音声からの未登録語切り出しと画像からの物体抽出の統合による新規物体の学習

○杉浦孔明[†,††]，水谷了[†††]，中村友昭[†††]，
長井隆行[†††]，岩橋直人[†,††]，岡田浩之[††††]，大森隆司[††††]

[†](独) 情報通信研究機構　[††](株) 国際電気通信基礎技術研究所　[†††] 電気通信大学　[††††] 玉川大学

Learning Novel Objects from Audio-Visual Input Based on Out-of-Vocabulary Word Segmentation and Object Extraction

*Komei Sugiura[†,††]，Akira Mizutani[†††]，Tomoaki Nakamura[†††]，
Takayuki Nagai[†††]，Naoto Iwahashi[†,††]，Hiroyuki Okada[††††]，and Takashi Omori[††††]

[†]NICT　[††]ATR　[†††]The University of Electro-Communications　[††††]Tamagawa University

Abstract— This paper presents a method for learning novel objects from audio-visual input. Objects are learned using out-of-vocabulary word segmentation and object extraction. We conducted experiments in which a user taught the names of objects by uttering and showing them to a robot implemented with our method. The results reveals that our method obtains an accuracy of 88% for the integrated recognition accuracy of vision and speech.

Key Words: out-of-vocabulary, object learning, HMM, SIFT

1. はじめに

日常生活環境でロボットが人間と自然に対話・行動するためには，自己位置推定と移動，把持，音声処理，画像処理などを頑健に行なうソフトウェア・ハードウェアの統合が必要であり，興味深い問題が山積している．ここで日常生活支援ロボットの音声対話技術に着目すると，既存のトップダウン手法においては言語知識があらかじめに与えられていることが多い．

全ての言語知識を用意することは不可能であるため，トップダウン手法には未登録語を発声できないという制約がある．例えば，案内ロボットがユーザの顔画像を学習できたとしても，人名が登録されていなければ名前を呼ぶことはできない．一方，画像と音声からボトムアップに語彙を獲得するロボットの研究も行なわれている [2,3]．しかし，これらのボトムアップ手法には実用性の問題がある．

これに対し我々は，言語知識を利用しつつ，未登録語の発声が可能なハイブリッド手法を目指す．具体的なタスクとしては，新規物体の学習を扱う．我々の提案する手法では，未登録語の登録をテンプレート文で行ない，通常の対話はルールベースで行なう．新規物体の学習タスクの実験条件は，RoboCup@Home を参考にする．例えば，ユーザがロボットに物体を見せながら「これの名前は X」と発話することにより物体を学習させ，ロボットに再度物体を見せた場合に「これは X です」と発話させるようなタスクである (Fig. 4 参照)．

いま，新規物体に関する画像とテンプレート文で発話された音声が与えられたとする．このとき，新規物体の学習のために解くべき問題は以下の 4 つに分類できる．すなわち学習時には，1) 雑音下での頑健な音声認識，2) 画像から学習すべき物体の抽出，が必要であり，認識時には，3) 照明条件の変化に対して頑健な物体認識，4) 未登録語の発声，が必要である．これに対し我々は，ノイズの逐次推定と雑音抑圧，音声からの未登録語の切り出しと声質変換，動きアテンションに基づく物体抽出，scale-invariant feature transform (SIFT) 情報によるマッチング，を組み合せてこれらの問題を解決する．

本論文の構成は以下の通りである．まず 2 節では，提案手法を音声および画像処理にわけて説明する．3 節で実験に用いるロボットを概説した後，4 から 6 節においてそれぞれ，音声処理，画像処理，統合システムに関して行なった実験の結果を述べる．最後に 7 節で結論を述べる．

2. 提案手法

提案手法の概略図を Fig. 1 に示す．

2·1 音声処理

音声処理において，フロントエンド部および音声認識部では，(株) 国際電気通信基礎技術研究所 (ATR) において開発された hidden Markov model(HMM) に基づく音声認識システム ATRASR を用いる．

まず，パーティクルフィルタに基づく非定常ノイズの逐次推定と MMSE(Minimum Mean Square Error) 推定に基づくノイズ抑圧を行なう [1]．発話区間の切り出しにおいては，フレーム内のエネルギーに基づき，endpoint detection(EPD) を行なう．

音声認識 (ASR) で用いる音響モデルのうち，「clean AMs」はクリーン音声のみで学習されたモデル (男声・女声)，「clean & noisy AMs」はクリーン音声に雑音を

