モジュラーホームロボットによる環境音を利用した家庭内作業のためのROARにおける環境音学習過程の自動化

○五十嵐大騎（都市大）　佐藤大祐（都市大）　金宮好和（都市大）

1. 緒言

ホームロボットが家庭内で作業を行うためには，周囲の環境状況を把握し，状況に応じた動作を実行する必要がある．家庭内環境には，家電製品の操作音や動作音など，さまざまな環境音が存在しており，これらを利用することによって環境状況を把握することは，ホームロボットにとって非常に有用であると考えられる．

我々は，機能ごとにモジュール化されたホームロボットによる家庭内作業の実現を目的に研究開発を行ってきており，ロボット用オープンソース・ミドルウェアである Robot Operating System（ROS）上で動作する ROS Open-source Audio Recognizer（ROAR）[1] を利用した環境音認識モジュールを追加し，環境音認識による作業動作生成システムを構築した [2]．そして，このシステムを利用した家庭内作業の動作計画を行い，家庭内作業の一つとして朝食準備作業を実現した [3]．しかし，ROAR にはホームロボットの動作時にマイクロフォンによる集音を行った際，環境音の録音データが変化し，その認識性能が低下することが考慮できない問題や，環境音を認識するためには人間が介在しなければならない問題など，ホームロボットに利用するための技術的課題点が複数存在する．

そこで本稿では，ROAR を利用した環境音認識において，ホームロボットに学習対象となる環境音を与えた後に必要となる環境音の学習過程を自動化する．そして，環境音認識実験により改善手法の有用性を示し，自動化による作業効率の向上を実現する．

2. 環境音認識システムとその問題点

2.1 ROAR を利用した環境音認識モジュール

ROAR には音の学習と認識の二つの機能がある．学習段階では，マイクロフォンを用いてユーザが取得した録音データを図 1 のようなスペクトログラム（時間，周波数，信号成分の強さの 3 次元グラフ）に変換し，これをユーザが目視で確認し，マウスを用いて学習させる環境音の領域を選択する．これらの作業を複数回繰り返し，環境音の特徴量を得る．機械学習の一つである One-Class Support Vector Machine（OCSVM）を利用して，取得した複数回分の環境音の特徴量から超球を生成し，学習結果として環境音のテンプレートを得る．認識段階では，学習によって得られたテンプレートを用い，マイクロフォンで集音される録音データの中から機械的に音の特徴量を抽出し，その特徴量が OCSVM によって得られた超球内に収まっているかを判別することによって，学習した環境音を認識している．

この ROAR を，開発しているホームロボットの環境音認識モジュールとして利用するために，図 2 のようなロボット手先に備えられたセンサモジュール内に

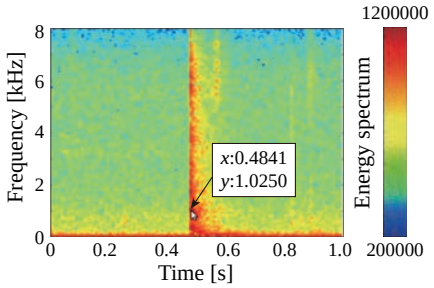

図 1　集音データのスペクトログラム上での環境音の選択

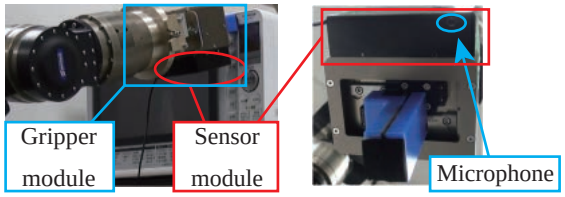

図 2　マイクロフォンを搭載したグリッパモジュール

Microsoft 社製 LifeCam HD-5000 を実装し，これに内蔵されたマイクロフォンで音を取得する．マイクロフォンがホームロボットのグリッパ近くに備わっているため，ホームロボットが操作する作業対象の周辺で発生する環境音の取得が容易になる．

