INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition

Mingxin Zhang, Tomohiro Tanaka, Wenxin Hou, Shengzhou Gao, Takahiro Shinozaki

Tokyo Institute of Technology
www.ts.ip.titech.ac.jp

Abstract

The process of spoken language acquisition based on sound-image grounding has been one of the topics that has attracted the most significant interest of linguists and human scientists for decades. To understand the process and enable new possibilities for intelligent robots, we designed a spoken-language acquisition task in which a software robot learns to fulfill its desire by correctly identifying and uttering the name of its preferred object from the given images, without relying on any labeled dataset. We propose an unsupervised vision-based focusing strategy and a pre-training approach based on sound-image grounding to boost the efficiency of reinforcement learning. These ideas are motivated by the introspection that human babies first observe the world and then try actions to realize their desires. Our experiments show that the software robot can successfully acquire spoken language from spoken indications with images and dialogues. Moreover, the learning speed of reinforcement learning is significantly improved compared to several baseline approaches.

Index Terms: spoken language acquisition, speech understanding, human-computer interaction

1. Introduction

The performance of automatic speech-recognition systems has become comparable with or even surpassed humans in several conditions. However, its performance relies on a large amount of transcribed speech data. Furthermore, this need for transcribed labels is an essential limitation in many more applications that require individual adaptations to new linguistic expressions and their semantic understanding. For example, for robots that coexist with humans, they will need this ability.

In principle, it is possible for an end-to-end neural-network-based dialogue system to learn a spoken language through dialogue without relying on a transcribed speech corpus by reinforcement learning. However, merely applying reinforcement learning to speech response learning requires an unrealistic number of trials because the action is a speech utterance whose dimension can be very high. Unsupervised learning such as word-unit learning and sound-and-image object grounding is another direction of automatic spoken-language learning. Although these methods have been proven to work in principle, most of them suffer from low accuracy. On the other hand, humans have a very flexible and strong capability for spoken language learning. A hypothesis in the neurobiology area states that humans combine various types of learning functions in a certain manner to form a superior learning system [1].

A software robot has been developed in [2] that learns spoken language by combining Deep Q-learning (DQN) [3]-based reinforcement learning and ES-KMeans [4]-based unsupervised word learning. The key idea of the research was to make a sound dictionary through the unsupervised word learning from speech and then use the sound dictionary as the action space in the reinforcement learning to reduce the search space significantly. However, one limitation of the robot is that the action space is linear to the dictionary size. With the increase in the size of the dictionary, which includes broken segments, the chance of making correct utterance decreases linearly, especially at the beginning of the learning.

In human spoken-language learning, vision plays an important role in guiding the focus. For example, young children learn object names from picture books and recitation by their parents. With the conversations about the objects in the picture, pronunciation accuracy, and the spoken language understanding is improved. In this paper, we propose a new learning algorithm that realizes the vision-aided, efficient spoken-language learning by introducing an unsupervised sound-image learning module to the sound-dictionary-based spoken-language learning robot [1].

2. Related Work

2.1. Grounded Language Learning

Grounded language learning is an area of research that is closely related to our task. The main idea is to associate an abstract term in language with tangible objects such as images or actions or to perform classification on this term [5, 6, 7]. In the early years, Siskind presented a non-statistical language-grounding model consisting of many handmade logics [8, 9]. Several reinforcement-learning-based methods have been proposed to automatically obtain high-performance grounding models. Hermann et al. [10] introduced a language acquisition model that moves the robot around in a virtual 3D environment following orders. Yu et al. [11] proposed a language acquisition model for question answering and sentence-directed navigation trained by interacting with the virtual 2D world. Sinha et al. [12] presented an attention-based language-grounding model that navigates the user to the place specified in a given description sentence. Sigurdsson et al. [13] proposed a model to improve unsupervised word translation by using visual grounding. However, those models are text-based and, therefore, do not simulate spoken language-learning processes per se.

