一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-057-13 (11/20)

言語獲得能力を備えた音声対話エージェントの検討

On Language Acquisition With Spoken Dialogue Agent

篠崎隆宏 * 高 聖洲 張 明鑫 侯 汶昕 田中 智宏

東京工業大学

Abstract: We propose an end-to-end neural network-based spoken language acquiring agent that combines unsupervised learning and reinforcement learning. The agent first learns sound words from unlabeled speech waveform and makes a sound dictionary. Then, the agent uses the sound dictionary as an action space during spoken dialogues. With the proposed method, reinforcement learning compensates the insufficient accuracy of unsupervised learning, while unsupervised learning significantly contributes to improve the efficiency of reinforcement learning. Simulation experiments demonstrate that the agent efficiently learns to speak appropriate word utterances based on the outside environment and its internal desire.

1 はじめに

生活環境において人と共存し人の生活をサポートするロボットが実現すれば，高齢化の進む将来社会において有用と期待される．ロボットが個別の環境下で柔軟な活動を行うためには，日々新し知識を継続的に取得するとともにそれを言語表現する高い学習能力が求められる．人にとって音声言語は一次的であり，文字言語を持たない言語はあっても音声言語を持たない言語は知られていない．また，道具を使用せずに身体のみを用いてパラ言語情報を伴った豊かな意思伝達を即座に行うことができるという，文字言葉にはない特徴がある．人の音声言語学習能力は強力で，特定の言語の知識を何も持たない状態で生まれた後社会生活を行う中で母国語を獲得することができる．しかし現在一般的な教師あり学習を用いた対話システムでは，音声対話を行う中で学習を行い言語知識を拡張するということができない．人と真に柔軟な音声対話を行えるロボットを実現するためには，人との日常会話の中で閉じた学習ループを形成する学習アルゴリズムの実現が不可欠である．

人がどのように音声言語を獲得しているのか未だ完全な理論や工学モデルは実現していないが，幾つかの試みは行われている．初期の研究として，スキナーは行動心理学の立場から人は単語に意味を結びつける強化理論に基づいて言語を学習しているとの仮説を提唱している [1]．近年では，ロボットに音声言語を自動獲得させることで構成的に音声言語獲得を理解すると共に工学的に応用しようとする研究が行われている．しかし，語彙獲得と意味理解および状況に応じた適切な発話発声の全てを同時に実現しているシステムは存在していない [2]．本研究では教師なし学習と強化学習により音声言語獲得におけるこれら全ての側面を同時に実現するニューラルネットワークシステムを提案し，評価実験を行う [3, 4] [1]．また，今後の課題について検討する．

*連絡先：東京工業大学工学院情報通信系
〒 226-8502 神奈川県横浜市緑区長津田町 4259-G2-2
www.ts.ip.titech.ac.jp

2 関連研究

Roy 等によるシステム [5] は，連続的に発話された音声から単語境界の推定を含めて語彙を学習することができる．また共起関係をもとに単語と画像オブジェクトの対応を学習することもできる．しかし事前に教師あり学習した音素認識器の利用が前提となっており，完全なゼロ知識からの学習にはなっていない．また単語の学習を目的としており，行動の学習は対象とされていない．事前学習された音素認識器としてであるが，ニューラルネットが HMM やヒストグラムとともにシステムの一部として用いられている．岩橋は，隠れマルコフモデル (HMM) と統計文脈自由文法 (SCFG) により構成されたシステムを提案している [6]．このシステムでは一単語ごと発話した単語の発声から語彙を獲得し，相互情報量に基づき共起関係を定式化することで単語と視覚を結びつける学習を行っている．ロボットへの要求発話の認識や行動も学習に基づいているが，認識結果を行動に結びつける方策はヒューリスティックなプログラムに依存している．杉浦等のシステムでは，単語学習結果をもとに確率推論結果の信頼度を学