2.2 環境音認識システムの問題点

ROAR を利用した環境音認識システムの大きな問題点は，学習過程においてユーザが介在する必要があることである．一つのテンプレートを作成するために，スペクトログラムから音の特徴量を選択する作業を高い学習結果が得られるまで要求されるため，操作に慣れたユーザであっても 1 音源に対して 1 時間弱の時間がかかり，非常に効率が悪い．

また ROAR は，集音時の周囲環境の変化や音源とセンサとの相対位置姿勢の変化を考慮した録音データ取得が可能なソフトウェアではないため，ホームロボットが作業中に環境音認識を高い精度で実現するためには，利用方法や学習過程の改善が必要である．

そこで本稿では，従来の ROAR では手動で行っていた学習対象となる環境音の選択を自動化し，学習結果である環境音のテンプレート作成にかかる時間を大幅に削減する．

3. 環境音学習過程の自動化

3.1 最大エネルギースペクトルに基づくテンプレートの作成

先に述べたとおり，ROAR において学習させる環境音の特徴量の決定は，ユーザが特徴量をスペクトログ

ラム上で視覚的に判断することで行う．よって，これをロボットに判断させるためには，その基準を数値的に決定する必要がある．そこで，環境音のデータを解析して特徴量となる要素を取り出すアルゴリズムを考案し，認識したい環境音の学習が自動的に進み，テンプレートが作成できるようシステムを変更する．

図 1 より，特徴量の抽出に必要となるパラメータは，音が発生した時間，音の周波数，そしてエネルギースペクトルの三つである．よって，環境音の録音データを時間領域と周波数領域で解析し，エネルギースペクトルが最大となる要素を特徴量として抽出することによって，テンプレート作成までの学習過程を自動化する．

3.2 特徴量の定義と抽出方法

音をマイクロフォンによって集音すると，時間領域における波形データが出力される．振幅を音波の特徴が表れるパラメータの一つとし，振幅が最大値となる時間を特徴量の時間成分として定義する．

また，周波数領域でのスペクトル解析には一般的に Fast Fourier Transform（FFT）が用いられる．ここで，FFT の理論は次式で表される．

$$
\begin{aligned}
F(s) &= \sum_{x=0}^{N-1} f(x)\exp(-j\frac{2\pi x}{N}s) \\
&= \sum_{x=0}^{N-1} f(x)\cos\frac{2\pi x}{N}s - j\sum_{x=0}^{N-1} f(x)\sin\frac{2\pi x}{N}s \\
&= Re - jIm \qquad (1)
\end{aligned}
$$

ただし，N がサンプリング数，x が標本点，$f(x)$ が振幅，s が周波数，$F(s)$ が周波数スペクトルである．また，エネルギースペクトルの理論式は以下の通りである．

$$E = F(s)^2 = |Re - jIm|^2 = Re^2 + Im^2 \qquad (2)$$

このことから，波形データを FFT にかけることで，エネルギースペクトルが得られる．エネルギースペクトルは，音の信号成分の強さを表しており，これを特徴を表す指標の一つと考え，エネルギースペクトルが最大となる周波数を特徴量の周波数成分として定義する．

3.3 FFT による特徴量抽出とその問題点

環境音の一例として，電子レンジの操作音が挙げられる．電子レンジの操作音は，スイッチを押したときの機械音と電子レンジの操作を確認するための電子音によって構成される．この音の波形データを FFT にかけると，図 3 に示す結果となる．時間領域における振幅グラフを上図に示し，周波数領域におけるエネルギースペクトルのグラフを下図に示す． 上図の波形が単発的に発生する箇所が，機械音によるものであり，波形が集中して発生する領域が，電子音によるものである．これより，時間領域では機械音が振幅最大であることから，機械音を特徴量として捉えている．一方で，下図のエネルギースペクトルにおいて，機械音のスペクトルは全体的に分散しており，電子音のスペクトルは 2000 Hz 付近で単発的に現れている．これより，周波数領域では電子音がエネルギースペクトル最大であることから，電子音を特徴量として捉えている．