2.2. Visual-Audio Correspondence Learning

Several works study word unit learning from raw sound and audio-visual grounding without using annotation labels. One approach is to apply clustering to audio-visual paired inputs [14]. Another approach is to train a binary classifier to judge if these audio and visual samples correspond to the same event or not [15, 16, 17, 18, 19]. Although spoken language acquisition is more than just learning correspondence, these ideas

[1] We will release our system and dataset at our web page www.ts.ip.titech.ac.jp.

http://dx.doi.org/10.21437/Interspeech.2020-2027

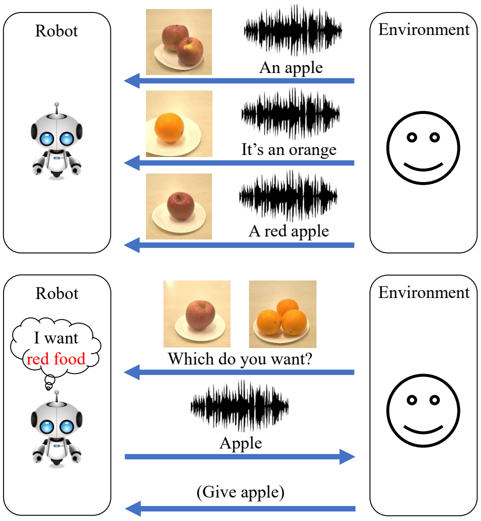

Figure 1: *The top part represents the indication phase while the bottom part represent the question answering dialogue phase of the designed task. The robot does not use any text label.*

serve as the foundation for our task's setting.

3. Spoken Language Acquisition Task

We designed a spoken language acquisition task with a software robot, in which the robot is analogized to be a newly-born infant who is an active language learner. The robot is shown with an image of a food item randomly selected from a finite set. At the same time, a vocal description of the image's content is given. Example images and their corresponding descriptions are shown in Figure 1. The description is only given in waveform without text. After certain rounds of "study", the robot is then given a pair of images with the question: "Which do you want?". The robot has an internal preference for the color of the food, which is randomly assigned at each one-turn episode. Speech recognition on the vocal answers from the robot is performed by the environment. The robot is rewarded when it speaks the correct food name with the preferred color; it is not rewarded when it speaks the wrong food name or says non-meaningful words. For example, the robot is supposed to answer "Orange" in a case when the internal color preference is yellow or "Apple" when the color preference is red, as in the situation shown in Figure 1.

4. Preliminaries

4.1. Deep Q-Learning

The Deep Q-learning (or Deep Q-Network, DQN) method is a variant of Q-learning. It was introduced by Mnih et al. [3] to tackle highly complex situations in reinforcement learning. DQN has been proven to be effective in many challenging tasks, such as computer resource management [20], robotics [21], and even chemistry [22].

4.2. Unsupervised Sound-Image Learning Algorithm

The unsupervised sound-image correspondence learning is introduced by Harwath et al. [15]. In their model, an image and the corresponding audio description are fed into the visual and audio extractors, respectively. The extracted feature vectors are then tied together by calculating their L2 distance, which is referred to as the similarity score. In this way, the model automatically learns associations between images and audio descriptions. To train the model in an unsupervised manner, Harwath et al. employed a variant of the triplet loss. The goal of triplet loss-based learning is to enlarge the differences of the similarity scores between the ground-truth pairs and others. We train the network by using the triplet loss shown in equation (1) and (2).

$$L = t\left(f_I\left(x_I\right), f_S\left(x_S^+\right), f_S\left(x_S^-\right)\right), \tag{1}$$

$$t\left(v_I, v_S^+, v_S^-\right) = \max\left(\left\|v_I - v_S^+\right\| - \left\|v_I - v_S^-\right\| + \Delta, 0\right), \tag{2}$$

where x_I and x_S^+ are the original image and the corresponding audio description, x_S^- is a randomly selected audio description from other than x_S^+, v_I is the image feature extracted by f_I, v_S^+ and v_S^- are paired and unpaired sound features extracted by f_S, and Δ is a positive constant.

4.3. Sound Dictionary

The sound dictionary is defined as a set of sound segments. The robot selects and plays one of them to tell its intentions. We make the sound dictionary following [2] from a continuous speech without label information. We apply random-cut based segmentation and ES-KMeans [4] to segment the continuous speech and use the segmented speeches as the sound dictionary.