[1] 提案法を実装したプログラムを `https://github.com/tttslab/spolacq.git` で公開している．

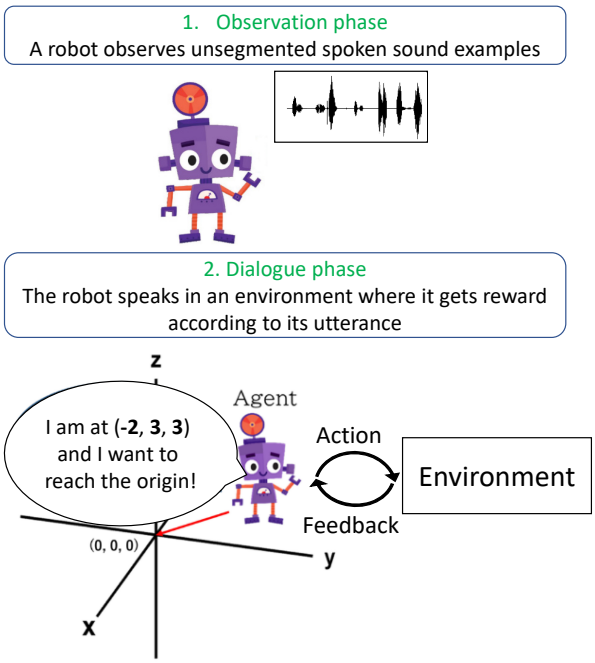

図 1: Spoken language acquisition system based on unsupervised word learning and reinforcement based dialogue leanring.

習して対話制御に用いている [7]．しかし，信頼度を用いた対話制御および確認発話は固定されている．谷口等は階層的なノンパラメトリックベイズ法と特徴量抽出にニューラルネットを組み合わせたシステムを提案している [8]．しかし学習済みの音素認識器が仮定されている点でゼロ知識からの学習ではなく，また発話方策の学習は含まれていない．Harwarth 等はニューラルネットを用いてラベル付けのされていない音声と画像から音声単語と対象物の関係を直接学習させる手法を提案している．しかし音声と画像のグランディングを対象としており，発話方策の学習は含まれていない．[9]．羽鳥等は，ロボットが強化学習による学習に基づき音声で指示された行動を実行するシステムを提案している [10]．このシステムでは行動が強化学習により学習されるが，音声入力は学習済みの音声認識器を用いてテキストに変換されており，音声処理は学習に含まれていない．また，音声の発話は対象とされていない．

3 提案方法

音声言語獲得機能を備えた音声対話エージェントは，原理的には感覚入力と音声合成器を備えたニューラルネットエージェントに強化学習を適用することで実現できるはずである．そこでは，音声発話が強化学習における行動に該当する．しかし音声発話は可変長の高次元連続データであるため，ランダムに初期化された方策関数が状況に応じた意味のある発話を生成する確率は限りなく 0 に近い．発話行動が成功しなければ強化学習による学習は進まない．そのため強化学習のみを用いて音声言語獲得をさせようとすると，学習の収束に非現実的な時間が必要になってしまう問題がある．同様の困難性は人間の子供が音声を学習する際にも存在するはずであるが，人間の場合は周囲が話す音声を観察して真似ることで対話における試行を効率化していると考えられる．そこで，周囲の音声やその他画像入力等を観察し教師なし学習を行う観察学習と，教師なし学習により得られた音声単語集合を行動空間として強化学習を行う対話学習を組み合わせて行う学習アルゴリズムを提案する．

エージェントの基本設計として，エージェントに内在する欲求を満たすように外部世界からの入力に応じて音声発話を行うシステムを想定する．エージェントの内部欲求をどのように設計するかは重要な点であるが，本研究では対象外である．本研究では，任意に設定された欲求をもとに音声言語獲得を行う一般性のある仕組みについて取り組む．

以下では，はじめに観察学習として教師なし単語学習のみを用いる基本システムについて説明し，ついで音声画像接地をもとにロボットの注意を視界中のオブジェクトに集中させることで単語学習効率を向上させた拡張システムについて説明する．

3.1 教師なし単語学習に基づく音声言語獲得システム

教師なし単語学習法として，ES-KMeans 法 [11] などが提案されている．現状単体で十分な精度の学習を行える手法は存在しないが，強化学習と組み合わせて使用する場合は行動空間を離散化し探索を大幅に削減する効果が期待できる．そこで，図 1 に示す教師なし単語学習と強化学習を組み合わせた自動音声言語獲得手法を提案する．単語の教師なし学習では間違った単語セグメントを排除するよりも必要語彙のすべての単語の正しいセグメントが含まれることを優先し，生成される音声辞書のサイズは十分に大きくとる．

対話学習では，エージェントが 3 次元空間内のランダムな初期位置に置かれた状況を想定する．エージェントは原点に到達したいという内部欲求を持っている．エージェントは自分で直接移動することはできないが，方向を示す音声コマンドを発声するとエージェントが乗っている乗り物がそちらの方向に一ステップ進むことを想定する．強化学習には，教師なし単語学習で得た音声辞書を行動空間とする Q 学習を用いる [3]．

