平成三十年度　特別研究報告書

AI スピーカーを利用した

状況推測システムの提案

龍谷大学　理工学部　情報メディア学科

T150477　北迫祐樹

指導教員　三好　力　教授

内容梗概

各家庭に音声 AI を備えた機器が増えていく中で,音 AI が取得する音をつかいユーザーの状況を推測し状況に合わせたサービスの提供が行われるシステムの提案を行う.

本システムではユーザーの命令より前の環境音を分析し,特徴的な環境音をシンボル音として分類しておく.シンボル音の種類と発生した時間とを記録し収集する.収集したデータを用いることで,特定のシンボル音の並びが類似した時間間隔で発生する度に特定の命令が行われているといったパターンを解析することが可能になると考えた.

シンボル音に関する実験として何をシンボル音とすることが出来るか,異なるシンボル音同士の識別が可能であるか検証を行うために 2 種類の実験を行った.

実験 1 は録音した日常の環境音の中からシンボル音の候補となる特徴音を調べた.

実験 2 は 3 つの特徴音を選び,それぞれ 120 個ずつ録音したデータをスペクトログラム画像に変換し CNN モデルを構築し学習を行った.学習の済んだモデルで未確認データの識別を行いその予測精度から識別が可能か検証した.

多数のシンボル音の候補を見つけることが出来た上に平均 97.8%と高い精度でシンボル音同士の識別が可能であることがわかった.

目次

第1章 緒言

1.1 研究背景

音声認識技術や対話技術などの音声AIを備えたオーディオスピーカー（AIスピーカー）が次々と登場している.2017年6月末までに,先行した米Amazon.com社の「Amazon Echo」や米Google社の「Google Home」などが米国だけで累計4000万台近く売れたという推定もある.日本にも普及が始まっており,日本のメーカーでも製品の販売を開始しつつある.こうして,各メーカーがAIスピーカーに注目している理由には,ホームIoT機器市場という大きな市場を開拓するための布石のひとつだとみられている.現状ではユーザーとホームIoT機器をつなぐ役割はAIスピーカーが担っているが,ホームIoT機器の普及が進むにつれIoT機器に直接マイクを2つ付けることで音声AIを機能として備えることも考えられている.［1］［2］

この先,家庭の中に音声AIを備えた機器が増えていく中で,音声AIが取得する音をつかいユーザーの状況を推測し,共有することが可能になれば,ユーザーの状況に合わせたサービスの提供や,室内見守りシステムの構築など幅広い分野への応用が期待される.本研究では,音声AIを搭載した機器がユーザーの状況を音を頼りに推測するシステムの提案と,システムの実現に必要な環境音の分析および学習と識別を行う.

第2章 既存技術

2.1 短時間フーリエ変換(STFT)

音声など時間変化する音響信号の周波数と時間(位相)の解析を行う際に,音響信号全体をフーリエ変換してスペクトルを得てしまうと周波数の時刻毎の変化を知ることが出来ないため,時刻毎のスペクトルを計算してその時間変化を知る処理のことを短時間フーリエ変換(STFT, short-time Fourier transform)という.

短時間フーリエ変換のアルゴリズムは,信号を前から順番に一定の範囲で切り出し,切り出した部分に窓関数をかけてからフーリエ変換を行い,切り出す範囲を少しずらして同じ処理を行うことを繰り返すものである.

スペクトル時間変化を表示した図をスペクトログラムと言う.

2.2 畳み込みニューラルネットワーク(CNN)

畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) とは,畳み込みを主な構成要素としたニューラルネットワークである.教師あり学習を前提とした学習を行い,画像認識の分野で非常に高い性能を発揮している.

畳み込み層とプーリング層などが積み重なることで多層のネットワークを構築する.

CNN での畳み込み処理は, フィルタと入力画像の内積をとり, ラスタスキャンにより繰り返し畳み込みを行うことで特徴マップを得る. 重みフィルタは, 誤差逆伝播法による勾配降下最適化法により学習される.

プーリングとは,入力される特徴マップの小領域から値を出力して新たな特徴マップに変換する処理である. プーリングを行う目的は 2 つある. まず, プーリングによりユニット数を減らすため, 調整するパラメータを減らすことが出来る. 2 つ目は, ある小領域から応答値を出力するため, 幾何変化などに対する不変性を獲得することが出来る

人間の手を介さずネットワークの学習を通して画像特徴量の自動抽出が可能になったことで,既存手法を著しく上回る精度を実現した.