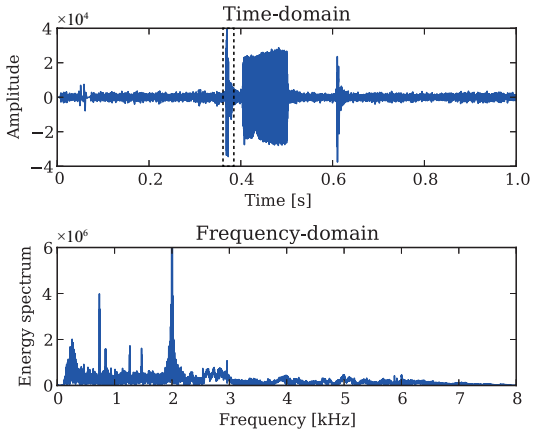

図 3　FFT による電子レンジの操作音の特徴量抽出結果

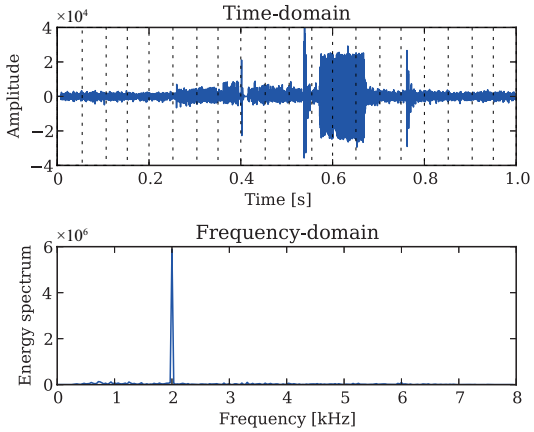

図 4　STFT による電子レンジの操作音の特徴量抽出結果

以上の結果から，時間領域と周波数領域で結果が異なるため，正確な特徴量抽出が行われない問題が生じる．電子レンジの操作音には，機械音や電子音が混在することが想定される．そのため，音源分離をせずに特徴量抽出を行うと，お互いの特徴量が競合し，正確な抽出ができない場合がある．

3.4 特徴量抽出の問題点解決

ここで，FFT 以外の解析方法として Short–Time Fourier Transform（STFT）を採用する．STFT は音声など時間変化する信号の解析に用いられ，音の波形データをサンプリング時間で細分化し，各サンプリングデータを FFT にかける方法である．この方法では，時間領域における特徴量を複数得られるため，電子音と衝撃音を分割した解析が可能になる．このことから，時間領域の対応が可能になると考えられる．

電子レンジの操作音を STFT にかけると，図 4 に示す結果となる．上の時間領域グラフのように時間軸をサンプリング時間で細分化して，それぞれ FFT をかけてエネルギースペクトルの最大値を比較する．その値が最も大きいときの周波数スペクトルが下の周波数領域グラフとなっている．

ここで，STFT でエネルギースペクトルが最大となる周波数は，2000.00 Hz であった．一方で，電子レン

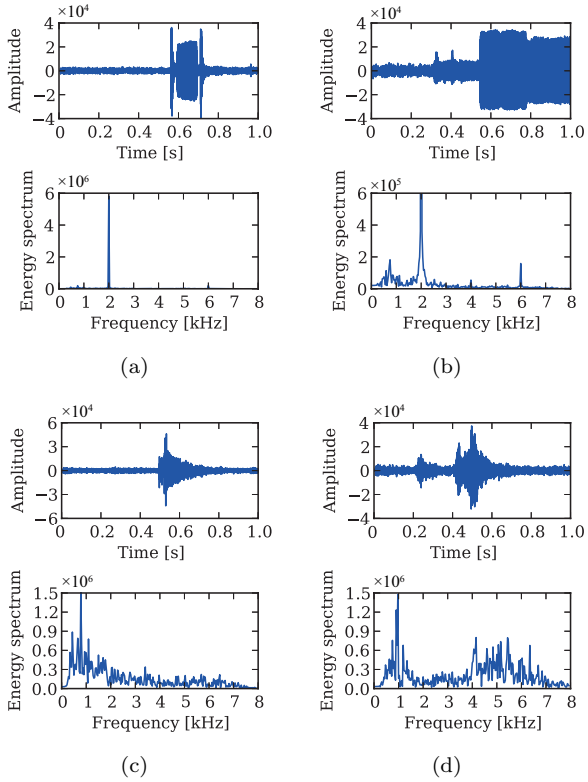

図 5 STFT による環境音の解析結果: (a) 電子レンジの操作音 (b) 電子レンジの終了音 (c) 電子レンジの扉の開閉音 (d) トースターの終了音

ジの操作音の周波数は 2000 Hz である．このことから，周波数領域において STFT は正確に特徴量抽出が行われているため，手動によるスペクトログラムの選択に代わるアルゴリズムとして利用が可能となる．