3.2 音声画像接地に基づく注意機構を導入したシステム

提案するエージェントの構成を図 2 に示す．提案法は，観察フェーズにおいて音声辞書の作成とともに音声と画像の関係を教師なし学習しておき，対話フェーズにおいて強化学習を行う際にロボットの視界にある画像と関連の高い音声単語の確率が高くなるように行動価値関数を補正することを原理とする．確率の補正は，行動価値関数を実装するニューラルネットに入力画像との類似度を入力することにより行う．さらに，教師なし学習により得た初期の音声辞書にある誤りを音声対話学習時に上書きし精度を高めるための機構として Action filter を提案し導入した [4].

想定する音声対話タスクを図 3 に示す．観察学習フェーズでは，ロボットは食べ物の写真を見ながら音声を聞く．音声は 1 から数単語からなる食べ物の名前についての発話である．発話ラベルや発話に含まれる単語数は，ロボットには一切与えられない．対話フェーズではロボットは 2 種類の食べ物の写真を提示される．ロボットは自身の内部状態に依存して，食べ物に対する嗜好を持っている．提示されている食べ物のうち自分の欲する方の食べ物の名前を正しく発声できれば，その食べ物が与えられ報酬を得る．反対の食べ物や提示されていない食べ物の名前を発声した場合や，発声が食べ物の名前として認識されない場合，報酬は得られない．

4 3D 空間での原点回帰対話実験

4.1 実験条件

教師なし単語学習には，Google Speech Commands Dataset の音声を用いた．データセットの発話のうち方向を表す 6 種類の単語 (up, down, left, right, forward, backward) および空間移動とは関係のない 1 種類の単語 (marvin) のそれぞれについて 200 サンプル，合計 1400 サンプルを連結して一つの音声ファイルとした．教師なし単語学習法は ES-KMeans 法とともに，比較のために単にランダムに音声をセグメント化するランダム法も用いた．教師なし単語学習により得た，壊れたセグメントを含む音声単語辞書のサイズはおよそ 2000 である．対話タスクにおいてエージェントが発話した音声は，一般タスクの音声認識器である Google Speech-to-Text API [2] を用いて認識した．教師なし学習した単語辞書中で音声認識器により方向を表す単語に認識されたセグメントの割合は，ES-KMeans 法を用いた場合が 22.4%，ランダム法を用いた場合が 11.8%であった．対話フェーズにおいて，ランダムな位置に初期配置されたエージェントが原点にたどり着くまでが 1 エピソードである．強化学習は初期値を変えながら 100 回繰り返し，エピソード数に対する報酬の平均と分散を求めた．

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

4.2 実験結果

教師なし単語学習を用いた音声言語学習エージェントの学習の様子を図 4 に示す．ES-KMeans 法を用いた方が学習効率が高いが，ランダム法を用いた場合でも学習が進めばほぼ同じ対話精度が得られている．

5 注意機構を導入した食べ物要求対話実験

5.1 実験条件

画像データとして，20 種類の食べ物（ 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, white radish) の写真をそれぞれ 120 枚撮影した．各食べ物画像に対して，4 種類のテンプレート（e.g., “Apple,” “An apple,” “A red apple,” and “It’s an apple.”）をもとに音声合成器 [3]を用いて説明音声を作成した．合成音声には 20dB のガウス雑音を重畳させた．ロボットの食べ物に対する嗜好としては，ロボットの内部状態として RGB カラーをランダムにセットし，その色に近い食べ物を欲するものとした．

5.2 実験結果

音声画像接地による意識集中機構を導入した音声言語学習エージェントの学習の様子を図 6 に示す．語彙数が増えているため意識集中機構を導入しない場合は学習がほとんど進行しない一方で，意識集中機構を導入することで学習が大幅に効率的に進むことがわかる．さらに Action filter 機構を導入することで初期音声単語辞書にある誤りを修正し，高精度な音声対話が学習されることが分かる．また意識集中機構や Action Filter 機構を導入しない場合は，教師なし音声辞書学習に ES-KMeans を用いる方法はランダム法と比べて学習速度の向上に貢献していることが分かる．しかし，意識集中機構や Action Filter 機構を導入するとそれらの効果が大きく，音声辞書の作成方法の違いの差は殆どなくなる結果となった．