2.3 AI スピーカーについて

AI スピーカーとは,対話型の音声操作に対応した AI アシスタントである音声 AI が搭載されたスピーカーである.音楽再生の機能の他,内蔵されているマイクでユーザーの音声を音声 AI に送ることが出来る.図 2.2 に AI スピーカーの動作例を示す.AI スピーカーに対して,ユーザーが何か発話をすると,音声 AI はユーザーの音声を認識し,情報の検索や連携した IoT 機器の操作を行うことが出来る.ユーザーからの発話に全て反応するわけではなく,ウェイクワードと呼ばれる各メーカーが決めた特定の言葉を認識した後,続く音声を命令として認識するようになっている.

図 2.2: AI スピーカーを通したユーザーの発話からサービスの提供までの図解

2.4 IoT 機器の普及と AI スピーカーの役割

IoT(:Internet of Things)は,主に家電などの製品をインターネットとつなぐことで,身の回りのもの同士のネットワークを作ることを指す.家電をネットワークに繋げることで遠隔操作を行う事を可能にしたものや,家電同士の相互通信によってデータの共有や,動作の連携を行う事が出来る.

今の IoT 機器の普及は,図 2.3 の様に AI スピーカーが,IoT 機器とユーザーとの仲介を行うことで,ユーザーの音声を”通訳”して音声操作行うことが可能になり利用が容易になった為である.しかし,”通訳”を行うのはスピーカーではなく音声 AI であることから,図 2.4 のように IoT 機器にマイクをつけることで音声 AI と通信を行い,直接ユーザーと IoT 機器をつなぐ”脱 AI スピーカー”といったことも行われている. [1]