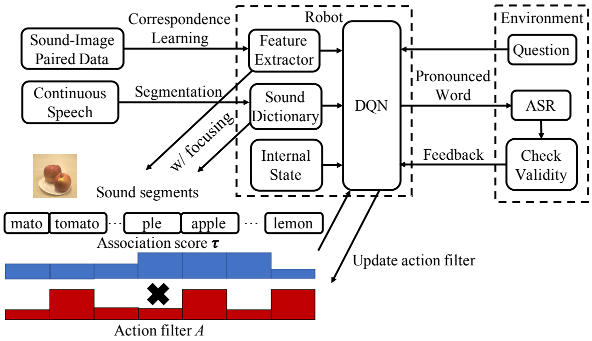

Figure 2: *System structure. The top part structure is the same between the baseline and the proposed systems. The bottom part illustrates the newly introduced focusing mechanism and the action filter.*

5. Spoken Language Acquisition Robots

We show the overall structure of the baseline and the proposed systems in Figure 2. Since the question in the dialogue phase is always the same (i.e., "Which do you want?") due to the task design, the robot receives only two images as the input; there is no explicit question input. We assume that the value range of Q function is 0 to 1.

5.1. Baseline System

The baseline system is a direct application of the sound-dictionary-based method [2] by adding two image front-ends to receive two images. The baseline robot simply selects a sound segment from the sound dictionary when given a pair of images x_{I1} and x_{I2}. According to the internal state s_{in} and the choice of the sound segment index a, the robot gets a reward r.

We use simplified deep Q-learning and adjust it to our task setting. Algorithm 1 describes the process, where x_{I1} and x_{I2}

Algorithm 1 Q-learning for language acquisition

```
Initialize action-value function $Q$ with random weights $\theta$
for episode $= 1$ to $S$ do
    Set a pair of images $x_{I1}, x_{I2}$. Set internal state $s_{in}$
    $s = x_{I1} \oplus x_{I2} \oplus s_{in}$
    Select $a = \arg\max_a Q(s, a; \theta)$
    Execute action $a$ in the environment and obtain reward $r$
    Perform gradient descent with
      $L(\theta) = (r - Q(s, a; \theta))^2$
end for
```

Algorithm 2 Learning of the action filter

```
Initialize action filter $A$ with matrix of all ones
for episode $= 1$ to $S$ do
    Execute action $a$ in environment and obtain reward $r$
     ($a$ is $l$-th sound segment in the $m$-th cluster)
    if $r \geq \beta$ then
        $A[:, m] = \lambda * A[:, m]$
        $A[l, m] = 1$
    else
        $A[l, m] = \lambda * A[l, m]$
    end if
end for
```

are vectorized images, and $\oplus$ indicates concatenation operation. Because the task is not sequential (only one turn interaction at a conversation), we do not need the target Q-network to estimate the reward corresponding to the state and action. The Q-value is just the reward the software robot gets after each interaction with the environment. The Q-network is defined as follows:

$$\acute{s} = g_I(x_{I1}) \oplus g_I(x_{I2}) \oplus g_p(s_{in}), \tag{3}$$

$$Q(s, a; \theta) = \sigma(g_c(\acute{s}))[a], \tag{4}$$

where g_I is a randomly initialized convolution neural network (CNN), and g_p and g_c are multi-layer fully connected neural networks (FCN). The output layer size of g_c is equal to the sound dictionary size, and $\sigma()[a]$ indicates a-th element of the softmax function. The parameters of the two image front-ends are shared. The Q-network is trained to output higher Q-values on the desired food given in the images.

5.2. Proposed System

The proposed system adds pre-training of the image front-end and a vision-based focusing mechanism to the baseline.

5.2.1. Pretraining of image front-ends

In the baseline system, the image front-end g_I is randomly initialized. In the proposed method, we initialize it by using parameters obtained as f_I in the sound-image unsupervised learning. Because f_I is trained so that the output is useful for sound-image pair/unpair judgment, we can expect that f_I is good at distinguishing the food types.

5.2.2. Vision based focusing

The sound dictionary size is usually large; letting the robot directly explore it is inefficient. We can make the learning process more efficient by having the robot focus on a limited number of sound words in the dictionary associated with objects in the robot's view field.