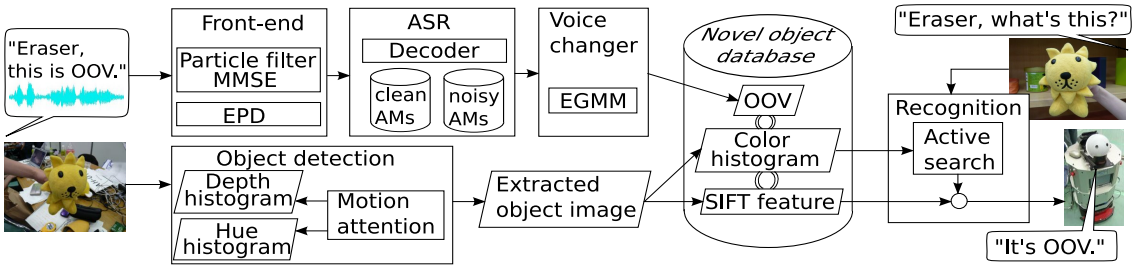

Fig.1 提案手法の概要

重畳した音声により学習されたモデル (男声・女声) である．これにより，雑音下であっても頑健な音声認識を行なうことが可能である．これにより前節で述べた問題 1) を解決できる．

本手法では，未登録語の登録は「これの名前は X」など決められた定型文で行なうものとする．音声認識の結果，入力音声の音素アラインメント情報が得られるので，音声から未登録語部分 (Fig. 1 の「OOV」) を切り出す．

ここで切り出された未登録語は，ユーザの声による音声である．そのため，ロボットに「これは X です」のような音声を出力させる場合には，そのままでは X の部分のみがユーザの声になり不自然である．そこで，切り出された未登録語音声を合成音声の声に変換してデータベースに登録する．不特定のユーザの声を特定の声に変換するために，eigenvoice Gaussian mixture model(EGMM) に基づく声質変換 [4] を行なう．つまり，ユーザからの入力音声の切り出しと声質変換を合わせて問題 4) を解決する．

2·2 画像処理

1 節で述べたように，画像処理の観点からは 2 つの問題がある．問題 2) は画像のセグメンテーションや物体抽出の問題である．本手法では，ユーザがロボットに物体を教示する場面を想定しているため，動きのあるひとかたまりの物体に注意を向けることで物体を抽出する動きアテンションに基づく物体抽出手法をベースとする [5]．これは，画像中の動きを検出し，その動きのある領域の色や奥行きの情報を基に最終的な物体領域を推定するものであり，ステレオ視差画像の計算を含めてもフレームレートに近い速度で動作する．

物体認識の際にも，シーン中のどこに認識すべき物体があるかを抽出する必要がある．但しこの際は必ずしも人が物体を持っている保証がないため，動きに注意を向けた抽出手法を用いることができない．そこで認識時の領域抽出には，色ヒストグラムと奥行き情報を併用した高速なアクティブ探索による領域抽出手法を用いる．

認識時には，SIFT を用いた局所特徴のマッチングを利用する．この際，色情報を用いて候補を絞った上で，学習時に様々な方向から見た物体の SIFT 情報とのマッチングを行い最終的な認識を行う．これにより問題 3) を解決する．認識結果は，前述の手法で得られた未登録語により発声される．

3. ハードウェア

Fig. 4 に実験に用いるロボット「eR@ser」を示す．ロボット上部には，マイクおよびカメラを搭載している．音声入力用に，三研マイクロホン製のショットガンマイクロホン CS-3e を用い，音声出力用に YAMAHA 製 NX-U10 スピーカ (20W) を用いる．また，アプライド・ビジョン・システムズ製ステレオビジョンカメラから得られた画像の学習を提案手法により行なう．

また，今回は実験に用いないものの，eR@ser はリビングルームの中でオブジェクトや人を探索することが可能である．音声・画像処理用計算機を搭載して移動するために，MobileRobots 製 PIONEER P3-DX をベースに用いている．また，SICK 製レーザレンジファインダから得られる情報に基づき，環境地図を構築する．