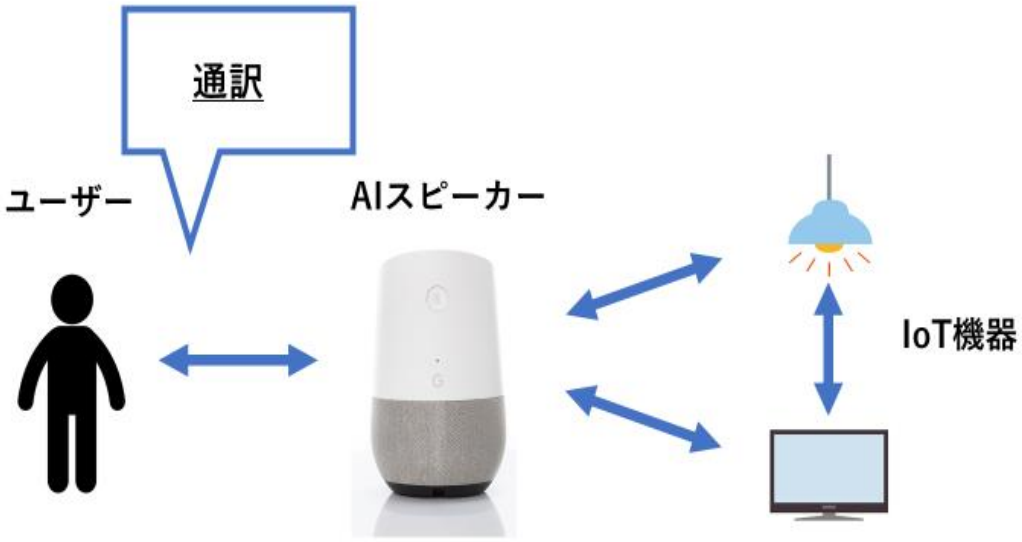

図 2.3: AI スピーカーがユーザーの言葉を通訳する図解

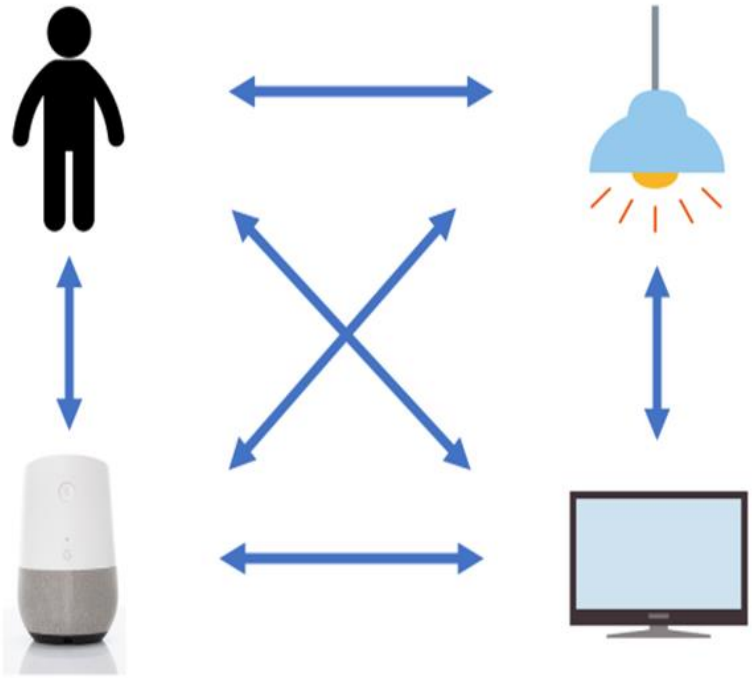

図 2.4: 各機器が音声対応になる脱 AI スピーカーの図解

2.5 マイクロフォンセンサを用いた在宅行動モニタリング

2006 年に行われた実験で,1 ヶ月間マイクを家の中の床や空間に 6 箇所設置し,在宅内で発生する音圧を計測し行動をモニタリングしたものである.日毎の室音パターンにまとめることで,室音の類似性を比べたところ,平日の行動パターンにある程度規則的な類似性が検出された. [3]

当時の実験では普通の家庭の中で,複数台のマイクが設置されることは想定されていないと思われる.これからは AI スピーカーを含んだ複数台のマイクが室内音をモニタリングすることでユーザーの行動を推測することが可能になると考えられる.

2.6 音状況認識技術（NEC）

NEC が 2016 年に,収集した雑多な音の中から目的の音を切り分け,起きている事象を認識する「音状況認識技術」開発したと発表した.これはあらかじめ目的音をマイクで収集し,目的音から細かい特徴を構成音として分けて学習させておき,学習していない未知の構成音を環境雑音として分けることで,環境雑音による影響を受けることなく高精度に構成音を抽出するというものである.ガラスの割れた音などを目的音とする事で災害検地や犯罪防止に役立てるという. [4]

異常を検知するために,日常では発生しにくいガラスの割れる音などを目的音として細かく特徴を抽出し,目的音以外の日常で発生する音は環境雑音として処理するという手法をとっている.しかし,この技術をそのまま日常で発生する音を目的音とする今回の提案システムで用いた場合,ユーザーごとの環境音を学習することが必要になることや,正例として学習したい命令を行うよりも前に発生する環境音と,負例として処理したい環境雑音との区別がつかないことが考えられる.

2.7 既存技術の問題点

アメリカでは,AI スピーカーを複数台家に設置することで,ユーザーが家のどこにいても,マイクがユーザーの声を拾い音声操作を可能とするといった,音声操作の普及が進んでいる.しかし,現状では AI スピーカーや音声 AI を搭載した IoT 機器は,ユーザーからの命令を音声 AI に送ることで動作を行う受動的なものである.これから音声 AI を搭載した機器が増えていくと考えられるが,受動的にしか動かない音声 AI を何台も家に置くだけでは不便な点が多く,ユーザーが発話し音声 AI に命令するまで,音声 AI はユーザーの状況を推測することが出来ない.そのため IoT 機器同士がユーザーの状況を共有することが出来ず,ユーザーの次の命令を待つことになる.

また,音声 AI を搭載した機器の普及が進むことで家庭の中に複数台のマイクが設置されようとしているが,これら複数台のマイクを利用した状況推測システムについて考慮した既存技術はほとんどなく改めて実験を行う必要がある.

既存技術ではガラスが割れる等の異常音を検知するといった緊急の状況を推測することを目的としたものが多いが,日常の動作からユーザー状況を推測するといった目的のためには,ユーザーごとの環境音を学習することが必要になることや,正例として学習したい命令を行うよりも前に発生する環境音と,負例として処理したい環境雑音との区別がつかないことが考えられる.

第3章 提案手法

3.1 既存技術の解決

これからは音声 AI を搭載した機器を複数台室内に設置することが普及していくと考えられるが,現状の音声 AI ではユーザーからの命令を待ち続ける受動的な動作しか行わない.ユーザーの状況に合わせたサービスの提案や提供といった能動的な動作を行うには,音をもとにユーザーの状況を推測することが必要になる.

ユーザーが音声 AI に命令する時の状況を推測するためには,命令より前の環境音を判別できるように機械学習を行う必要がある.しかしユーザー毎に違う正例となる環境音と,負例となる正例以外の環境音との区別をつけることは容易ではない.ユーザー毎の正例データは時間をかけることで蓄積されていくが,決まった目的音の細かい特徴量をあらかじめ抽出しておく既存技術の手法が行えないために,負例をそれ以外とすることが出来ない.そこで,あらかじめ命令が行われる際の環境音を分析し,数種類のシンボル音を定め分類しておく.ユーザー毎に異なる環境音のなかから,シンボル音をいくつか識別し,シンボル音が発生する順番と時間間隔のパターンを学習することで,命令が行われる状況とパターンとを結びつけることが出来ると考えた.