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

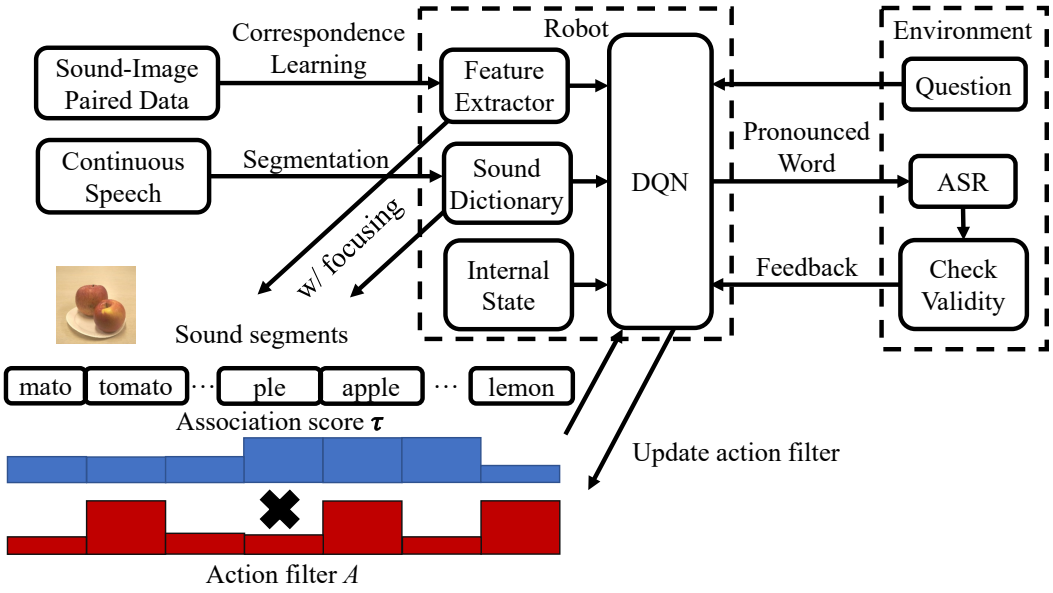

図 2: Proposed spoken language acquisition agent with sound-image grounding based focusing mechanism and the learning environment.

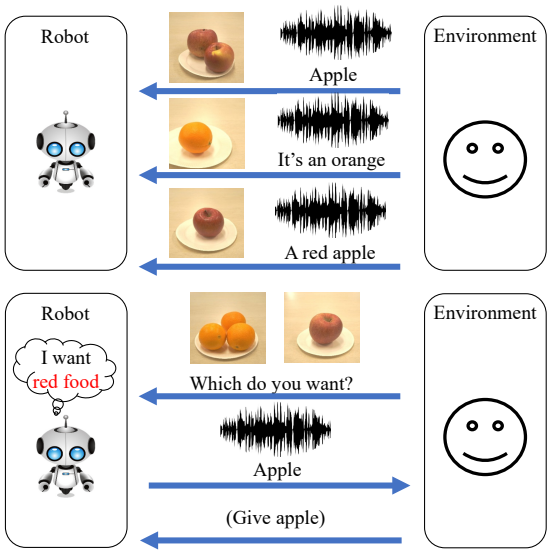

図 3: Spoken language acquisition task with image presentation.

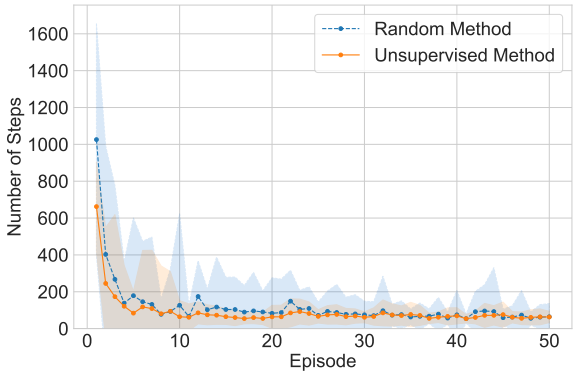

図 4: Learning process of the agents by the spoken dialogue with the environment that recognizes the speech commands.