3.5 環境音の特徴量抽出実験

自動化した学習システムを使って，家庭内作業中に発生する環境音を集音し，STFT による特徴量抽出が可能か検証実験を行う．解析結果を図 5 に示す．上図が時間領域における環境音の波形グラフであり，下図が周波数領域における周波数スペクトルのグラフである．これより，電子レンジの操作音や終了音といった単発の電子音は，エネルギースペクトルが明確に現れているため，このシステムによる特徴量抽出が有効であると考えられる．一方で，開閉音などの衝撃音に属する音はエネルギースペクトルが分散しているため，エネルギースペクトルの最大値のみを特徴量として定義すると，音としての特徴が欠落する問題が生じる．

以上のことから，このシステムを利用することで電子音の特徴量抽出が可能になり，ロボットに学習させることができる．しかし，衝撃音のテンプレート生成は技術的課題が存在する．そのため，特徴量の抽出方法に更なる改良の必要がある．

4. 改善したシステムの有用性

4.1 作業効率の向上

環境音モジュールを用いて従来手法と改善手法で環境音のテンプレートを作成し，作成にかかる時間を測定する．ここでは，学習対象の環境音を電子レンジの操作音として，50 回の学習を行った．

以前までででのテンプレート作成時間は，40 分であった．一方で，改善手法によるテンプレート作成時間は，5 分である．これより，13 %程度まで作業時間を減らすことができたため，改善手法によって作業効率が向上したと言える．また，短時間で環境音を学習させることが可能になることから，より多くの環境音のテンプレートが簡単に作成できるようになる．これによって，ロボットの動作に応じた環境音のテンプレートを複数保有することがが可能になるため，様々な環境状況に対応ができるようになると考えられる．

表 1　学習過程の自動化前後での環境音認識率の比較

Type of sound	Recognition Rate [%]	
	Conventional approach	Improvement approach
Operation	92	96
Finish	78	76

4.2 改善手法による環境音の認識率の変化

従来手法と改善手法で，認識率にどう影響するかについて検証を行う．電子レンジの操作音と終了音のテンプレートを，二つの手法でそれぞれ作成し，同じ環境音を 50 回音を鳴らしたときの認識率を測定する．

表 1 に作成時間の結果を示す．このように，学習システムを改善したことによる認識率は変化が見られないが，従来と同様な精度の環境音のテンプレート作成が可能となる．よって，従来と同様なテンプレートが素早く得ることができるため，改善方法によって作業効率の向上が実現できる．

5. 結言

従来の環境音認識システムの問題点を挙げ，ロボットが自動で学習と認識ができるようシステムの改善を行った．その結果，電子音に属する環境音の学習には効果があることを示し，環境音の学習のための作業効率の向上を実現した．

今後の課題として，どのタイミングに鳴った環境音でも対応できるように，学習システムをさらに改善する．また，家庭内作業時に想定されるあらゆる環境音を学習させるため，特徴量抽出のパターンを増やす．

参　考　文　献

[1] J. M. Romano, J. P. Brindza and K. J. Kuchenbecker: “ROS open–source audio recognizer: ROAR environmental sound detection tools for robot programming,” Autonomous Robots, vol. 35, no. 3, pp. 207–215, 2013.

[2] 夏目彬弘, 関戸佐知, 佐藤大祐, 金宮好和：“環境音認識を用いたモジュラーホームロボットによる朝食準備作業”, 日本機械学会ロボティクスメカトロニクス講演会’14 講演論文集, 2014.

[3] 関戸佐知, 黒山佑太, 新良貴陽平, 佐藤大祐, 金宮好和：“モジュラーホームロボットによる環境音を利用した動作計画に基づく朝食準備作業”, 第 32 回日本ロボット学会 学術講演会, 2014.