3.2 提案システム

あらかじめ,日常的に音声 AI に命令する中で毎日繰り返し命令が行われるような比較的頻度の高い命令を想定し,その時のユーザーの命令より前の環境音を分析する.特徴的な環境音を数種類シンボル音として分類しておく.シンボル音の例としては,ドアの鍵が開く音,ドアが開く音,足音,靴を脱ぐ物音などをそれぞれシンボル音とする.

ユーザーの家に設置された複数台の音声 AI が搭載された機器をつかい,命令や一定以上の確度のシンボル音が発生するたびに,命令やシンボル音の種類と発生した時間とを記録し収集する.収集したデータを用いることで,特定のシンボル音の並びが類似した時間間隔で発生する度に特定の命令が行われているといったパターンを解析することが可能になる.データ解析あるいは機械学習により判明した特定のシンボル音の発生パターンと類似した発生パターンを認識したとき,特定の命令を行うかユーザーに提案する.

このシステムが完成すれば,『日毎の室音パターンを比べるために 1 月録音し続ける』といった既存研究の手法と比べて,記録するのは発生したシンボル音や命令の種類とその時間のデータのみと記録データ量が少なく済む上に,ユーザーの状況を推測することで,ユーザーに対して命令よりも先にサービスの提案や提供が可能

になるほか,家の中の全ての IoT 機器にユーザーの状況を共有することが可能になるため,家電の管理や制御といった点でも役に立つと考えられる.

3.3 提案システムの技術的問題点

音声 AI はユーザーからの命令を受ける際,最初にウェイクワードと呼ばれる特定のワード（OK, Google ！ やへイ Siri !）を認識するまでは命令を受け付けない仕組みになっている.提案システムではこの仕組みをそのままシンボル音に当てはめることでシンボル音を識別することを考えたが,ウェイクワードは音声認識技術を使い複数の音声データを認識後のシンボルデータを用いてあらかじめ辞書登録することで識別を行う.環境音であるシンボル音に音声認識を当てはめることは困難であるため,環境音は短時間フーリエ変換を行い周波数軸と時間軸によるスペクトログラム画像への変換を行うことで学習を行い,それぞれのシンボル音の特徴を抽出し,シンボル音の識別を行う.

第4章 実験内容と考察

4.1 実験概要

本章では,提案システムがユーザーの状況を推測するために行うパターンマッチングで機械学習の手がかりとして使用するシンボル音の識別に関する実験と考察について述べる.提案システムに必要なシンボル音に対して①何をシンボル音とすることが出来るか,②それぞれ異なるシンボル音同士の識別が可能であるか検証を行うために2種類の実験を行った.実験1は①について,実験2は②について検証するために実験を行った.

4.2 実験環境

本実験にて使用した実験環境を以下に示す.録音環境は AI スピーカーの代わりとしてマイクが 2 つ付いた IC レコーダーを使用し,録音したデータの正規化と形式変換はAudacityを使用した.音声データは全てwaveファイルで構成されている.実装は Python にて行い,配列処理等は Numpy,CNN 等のモデル構築と学習はTensorflow,Keras,音声データのスペクトログラム画像への変換はLibrosaによって行った.

Audacity
Ubuntu 16.04 LTS
NVIDIA GeForce 960M
CUDA 9.0
cuDNN 7.0.5
Python 3
OPENCV 2.0

Numpy 1.15.4
Tensorflow-gpu 1.12.0
Keras 2.2.4
Librosa

4.3 実験方法

4.3.1 実験 1

実験 1 は日常の環境音の中から何をシンボル音とすることができるかを調べるために,録音した日常の環境音の中からシンボル音の候補となる特徴音を調べた.

録音環境は音声 AI の代わりとして,マイクの 2 つ付いた IC レコーダー(LS-14 LINEAR PCM RECORDER, OLYMPUS)を使用し,玄関とリビングの 2 箇所にそれぞれ 1.0 メートル程の高さになるように三脚を設置し収録を行った.

期間は 12 月 20 日から 1 月 10 日までの 3 週間,1 日 6 時間以上の録音を行い,合計約 152 時間分(約 113.5GB)の録音を行った.

録音した環境音を実際に聞くことで環境音の中からシンボル音の候補となるまとまった音をリストにまとめた.

4.3.2 実験 2

実験 2 は実験 1 でシンボル音の候補として挙げられた特徴音が,実際にそれぞれ異なるシンボル音として識別が可能であるか検証を行うために,実験 1 で調べたシンボル音の候補の中から人間でも異なるシンボル音同士の識別が行えているかが判りやすい音を選び出し,機械学習を行い,識別が正しく行えているかの検証を行った.