6 エージェントの発話機構について (ディスカッション)

提案した２つの音声言語獲得エージェントの発話はどちらも，教師なし学習で得た音声辞書から選択した音声セグメントの再生に限られている．そのため，声質は観察学習の時に観察した音声の話者のままである．また感情その他を表現する抑揚の制御も音声辞書の要素の選択を通してしか行えず，自由度が低い．より柔軟な音声発話を行えるようにするためには，音声生成部を自由度の高い音声合成器に置き換える必要がある．一方強化学習を効率的に進める観点からは，エージェントのランダムな試行が音声発話として意味のあるものになる確率がある程度の大きさを持つ必要がある．そのためには，コンパクトな連続空間から音声発話を合成する機構が必要である．このために，今後の課題として音素ラベルによる条件付けを必要とせず音声を生成できる WaveGAN [12] を応用することなどが考えられる．また，人の場合は声帯や声道などから構成され

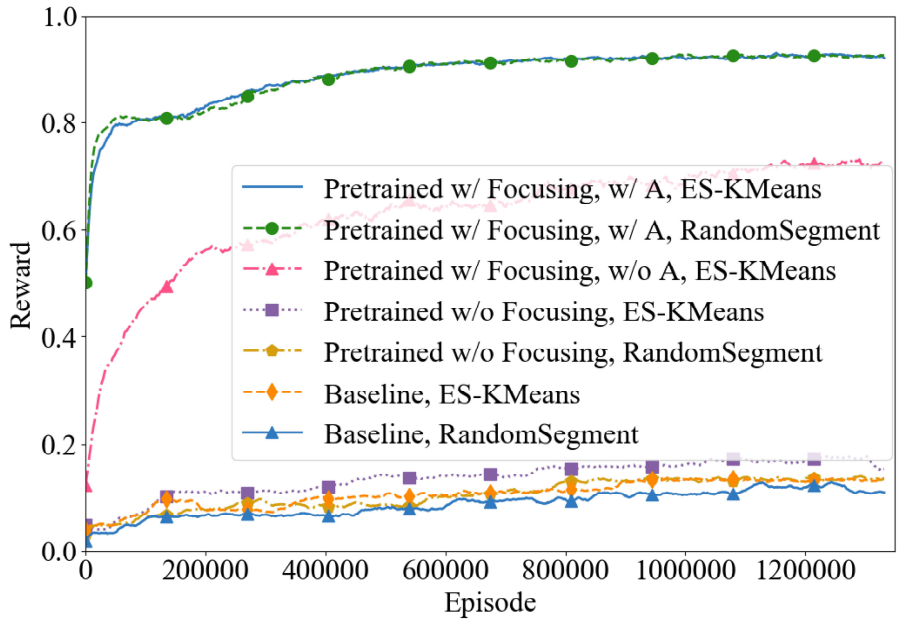

図 5: Learning performance of the language acquisition agent in the dialogue phase.