We perform K-means clustering for the image features of the image data using f_I after the indication phase, and obtain M clusters with centroids c_1 to c_M that are expected to correspond to the food type as shown in Equation (5).

$$[c_1, ...c_M] = \text{KMeans}(f_I(x_{I,1}), ...f_I(x_{I,n})), \tag{5}$$

where n is the size of the sound-image pair data. We also apply the sound front-end f_S to each of the sound dictionary entries. The image and sound features have the same dimension of f, and share the same meaning in the space because of the triplet loss-based learning. For each cluster centroid c_m, we select L closest sound segments based on Euclidean distance. We obtain a list of LM sound segments and use it as a new sound dictionary. We define an association score vector $\tau^{x_I} = [\tau_1^{x_I}\tau_2^{x_I}, \cdots, \tau_M^{x_I}]$ between the clusters and an input image x_I by Equations (6) and (7).

$$\hat{\tau}_m^{x_I} = \sigma(-d(f_I(x_I), [c_1, ...c_M]))[m], \tag{6}$$

$$\tau_m^{x_I} = \hat{\tau}_m^{x_I} / \max(\hat{\tau}^{x_I}), \tag{7}$$

where d is the Euclidean distance.

Then, we integrate the association score into the Q-network as a focusing mechanism to boost the probability of choosing the sound segment entries corresponding to images shown to the robot. The network represented by Equations (3) and (4) are extended as Equations (8) to (10).

$$\acute{s} = g_I(x_{I1}) \oplus g_I(x_{I2}) \oplus g_p(s_{in}), \tag{8}$$

$$[\alpha_1, \alpha_2, \alpha_3] = \sigma(\acute{g}_c(\acute{s})) \tag{9}$$

$$Q(s, a, \theta) = \Big(\alpha_1 \operatorname{vec}(\mathbf{1}\tau^{x_{I1}})^{\mathsf{T}} + \alpha_2 \operatorname{vec}(\mathbf{1}\tau^{x_{I2}})^{\mathsf{T}} + \alpha_3 \hat{g}_c(\acute{s})\Big)[a], \tag{10}$$

where $\mathbf{1}$ is a L-dimensional ones vector, vec is vectorization operation, $^{\mathsf{T}}$ is the transpose and $\acute{g}_c, \hat{g}_c$ are randomly initialized FCNs.

5.2.3. Action filter

We introduce a learnable action filter A that helps the robot to select the right sound segment corresponding to the intended object overriding the initial noisy association score through the trials of the question answering. A is a $L \times M$ matrix estimated by an algorithm shown in Algorithm 2, where λ is a coefficient to control the learning speed of A, and β is a threshold for success judgment. We integrate the action filter into the Q-network by extending Equations (3) and (4) to Equations (11) to (14).

$$\acute{s} = g_I(x_{I1}) \oplus g_I(x_{I2}) \oplus g_p(s_{in}), \tag{11}$$

$$[\alpha_1, \alpha_2] = \sigma(\acute{g}_c(\acute{s})), \tag{12}$$

$$h = \alpha_1 \operatorname{vec}(\mathbf{1}\tau^{x_{I1}})^{\mathsf{T}} + \alpha_2 \operatorname{vec}(\mathbf{1}\tau^{x_{I2}})^{\mathsf{T}}, \tag{13}$$

$$Q(s, a; \theta) = \Big(\operatorname{vec}(A)^{\mathsf{T}} * h\Big)[a] + \epsilon, \tag{14}$$

where $\acute{g}_c$ is a randomly initialized FCN, and ϵ is a random noise from 0 to 0.1 added when selecting an action to mimic epsilon-greedy strategy. The network structure represented by Equations (11) to (13) is to select a desired object, and Equations (14) is to select an appropriate sound segment.