4. 実験 (1): 音声からの未登録語の切り出し

本実験の目的は，1) 区間検出精度の評価，2) 未登録語切り出しの誤差の評価，の 2 つである．

4·1 実験条件

まず，提案手法を評価するためのデータベースを構築した．雑音および発声変形の影響を調べるため，ロボットを用いる環境と同等のノイズ環境を再現して収録を行なった．ノイズは展示会場において収録したものを用い，ノイズレベルは 60dBA, 70dBA, 80dBA とした．ノイズレベルはロボカップ大会時の会場における雑音を参考に決定した．被験者とマイクまでの距離は 30cm とし，マイク周辺において全方向からの雑音がほぼ一定の大きさになるように調整する．

被験者は，8 名 (20 代から 40 代, 男女各 4 名) である．1 回の収録では，被験者に一定のノイズ環境の下，2 秒間隔で 8 文を発話を行なわせた．このとき被験者は 2 秒間隔で「イレイサー，これの名前は X」と発話する．X は以下の 8 単語のいずれかである．

- スリッパ，ライオン，虎のぬいぐるみ，ペンたて，アルバム，ウェットティッシュ，お茶，ごみ箱

雑音モデルの学習のために，各ノイズレベルにおける最初の発声前に，20sec の無発話区間を設けた．

音声データを 16kHz，16bit でディジタル化し，各フレームごとに，25 次元の特徴量ベクトルを計算する．特徴量として，12 次元のメル周波数ケプストラム係数 (MFCC)，Δ MFCC，対数パワーを用いた．フレーム長は 20msec, シフト長は 10msec とした．認識用の文法として，以下の 2 つを用意した．X の部分は音節の自由遷移である．

- 「イレイサー，これの名前は X」
- 「イレイサー，これは何ですか」

4·2 評価方法

以上により得られたデータベースに対して，人手で発話区間を指定した場合と EPD により発話区間を検

Table 1 発話単位の認識精度 [%]

	clean AMs	clean & noisy AMs
manual	99.5	99.5
EPD	82.8	83.3

出した場合の比較評価に用いる手法について述べる．

音声区間の検出性能を調査するために，各発話の検出精度を調査する．ここで，ロボットの音声モジュールとして区間検出を用いることを前提とすると，検出精度としては音声認識に必要な部分が検出されていることを示す指標であることが望ましい．そのため，検出精度として以下に定義される発話単位の認識精度 (Acc') を用いる．すなわち Acc' = (正解発話検出数/全発話数) である．ただし，区間検出により切り出された発話区間が正しく認識された場合に正解とした．また，誤検出など複数の区間を検出した場合は，正しい認識結果が得られる区間が 1 つだけある場合に正解であるとした．

未登録語切り出し精度は，人手でラベリングした未登録語開始時刻と推定値の平均絶対誤差 (MAE) により評価する．ただし，発話区間が正しく認識された場合のみ比較を行なう．用いた文では，未登録語開始時刻から文末までを切り出すことにより，未登録語部分の音声を得ることができる．

4·3 実験結果

Table 1 に発話単位の認識精度を示す．表において，manual は人手による区間指定，EPD は音声区間検出による認識精度を示す．また，通常の音響モデルのみを用いた場合を「clean AMs」とし，ノイズを含む音響モデルを加えた場合を「clean & noisy AMs」とする．表より manual 条件では，ほぼ 100%の認識率が得られていることがわかる．本タスクは，未登録語を含む文と既登録語のみの文からなる小語彙音声認識タスクであるので，妥当な結果が得られたと考えられる．

しかしながら EPD 条件では，Acc' は 83%程度であった．Acc' で不正解となった原因は，発話全体や途中部分の誤棄却など，区間検出によるものがほとんどであった．つまり，不正解は音声認識誤りに起因するのではなく，区間検出誤りに起因している．これは，人手で音声を区間を指定した場合に，ほぼ 100%の認識率が得られることからもわかる．

未登録語切り出しの定性的な結果として，Fig. 2 に，各雑音レベルにおける音声波形 (「これの名前は『虎のぬいぐるみ』」) および雑音抑圧後の音声波形の例を示す．(a)-(c) はノイズ抑圧前の音声波形を表し，(d)-(f) はノイズ抑圧後の音声波形を表す．また，(d)-(f) において×は区間検出の開始・終了時刻を表し，○は未登録語の開始位置の推定値を表す．Fig. 2 より，音声区間が正しく検出されていることがわかる．