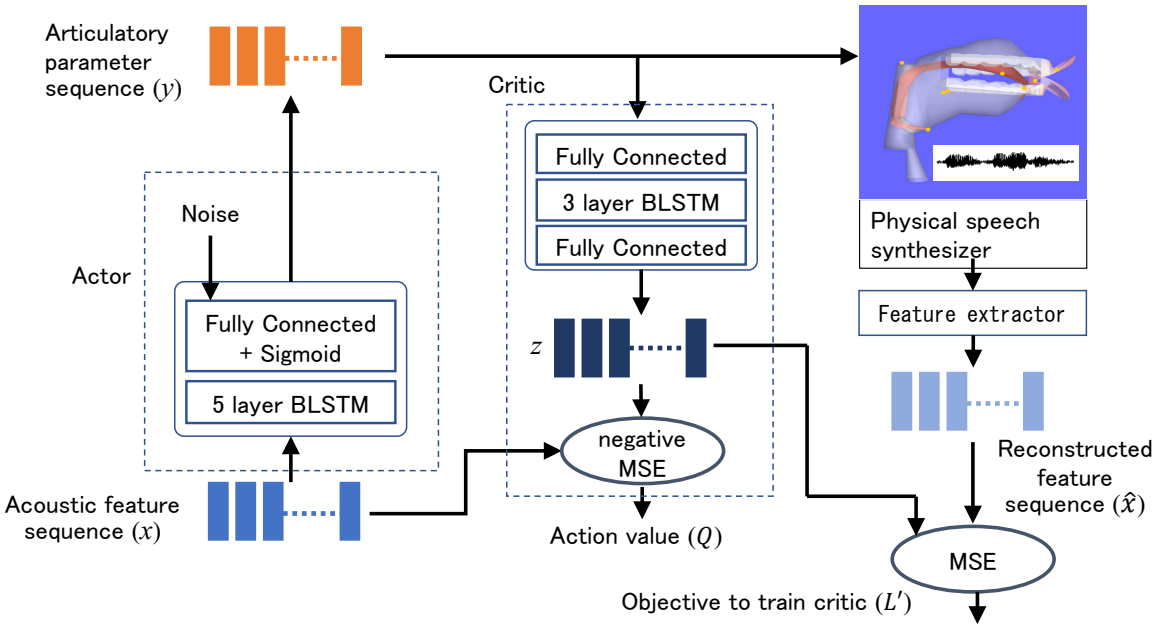

図 6: Hybrid autoencoder based self-learning of physical speech synthesis system using deterministic policy gradient.

る音声生成機構の限定された自由度が効率的な音声発話学習に貢献している可能性も考えられる．また発声器官の物理的な形状が，音声の話者性を生じさせる重要な因子となっている．音声言語獲得エージェントにおいても，人間の音声生成機構をそのまま物理的に模擬する物理音声合成器 [13] を事前知識の一形態として用いることも考えられる．制御器となるニューラルネットと物理音声合成器を接続してハイブリッドなオートエンコーダを構成し決定的方策勾配法を用いることで自己教師あり学習を実現することができ [14]，これを応用することが考えられる．

7　まとめ

教師なし単語学習と強化学習を用いた，音声言語獲得エージェントの仕組みを提案した．教師なし単語学習を用いて強化学習の探索空間を離散化することで，現実的な試行回数で学習が進むことを示した．辞書の精度は高いほうが望ましいが，音声辞書が必要単語の正しいセグメントを含んでいれば精度が低くても学習が進むことを示した．また，視覚に基づき音声辞書中の特定の単語に注意を向ける仕組みを提案した．ロボットが自分の望む食べ物を音声発話により得るタスクを設定したシミュレーション実験を行い，提案法の有効性を示した．今後の課題として，音声発話の柔軟性をより高めることが挙げられる．

謝辞

本研究は東レ科学振興会の助成を受けたものです．

参考文献

[1] B. Skinner, “Verbal behavior,” in *Verbal behavior*. New York: Appleton-Century-Crofts, 1957.

[2] T. Tangiuchi, D. Mochihashi, T. Nagai, S. Uchida, N. Inoue, I. Kobayashi, T. Nakamura, Y. Hagiwara, N. Iwahashi, and T. Inamura, “Survey on frontiers of language and robotics,” *Advanced Robotics*, vol. 33, no. 15-16, pp. 700–730, 2019. [Online]. Available: https://doi.org/10.1080/01691864.2019.1632223

[3] 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.

[4] M. Zhang, T. Tanaka, W. Hou, S. Gao, and T. Shinozaki, “Sound-image grounding based focusing mechanism for efficient automatic spoken language acquisition,” in *Proc. Interspeech*, 2020, pp. 1436–1440.

[5] D. ROY, “Learning words from sights and sounds : A computational model,” *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[6] N. IWAHASHI, “Language acquisition through a human-robot interface,” *Proc. Int. Conf. Spoken Language Processing*, 2000.

[7] 杉. 孔明, 岩. 直人, 柏. 秀紀, and 中. 哲, “言語獲得ロボットによる発話理解確率の推定に基づく物体操作対話,” **日本ロボット学会誌**, vol. 28, no. 8, pp. 978–988, 2010.

[8] T. Taniguchi, T. Nakamura, M. Suzuki, R. Kuniyasu, K. Hayashi, A. Taniguchi, T. Horii, and T. Nagai, “Neuro-serket: Development of integrative cognitive system through the composition of deep probabilistic generative models,” *New Generation Computing*, vol. 38, no. 1, Jan 2020.

[9] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, *Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input: 15th European Conference, Munich, Germany, September 8 – 14, 2018, Proceedings, Part VI*, 09 2018, pp. 659–677.

[10] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, “Interactively picking real-world objects with unconstrained spoken language instructions,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3774–3781.

[11] 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.

[12] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” in *Proc. ICLR*, 2019.

[13] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” *PLOS ONE*, vol. 8, no. 4, pp. 1–17, 04 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0060603

[14] H. Shibata, M. Zhang, and T. Shinozaki, “Unsupervised acoustic-to-articulatory inversion neural network learning based on deterministic policy gradient,” in *Proc. IEEE Spoken Language Technology*, 2020, accepted.