実験 1 のシンボル音の候補の中から人間でも識別が正しく行われているかを判別できる特徴音として, ユーザーが帰宅した際に音声 AI に電気をつけるように命令する状況を想定し,外側から鍵を開ける音,ドアの開閉音,内側から鍵を閉める音の 3 つの特徴音を選んだ.

3 つの特徴音をそれぞれ 120 個ずつ録音したデータを人の手で用意した.録音方法は実験 1 で使用した IC レコーダーを玄関の扉から 1.2 メートル離れた位置に高さ 1.0 メートル程の高さで設置し,シンボル音候補となる 3 つの特徴音をそれぞれ数秒間ずつ収録した.

録音したシンボル音は図 4.1 で示すような縦軸が周波数,横軸が時間のスペクトログラム画像になるように短時間フーリエ変換を行い,図 4.2,図 4.3,図 4.4 で示すような 380 * 380 pix の画像データに変換した.

鍵を開ける音,ドアの開閉音,鍵を閉める音を画像データ毎にラベリングし,それぞれ 120 枚のうち 80 枚を学習用データに 20 枚を検証用データにして CNN モデルを構築し学習を行った.このとき CNN の学習結果として識別の精度と損失値をグラフにした.

学習の済んだモデルで未確認データ各 20 個ずつの識別を行いその予測精度から識別が可能か検証した.