次に，未登録語切り出しの精度を定量的に示す．Fig. 3 は，未登録語開始時刻の正解と推定値の MAE を示したものである．図においてエラーバーは標準偏差を表す．(a)(b) は人手による音声区間の切り出しに対応し，(c)(d) は EPD による切り出しに対応する．ただし，正しく認識された発話に対して誤差を示しているので，(a)(b) と (c)(d) では母数が異なる．Fig. 3 より，60-80dBA の環境において，MAE は 20-30msec 程度であることがわかる．実験に用いた文の未登録語部分は，

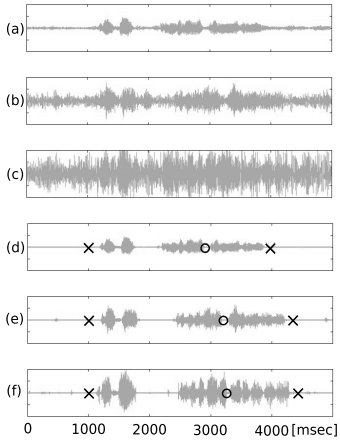

Fig.2 ノイズ抑圧前と抑圧後の音声波形．(a) 60dBA，(b) 70dBA，(c) 80dBA，(d) 60dBA ノイズ抑圧後，(e) 70dBA ノイズ抑圧後，(f) 80dBA ノイズ抑圧後

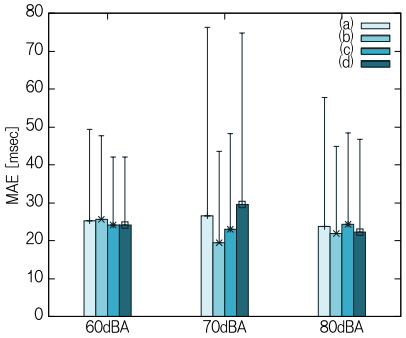

Fig.3 未登録語切り出し開始位置の平均絶対誤差．(a) clean AMs (manual)，(b) clean & noisy AMs (manual)，(c) clean AMs (EPD)，(d) clean & noisy AMs (EPD)

Fig.4 上段左:実験環境．上段右:物体教示の様子．下段:実験に使用した物体．右:実験に使用したロボット

平均 670msec 程度であるので，実用上十分な精度が得られているといえる．

5. 実験 (2): 画像からの物体学習

5·1 物体抽出の評価実験

物体抽出精度を評価するために行なった実験について述べる．実験では，被験者 8 名に対し 12 個の物体をロボットに見せて教示するよう指示した．環境は Fig. 4 のような一般的なリビングルームであり，使用した物体は Fig. 4 に示すような，ぬいぐるみや本，ペットボトルといった一般的なものである．

ただし被験者は，入力画像や抽出結果などをモニタを通して確認することができないものとするが，ロボットが動きを手がかりとして物体に注目することは予め口頭で説明することとした．

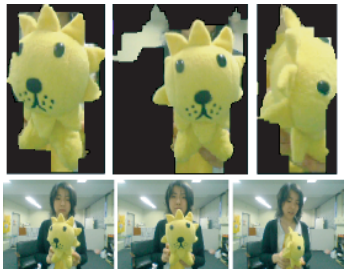

Fig.5 物体切り出しの例

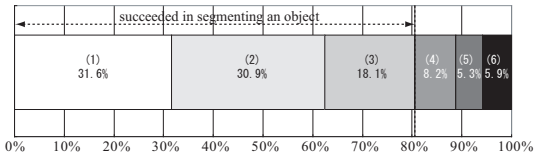

Fig.6 物体抽出実験の結果

ロボットは被験者が物体を見せ始めてから 100 フレーム分の画像を取得し，その間注目物体の切り出しを行う．各フレームにおける切り出し結果を，次の 6 つに分類することで評価した．

1. 90%以上の領域が抽出されている
2. 物体以外の領域を若干含んでいる
3. 物体の一部が欠けている
4. 物体以外の領域を大きく含んでいる
5. 物体領域が大きく欠けている
6. 物体の領域とは異なるところが抽出されている

1〜3 は認識に問題ないので切り出し成功とし，4，5 は切り出し失敗とする．Fig. 5 に，実際の切り出し例を示す．図の上段は，左から抽出成功，物体以外の領域を若干含む場合，物体の一部が若干欠けた場合である．また，下段はそれぞれの入力画像を示している． 実験結果を Fig. 6 に示す．図は全ての被験者および物体の結果を合計したものであり，(1)〜(6) の数字は上記切り出し結果の分類 1〜6 に相当している．つまり，全体として物体抽出に成功した割合は 80.6%である．被験者による抽出精度のばらつきはそれほど大きくなかったが，着ている服による違いが若干見られた．一方，物体によるばらつきは大きく，特に表面に光沢がある物体や色味の少ない物体は，背景が大きく含まれてしまう場合や，物体が大きく欠けて抽出される場合があった．

