was more than 94%.



(a) VWRR (b) Average reward

Figure 6: Valid word recognition rate (VWRR) and reward when the environmental noise changes at 120000 episodes. After 120000 episodes, the 10 dB sine wave noise is added, along with the 30 dB white noise.

Figure 6 shows the results when the noise environment changes at 120000 episodes, where we add 10 dB sine wave noise. The VWRR and the reward largely dropped at 120000 episodes because of the sine wave noise. The VWRR recovers with the repetition of the episodes. However, the baseline system had limited improvement because the agent could only choose alternative entries in the sound dictionary. The WaveGrad system had a more significant improvement because of the flexible learning ability. When we listened to the agent's utterances, they were natural. The average VWRRs at 360000 episodes by the proposed and baseline systems were 98.5% and 92.8%, respectively. We observed the same trend with the reward, where they were 0.737 and 0.652 at 360000 episodes. These results prove the superior adaptability of the WaveGrad based speaking agent to the baseline.

Conclusion

We have proposed a spoken language acquisition system using a WaveGrad speech organ. Compared to existing sound dictionary-based systems, the proposed approach has higher flexibility with the utterance pronunciation and superior adaptability to the changing environment. While recent unsupervised neural speech synthesis studies can randomly generate word-like sounds, our work is the first to synthesize sound utterances based on their meaning. Future work includes extending the utterance's expressive power, such as realizing emotional expression, longer sentences, and few-shot learning for quick vocabulary expansion in the real world. Additionally, many existing acoustic model adaptation techniques will be helpful to improve the pronunciation adaptation efficiency.

Acknowledgments

This work was supported by Toray Science Foundation.

References

Bińkowski, M.; Donahue, J.; Dieleman, S.; Clark, A.; Elsen, E.; Casagrande, N.; Cobo, L. C.; and Simonyan, K. 2020. High fidelity speech synthesis with adversarial networks. In *Proc. ICLR*.

Chen, N.; Zhang, Y.; Zen, H.; Weiss, R. J.; Norouzi, M.; and Chan, W. 2021a. WaveGrad: Estimating Gradients for Waveform Generation. In *Proc. ICLR*.

Chen, N.; Zhang, Y.; Zen, H.; Weiss, R. J.; Norouzi, M.; Dehak, N.; and Chan, W. 2021b. WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis. In *Proc. Interspeech*.

Cong, J.; Yang, S.; Xie, L.; and Su, D. 2021. Glow-WaveGAN: Learning Speech Representations from GAN-based Variational Auto-Encoder For High Fidelity Flow-based Speech Synthesis. In *Proc. Interspeech*.

Dupoux, E. 2018. Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173: 43–59.

Gao, S.; Hou, W.; Tanaka, T.; and Shinozaki, T. 2020. Spoken Language Acquisition Based on Reinforcement Learning and Word Unit Segmentation. In *Proc. ICASSP*, 6149–6153. IEEE.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proc. NIPS*, 2672–2680.

Gorin, A.; Levinson, S.; Gertner, A.; and Goldman, E. 1991. Adaptive acquisition of language. *Computer Speech and Language*, 5(2): 101–132.

Gorin, A.; Levinson, S.; and Sankar, A. 1994. An experiment in spoken language acquisition. *IEEE Transactions on Speech and Audio Processing*, 2(1): 224–240.

Harwath, D.; Torralba, A.; and Glass, J. 2016. Unsupervised learning of spoken language with visual context. *Advances in Neural Information Processing Systems*, 29: 1858–1866.

Hatori, J.; Kikuchi, Y.; Kobayashi, S.; Takahashi, K.; Tsuboi, Y.; Unno, Y.; Ko, W.; and Tan, J. 2018. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *Proc. ICRA*, 3774–3781.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Proc. NeurIPS*, 6840–6851.

Hou, W.; Dong, Y.; Zhuang, B.; Yang, L.; Shi, J.; and Shinozaki, T. 2020. Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning. In *Proc. Interspeech*, 1037–1041.

Ilharco, G.; Zhang, Y.; and Baldridge, J. 2019. Large-Scale Representation Learning from Visually Grounded Untranscribed Speech. In *Proc. CoNLL*, 55–65.

Iwahashi, N. 2000. Language acquisition through a human-robot interface. In *Proc. ICSLP*, volume 3, 442–447.

Jang, W.; Lim, D.; Yoon, J.; Kim, B.; and Kim, J. 2021. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. Interspeech*.

Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; van den Oord, A.; Dieleman, S.; and Kavukcuoglu, K. 2018. Efficient neural audio synthesis. In *Proc. ICML*, 2415–2424.