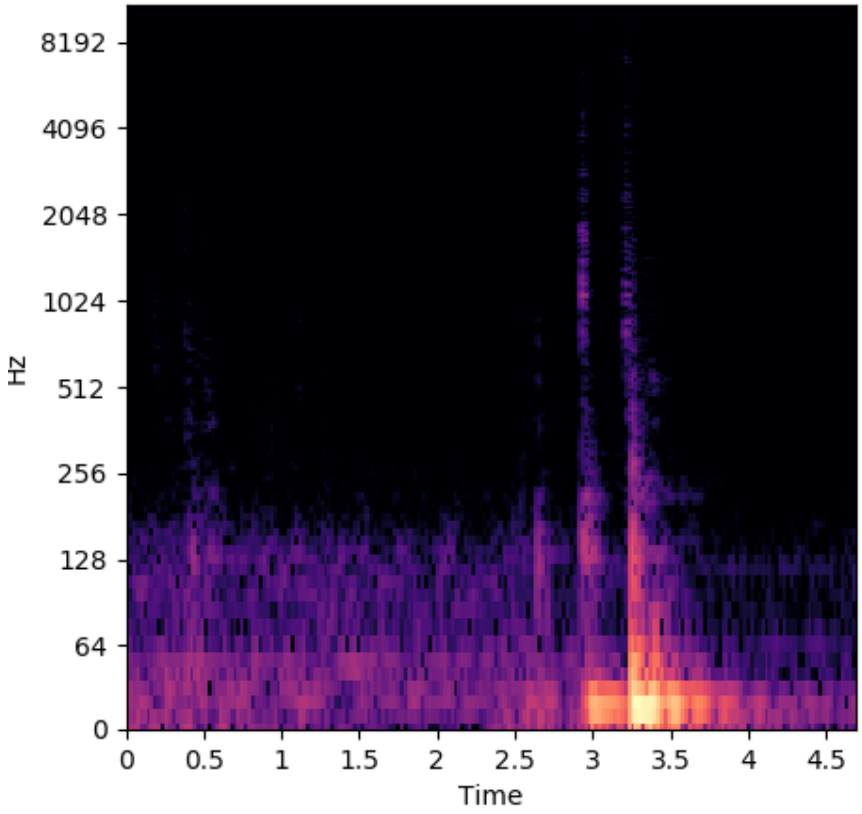

図 4.1: 短時間フーリエ変換を行いスペクトログラム画像に変換したドアの開閉音

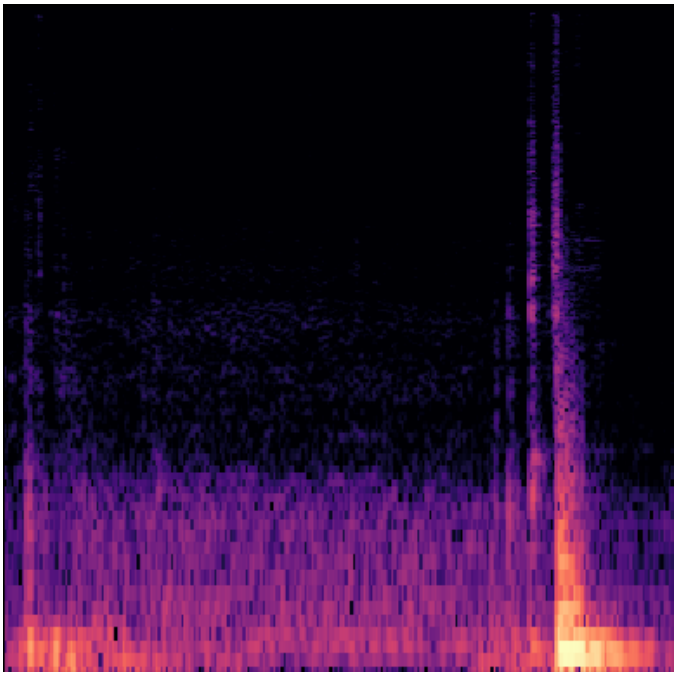

図 4.2: 切り出したドアの開閉音のスペクトログラム画像

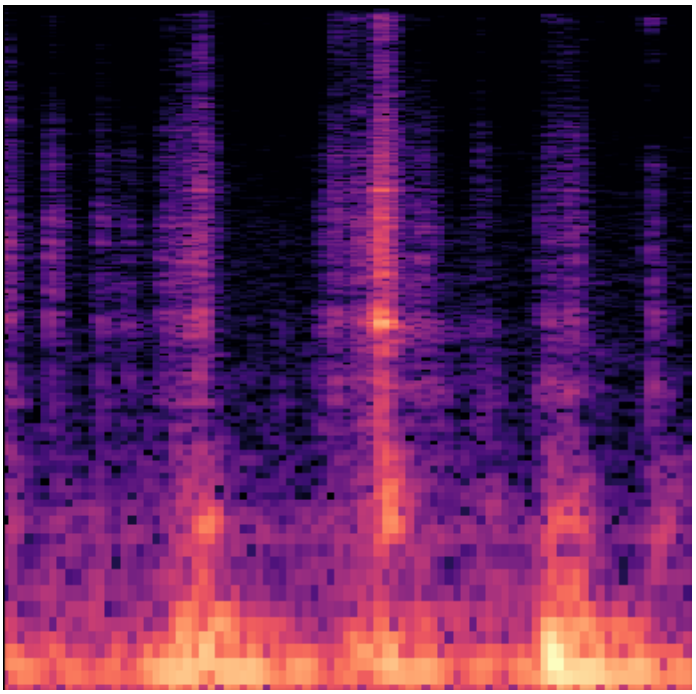

図 4.3: 切り出した外側から鍵の開く音のスペクトログラム画像

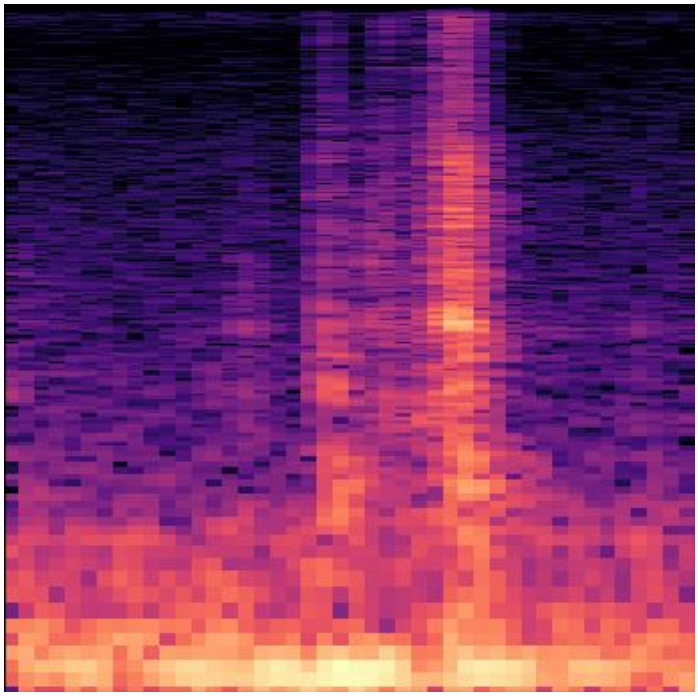

図 4.4: 切り出した内側から鍵の閉める音のスペクトログラム画像

4.4 実験結果

4.4.1 実験 1

実験 1 では玄関とリビングでそれぞれ収録した環境音の中からシンボル音の候補となる特徴音を調べた.

玄関で収録した環境音の中からシンボル音の候補となる特徴音を下記に列挙する.