5·2 物体認識実験

次に，学習した物体の認識精度を評価する実験を行なった．環境は前節と同様であるが，物体数を 25 個に増やし (Fig. 4 下段)，そのうち 20 個を学習させ，5 個を未学習とした．従って，未学習の 5 個に関しては，不明な物体と答えるのが正解となる．学習は，リビングルームのある一箇所で行うが，認識は 4 箇所の異なる場所で行うこととした．被験者 (1 名) は，物体の学習 (切り出し) が正しく行われたことを，ロボットに搭載したモニタで確認した上で認識を行わせた．

認識率の平均は 90%であった．結果の内訳を Table 2 に示す．場所 2 は，学習を行った場所と照明環境が大きく変化しており，未知物体を既知の物体と誤ることが多く起こった．

6. 実験 (3): 統合システムの評価実験

音声・画像処理を統合したシステムをロボットに実装し，リビングルーム環境において実験を行なった．実

Table 2 物体認識実験の結果

	場所 1	場所 2	場所 3	場所 4
正解/学習物体数	19/20	18/20	19/20	18/20
正解/未知物体数	5/5	2/5	4/5	5/5
認識率	96%	80%	92%	92%

験の目的は,「ある新規物体に対し，学習フェーズでは未登録語の切り出しおよび物体抽出が成功し，かつ認識フェーズでは画像の認識が成功する確率 (統合精度)」を評価することである．学習フェーズでは，ユーザが物体を見せながら未登録語 X を発声した後，ロボットが X を返答すれば成功とする．認識フェーズで，ユーザが物体を見せながら音声で名前を質問した後，ロボットが X を返答すれば認識フェーズ成功とする．登録物体 N 個全てに対し学習フェーズを行なった後，認識フェーズを N 個に対し行なう．

実験環境は前節と同じものを用い，物体は Fig. 4 下段の物体から 10 個を選択した ($N = 10$)．本実験では，ロボットに搭載したモニタより区間検出と学習の経過情報をユーザにリアルタイムでフィードバックしている．これは，発話中に区間検出誤りに関する情報をユーザに与えることで，区間検出精度を改善できるためである．学習フェーズでユーザの言い直しがなく成功した場合と，未登録語切り出しの失敗のためにユーザが言い直した場合について統合精度を調べた．

実験の結果，統合精度は 88%であった．学習フェーズにおいて言い直しを許した場合，統合精度は 94%であった．

7. おわりに

本論文では，日常生活環境において新規物体を学習する手法を提案した．提案手法では，新規物体を学習・認識するために，1) ノイズの逐次推定と雑音抑圧，2) 音声からの未登録語の切り出し，3) 動きアテンションに基づく物体抽出，4)SIFT 情報によるマッチング，を組み合せている．本研究で構築したロボットは，外部の計算機を用いずに，一連の処理を実時間で行なうことができるという特徴を持つ．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金 (基盤研究 (C) 課題番号 20500186) による研究助成を受けて実施されたものである．

参考文献

[1] Fujimoto, M. et al.: Sequential Non-Stationary Noise Tracking Using Particle Filtering with Switching Dynamical System, *Proc. of ICASSP 2006*, pp. 769–772 (2006).

[2] Iwahashi, N.: Robots That Learn Language: Developmental Approach to Human-Machine Conversations, *Human-Robot Interaction* (Sanker, N. et al.(eds.)), I-Tech Education and Publishing, pp. 95–118 (2007).

[3] Roy, D.: Grounding Words in Perception and Action: computational insights, *Trends in Cognitive Science*, Vol. 9, No. 8, pp. 389–396 (2005).

[4] Toda, T. et al.: One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices, *Proc. of ICASSP 2007*, pp. 1249–1252 (2007).

[5] 中里, 長井, 樽松: 動きアテンションによる物体の抽出とオンライン教師なし学習による物体認識, 信学技報, パターン認識・メディア理解研究会 PRMU2003-274, pp. 109–114 (2004).