6. Experimental Setup and Results

6.1. Dataset Construction

We developed and used our own dataset rather than using existing one because Flickr 8K [23] did not have corresponding questions. The VQA dataset [24] had many human-made questions, but they were too complex for the current system. Our dataset consisted of food images and corresponding audio descriptions. We took photos of 120 sets of food in 20 categories: apple, banana, carrot, cherry, cucumber, egg, eggplant, green pepper, hyacinth bean, kiwi fruit, lemon, onion, orange, potato, sliced bread, small cabbage, strawberry, sweet potato, tomato, and white radish. Audio descriptions were generated using Google Text-To-Speech library [2]. For each category, we applied four templates to generate descriptions of the contents: e.g., "apple," "An apple," "A red apple," and "It's an apple." To make the synthetic data more realistic, we added 20dB of Gaussian noise to audio descriptions. Among the 120 sets of food images, we used 90 to form a training set and 30 for a test set. The training set consisted of 7,200 sound-images pairs (= 20 food categories × 90 image samples × 4 sound descriptions). All the 7,200 waveforms were different because we added random noises. The test set had 600 images samples(= 20 food categories × 30 samples) .

6.2. Task Setting Details

In the indication phase, we used two types of data. One is the sound-image paired data, which we used to train feature extractors f_I and f_S. Another is the continuous speech that is made by concatenating 720 random samples of the audio descriptions, with 1-3 seconds of random intervals and 20dB of Gaussian noise. We used it to make the sound dictionary. In the dialogue phase, the robot is assigned a random preference color at each one-turn episode as an internal state and is randomly shown two food images in the test set as input. The robot wants a food object with an average color closer to its preference, where the Euclidean distance is measured in the RGB space. As the spoken dialogue environment, we used a general-purpose ASR system provided by Google [3]. If the recognition result of the robot's utterance is the name of the preferred food, the robot gets a reward, $r = 1$. If the recognition result is the name of an unintended object or something else, it gets no reward, $r = 0$.

We constructed the sound dictionary in two different ways: random segmentation and ES-KMeans. We ran the reinforcement learning experiments with three different settings: Q-network without focusing with g_I's parameter initialized randomly, Q-network without focusing with g_I's parameter initialized using the pre-trained model, and Q-network with focusing and g_I's parameter initialized using a pre-trained model. We called the first one "Baseline," the second one "Pretrained w/o Focusing," and the last one "Pretrained w/ Focusing." We referred to with and without action filter as "w/ A" and "w/o A". We ran the reinforcement learning three times with different random seeds and obtained averaged rewards.

6.3. Model Hyperparameters

We set K to 2 for the ES-KMeans clustering in the sound dictionary learning. Features size f is 50. For the focusing mechanism, we set the number of K-means clusters M to 40 and the number of sound segments L chosen by each cluster to 500. The updating coefficient λ of Action filter is 0.9, and β is 1. The g_I is ResNet-50 [25].

The shape of the internal state front-end network g_p is (3, 50). For the baseline system, g_c is (150, 75, ds) where the sound dictionary size ds is 2306 when the random segmentation is performed and 1996 when the ES-KMeans is used. The sound dictionary size is $LM = 20000$ when the focusing mechanism is used. When the action filter is not used, $\hat{g}_c$ has size of (150, 75, LM) and $\acute{g}_c$ is (150, 75, 3). When the action filter is used, $\acute{g}_c$ is (150, 75, 2). We determined these settings based on a preliminary experiment.

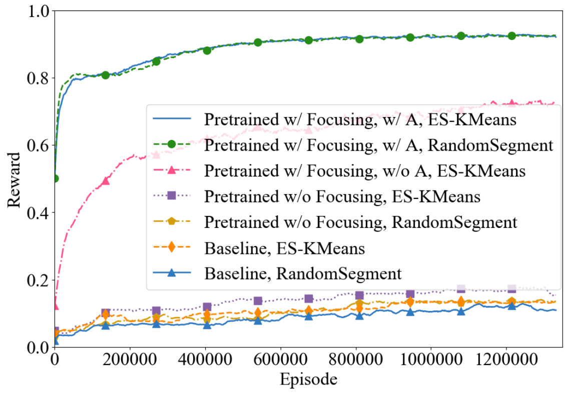

Figure 3: *Average reward in the dialogue phase.*

Table 1: *Final rewards*

Pretrained	Focusing	Action Filter	Segment	Final reward
○	○	○	ES-KMeans	0.920
○	○	○	Random	0.925
○	○	×	ES-KMeans	0.729
○	×	×	ES-KMeans	0.154
○	×	×	Random	0.135
×	×	×	ES-KMeans	0.136
×	×	×	Random	0.109