・自動車やバスが外を走る音
・自動車のエンジン音
・飛行機やヘリの音
・外から鍵を挿す音
・鍵を開ける音
・鍵を抜く音
・ドアが開閉する音
・内側から鍵を開閉する音
・人の話し声
・足音
・荷物が壁に当たる物音
・靴を脱ぐ物音
・靴を直す物音
・インターホンの音

リビングで収録した環境音の中からシンボル音候補となる特徴音を下記に列挙した.

・自動車やバスが外を走る音
・自動車のエンジン音
・飛行機やヘリの音
・人の話し声
・人の足音
・衣擦れの音
・ベルトの金属音
・鍵や鍵束がすれる金属音
・ベランダや廊下等のドアを開閉する音
・引き戸等を開閉する音
・テレビの音声
・椅子が軋む音

・テーブルに物がぶつかる音
・ドライヤーの音
・洗面台の水音
・台所で洗いものを行う音
・食品等の包装をあける音
・まな板を包丁で叩く音
・冷蔵庫を開ける音
・エアコンの操作音
・電子レンジ等の家電を使用する音
・猫の鳴き声や足音
・電気をつける音
・電気を消す音
・目覚まし時計に関する音

4.4.2 実験 2

実験 2 の結果を図 4.5, 図 4.6, 図 4.7 に示す. 図 4.5 の縦軸はモデルの識別精度, 図 4.6 の縦軸はモデルの損失値, 横軸は epoch(データセット全体の処理の単位, その反復回数), 折れ線グラフは学習用データ, 点線は検証用データの結果である.