Kamper, H.; Livescu, K.; and Goldwater, S. 2017. An embedded segmental K-means model for unsupervised segmentation and clustering of speech. In *Proc. ASRU*, 719–726.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proc. NeurIPS*, 17022–17033.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *Proc. ICLR*.

Kuhl, P. 2004. Early Language Acquisition: Cracking the Speech Code. *Nature reviews. Neuroscience*, 5: 831–43.

Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W. Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; and Courville, A. C. 2019. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In *Proc. NeurIPS*, 14910–14921.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Okamoto, T.; Toda, T.; Shiga, Y.; and Kawai, H. 2021. Noise level limited sub-modeling for diffusion probabilistic vocoders. In *Proc. ICASSP*, 6029–6033.

Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; and Courville, A. C. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proc. AAAI*, 3942–3951.

Ping, W.; Peng, K.; and Chen, J. 2019. ClariNet: Parallel wave generation in end-to-end text-to-speech. In *Proc. ICLR*.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. WaveGlow: A flow-based generative network for speech synthesis. In *Proc. ICASSP*, 3617–3621.

Roy, D. K.; and Pentland, A. P. 2002. Learning Words from Sights and Sounds: a Computational Model. *Cognitive Science*, 26(1): 113–146.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.; Saurous, R. A.; Agiomyrgiannakis, Y.; and Wu, Y. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP*, 4779–4783.

Silver, D.; and Huang, A. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529: 484–489.

Skinner, B. 1957. *Verbal Behavior*. New York: Appleton-Century-Crofts.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Proc. NeurIPS*, 11918–11930.

Taguchi, R.; Yamada, Y.; Hattori, K.; Umezaki, T.; Hoguro, M.; Iwahashi, N.; Funakoshi, K.; and Nakano, M. 2011. Learning Place-Names from Spoken Utterances and Localization Results by Mobile Robot. In *Proc. Interspeech*, 1325–1328.

Tamamori, A.; Hayashi, T.; Kobayashi, K.; Takeda, K.; and Toda, T. 2017. Speaker-dependent WaveNet vocoder. In *Proc. Interspeech*, 1118–1122.

Taniguchi, T.; Nagai, T.; Nakamura, T.; Iwahashi, N.; Ogata, T.; and Asoh, H. 2016. Symbol emergence in robotics: a survey. *Advanced Robotics*, 30(11-12): 706–728.

Taniguchi, T.; Nakamura, T.; Suzuki, M.; Kuniyasu, R.; Hayashi, K.; Taniguchi, A.; Horii, T.; and Nagai, T. 2020. Neuro-SERKET: Development of Integrative Cognitive System Through the Composition of Deep Probabilistic Generative Models. *New Generation Computing*, 38(1).

Valin, J.-M.; and Skoglund, J. 2019. LPCNet: Improving neural speech synthesis through linear prediction. In *Proc. ICASSP*, 5826–7830.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Viñyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. . 2016. WaveNet: A generative model for raw audio. In *Proc. SSW9*, 125.

Vincent, P. 2011. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7): 1661–1674.

Wang, X.; Takaki, S.; and Yamagishi, J. 2020. Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 28: 402–415.

Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.-E. Y.; Heymann, J.; Wiesner, M.; Chen, N.; Renduchintala, A.; and Ochiai, T. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proc. Interspeech*, 2207–2211.

Watkins, C. J. C. H.; and Dayan, P. 1992. Q-learning. *Machine Learning*, 8(3): 279–292.

Xiong, W.; Droppo, J.; Huang, X.; Seide, F.; Seltzer, M.; Stolcke, A.; Yu, D.; and Zweig, G. 2017. Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 25(12): 2410–2423.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. ICASSP*, 6199–6203.

Yang, G.; Yang, S.; Liu, K.; Fang, P.; Chen, W.; and Xie, L. 2021. Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech. In *Proc. SLT*, 492–498.

Yang, J.; Lee, J.; Kim, Y.; Cho, H.; and Kim, I. 2020. VocGAN: A High-Fidelity Real-time Vocoder with a Hierarchically-nested Adversarial Network. In *Proc. Interspeech*, 200–204.

Zhang, M.; Tanaka, T.; Hou, W.; Gao, S.; and Shinozaki, T. 2020. Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition. In *Proc. Interspeech*, 4183–4187.

Zuo, X.; Sumii, T.; Iwahashi, N.; Nakano, M.; Funakoshi, K.; and Oka, N. 2013. Correcting phoneme recognition errors in learning word pronunciation through speech interaction. *Speech Communication*, 55(1): 190–203.