6.4. Results and Analysis

We show the learning curves of models in Figure 3. The horizontal axis is the number of episodes, and the vertical axis is the reward. We also show the final rewards in Table 1. We can confirm that pretraining of image front-ends improves the learning speed for both segmentation conditions. We can also confirm that the focusing module strongly improved the learning speed. Note that unlike the other results, the ES-KMeans is not effective on the proposed method with the focusing module. This means the focusing module can help the robot select meaningful speech from action space, regardless of the number of obtained meaningful words.

7. Conclusion

We demonstrated the feasibility of a robot in zero-resource spoken language acquisition tasks by leveraging sound-image correspondence. We also took one step further by proposing a novel framework to accelerate this learning process with the combination of unsupervised multimodal pre-training and vision-based focusing strategy. Future works include generalizing the concept with a larger and more realistic dataset, using video data instead of images, and extending the applicability to more general cases by combining dialogue system strategies.

8. Acknowledgements

This work was supported by Toray Science Foundation.

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

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

9. References

[1] K. Doya, “Complementary roles of basal ganglia and cerebellum in learning and motor control,” *Current Opinion in Neurobiology*, vol. 10, no. 6, pp. 732–739, 2000.

[2] S. Gao, W. Hou, T. Tanaka, and T. Shinozaki, “Spoken language acquisition based on reinforcement learning and word unit segmentation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6149–6153.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14236

[4] H. Kamper, K. Livescu, and S. Goldwater, “An embedded segmental k-means model for unsupervised segmentation and clustering of speech,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 719–726.

[5] S. Harnad, “The symbol grounding problem,” *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[6] C. Matuszek, “Grounded language learning: Where robotics and NLP meet,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 5687–5691. [Online]. Available: https://doi.org/10.24963/ijcai.2018/810

[7] A. Chauhan and L. S. Lopes, “Using spoken words to guide open-ended category formation,” *Cognitive Processing*, vol. 12, no. 4, p. 341, 2011.

[8] J. M. Siskind, “Grounding language in perception,” *Artificial Intelligence Review*, vol. 8, no. 5-6, pp. 371–391, 1994.

[9] ——, “Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic,” *Journal of artificial intelligence research*, vol. 15, pp. 31–90, 2001.

[10] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin *et al.*, “Grounded language learning in a simulated 3D world,” *arXiv preprint arXiv:1706.06551*, 2017.

[11] H. Yu, H. Zhang, and W. Xu, “Interactive grounded language acquisition and generalization in a 2D world,” in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1UOm4gA-

[12] A. Sinha, B. Akilesh, M. Sarkar, and B. Krishnamurthy, “Attention based natural language grounding by navigating virtual environment,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 236–244.

[13] G. A. Sigurdsson, J.-B. Alayrac, A. Nematzadeh, L. Smaira, M. Malinowski, J. Carreira, P. Blunsom, and A. Zisserman, “Visual grounding in video for unsupervised word translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] D. Roy, “Grounded spoken language acquisition: experiments in word learning,” *IEEE Transactions on Multimedia*, vol. 5, no. 2, pp. 197–209, 2003.

[15] D. Harwath, A. Torralba, and J. Glass, “Unsupervised learning of spoken language with visual context,” in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[16] D. Harwath and J. Glass, “Learning word-like units from joint audio-visual analysis,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 506–517. [Online]. Available: https://www.aclweb.org/anthology/P17-1047

[17] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 609–617.

[18] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[19] D. Dov, R. Talmon, and I. Cohen, “Sequential audio-visual correspondence with alternating diffusion kernels,” *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3100–3111, 2018.

[20] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, “Resource management with deep reinforcement learning,” in *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, 2016, pp. 50–56.

[21] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[22] Z. Zhou, X. Li, and R. N. Zare, “Optimizing chemical reactions with deep reinforcement learning,” *ACS central science*, vol. 3, no. 12, pp. 1337–1344, 2017.

[23] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 05 2013.

[24] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual Question Answering,” in *International Conference on Computer Vision (ICCV)*, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.