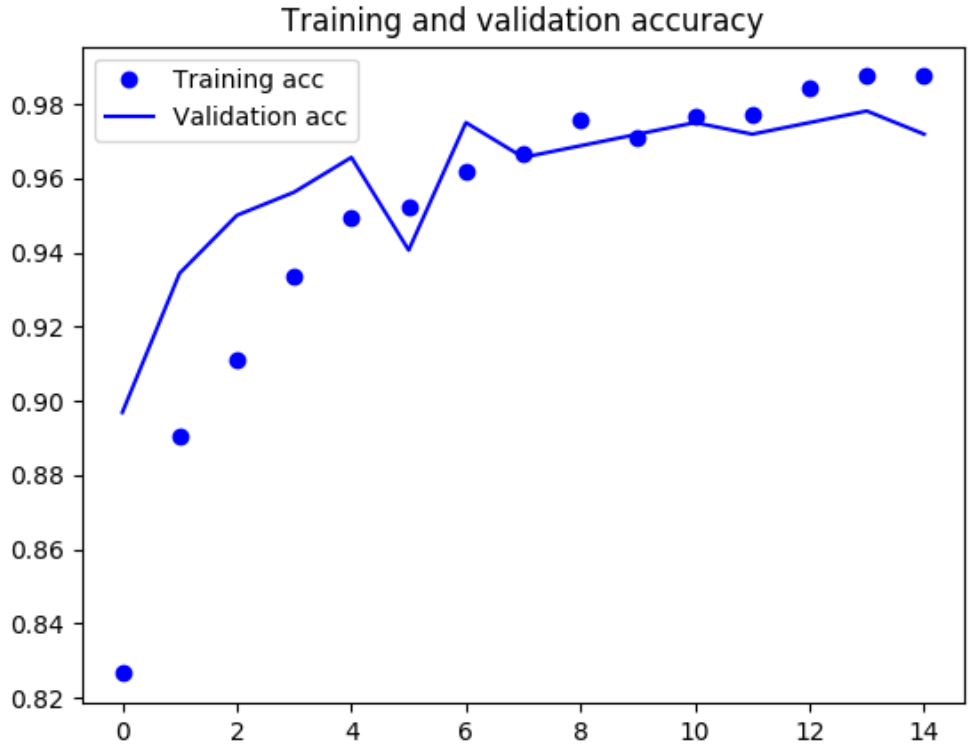

図 4.5: 精度を表すグラフ

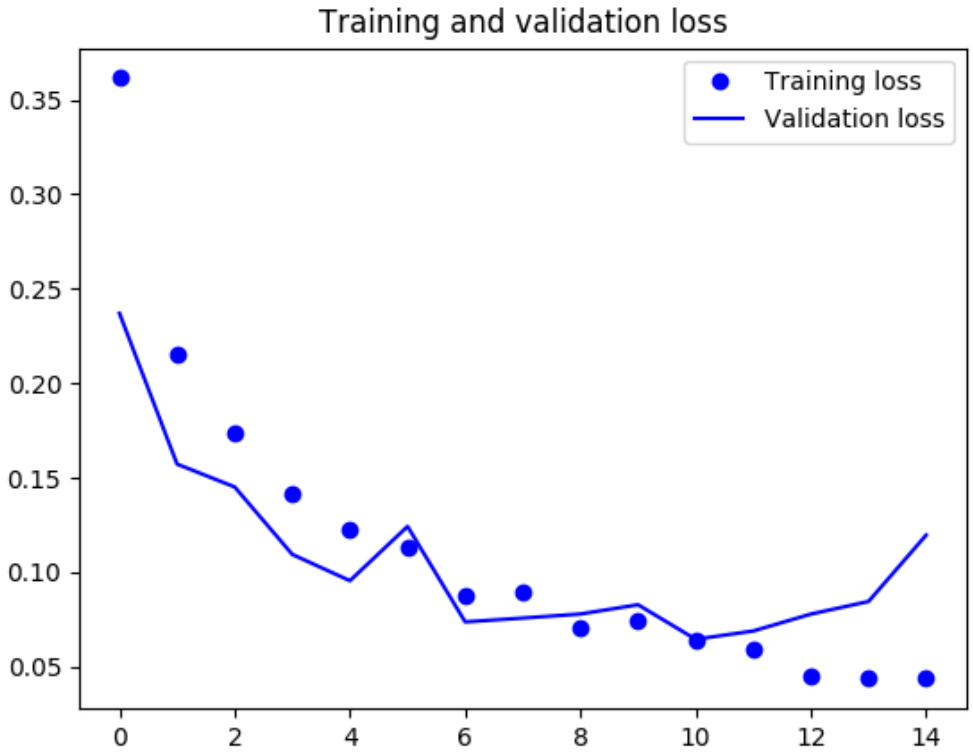

図 4.6: 損失値を表すグラフ

```
loss= 0.35093609710434637
accuracy= 0.978125
```

図 4.7: 未確認データを識別したときのモデルの予測精度の平均

4.5 考察

実験 1 の結果から, 日常の環境音の中からシンボル音の候補となる特徴音を調べたところ, シンボル音の候補となるものが多数挙げられることがわかった. シンボル音の候補として挙げた特徴音の中には一見するとノイズと思われる音もあるが, 提案システムで行うパターン解析あるいは機械学習の手がかりになるシンボル音は, 人間から見ると規則性のない特徴音であっても人工知能が長期的な視点で解析を行う中で規則性が見つかる可能性があるため, 少しでも多くのシンボル音の候補がある方が望ましいため, まとまりのある音はシンボル音の候補として上げられると考えられる.

実験 2 は実験 1 でシンボル音の候補として挙げられた特徴音が, 実際にそれぞれ異なるシンボル音として識別が可能であるか検証を行った. 1 つの音につき学習用データ 80 個, 検証用データ 20 個のスペクトログラム画像で学習を行った結果 CNN モデルの識別精度が未確認データに対して平均 97.8%と高い精度で識別が可能であることがわかった. 図 4.4, 図 4.5 のグラフから検証用のデータは画像データが少なく過学習が起こっていることがわかるが, 図 4.6 の未確認データに対する識別精度の高さから 80 枚で行った学習で十分に学習が進んでいたことがわかった.

個別のシンボル音として識別が可能であることがわかったため, シンボル音をシンボルデータの列として扱うことが出来るようになった. シンボルデータの列として扱うことで, データとして軽いものになった. 今後はシンボルデータの列からユーザーの命令より前の環境音に含まれているシンボルのパターンと類似するシンボルのパターンを見つけることができないか検証を行う必要がある.

第5章　謝辞

本論文を作成するにあたり,多くのご指導,ご助言を頂きました三好力教授に厚くお礼を申し上げます.また,議論に協力してくださった三好研究所の皆様や学友に心から感謝致します.

第6章 参考文献

[1]： “AI スピーカー日本上陸,API が家電を支配する”
野澤哲生, 日経エレクトロニクス, 2017/10/19
https://tech.nikkeibp.co.jp/dm/atcl/mag/15/00176/00001/

[2]： “IoT の原点,センサーネットワークの本質”
猿渡俊介, 日経エレクトロニクス, 2015/12/11
https://tech.nikkeibp.co.jp/dm/atcl/mag/15/100600071/

[3]： “マイクロフォンセンサを用いた在宅行動モニタリング”
川崎医療福祉学会誌 Vol.15 No.2 2006 615-620
品川佳満, 岸本俊夫, 太田茂, 2005/10/31

[4]： “NEC 音状況認識技術”
日本電気株式会社, 2016/11/28
https://jpn.nec.com/press/201611/20161128_03.html

[5]： “CNN 画像認識”
2018/06/26
https://qiita.com/tomo_20180402/items/e8c55bdca648f4877188