

# モバイル携帯端末を用いた環境音収集とその認識手法の検討

津田 貴彦<sup>1,a)</sup> 中西 恭介<sup>1</sup> 松山 みのり<sup>1</sup> 西村 竜一<sup>1,b)</sup> 山田 順之介<sup>2</sup> 河原 英紀<sup>1</sup> 入野 俊夫<sup>1</sup>

**概要:** 本研究では、環境音を入力とするインターフェースを有するモバイルアプリケーションの開発を行っている。実現に必要なのは、環境音認識手法の開発と、環境音サンプルの収集及び、クライアントアプリケーションの実装である。認識システムを予備評価した結果、アルゴリズムの改良と学習用データの拡充が必要であることを確認した。この問題に対し、データ収集用の Android アプリケーションを作成し、学内ではサークル等の活動に伴う音を 29 時間 24 分、学外では電車の走行音や救急車のサイレン等の音を 10 時間 36 分にわたって集めることに成功した。本発表では、収集データの分類と、その認識手法について議論する。

**キーワード:** 環境音認識, Android 携帯端末, 音データベース, クラウドソーシング

## Development of Collection and Recognition Method for Environmental Sound Samples using Android Mobile Devices

TSUDA TAKAHIKO<sup>1,a)</sup> NAKANISHI KYOSUKE<sup>1</sup> MATSUYAMA MINORI<sup>1</sup> NISIMURA RYUICHI<sup>1,b)</sup>  
YAMADA JUNNOSUKE<sup>2</sup> KAWAHARA HIDEKI<sup>1</sup> IRINO TOSHIO<sup>1</sup>

**Abstract:** We have been developing an Android mobile application which can recognize environmental sound signals. This report describes environmental sound signal recognition method, our collection of environmental sounds, and an overview of the prototype system. In order to collect further samples of environmental sounds, Android applications for data collection was developed.

**Keywords:** Environmental Sound Recognition, Android Mobile Device, Sound Database, Crowdsourcing

### 1. はじめに

iPhone の『Siri』[1] や docomo のスマートフォンで利用できる『しゃべってコンシェル』[2] といった、音声認識機能を持つ音声エージェント機能の実用化が始まっている。これらのシステムは、音声から利用者の意思を理解し、それに即したアプリケーションの起動や情報の提供を可能にした。一方、我々は日常生活において、周辺の環境音からも多くの情報を得ている。すなわち、警報音や環境音のような非言語音には、ユーザーに気づきを与える重要な情報

が含まれていると考えることができる。

本研究では、環境音からその場の状況を判断し、それに応じて適切な情報を利用者に提示するモバイルアプリケーションの開発を目指す。例えば、図 1 のように、野球部が練習しているグラウンド付近の環境音を認識する事で、その野球部に関する詳しい情報を取得することができる。また、警告音等を認識し、利用者に適切な情報を提示する事で、利用者の行動をサポートすることができる。現在、環境音認識の検討 [3] と、音響信号サンプルの収集および Android アプリケーションの実装を行っている。現在までに、環境音を収集するための Android アプリケーションを作成し、収集した音響信号サンプルを用いて認識の予備実験を行った。本稿では、その結果を報告する。

<sup>1</sup> 和歌山大学  
Wakayama University

<sup>2</sup> 日本電信電話株式会社  
NTT

<sup>a)</sup> s135032@sys.wakayama-u.ac.jp

<sup>b)</sup> nisimura@sys.wakayama-u.ac.jp



図 1 提案システムの利用シーン

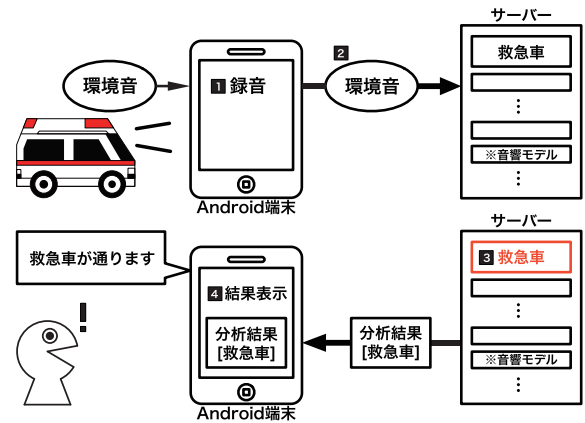


図 2 提案システムの構成

## 2. 提案システムの概要

日本国内で普及しているスマートフォンの OS の割合は、Android が 50%以上を占めている。また、Android アプリケーションの開発環境は無償で公開されており、アプリケーションを公開する際も制限が無い。そこで、本研究では、提案システムを Android OS で動作するアプリケーションとして実装している。

これまでに、本研究では、提案システムのプロトタイプを作成した。図 2 に示すように、クライアントプログラムとウェブサーバ上で動作するサーバプログラムの連動によってシステムを構成する。このアプリケーションでは、環境音を 1 秒間録音し記録したデータをサーバに送信する。次にサーバから返却された認識結果に関する情報が記載されているウェブページを提示する。プロトタイプでは、認識結果に対して、図 3 で示したアイコンを含めたウェブページを提示することができる。

システムの内部処理の概要は以下になる。

- (1) Android 端末の内蔵マイクロホンで周囲の環境音を収録
- (2) 収録信号のデータと、収録時間及び収録位置情報をインターネットを介して分析用サーバに逐次アップロード
- (3) 受信データを分析用サーバで統計的認識処理
- (4) 認識結果をクライアント側に返却
- (5) 結果に関する情報を Android 端末上で提示

また、サーバ上に収録信号を蓄積することで、認識処理の学習用データとし、利用することを予定している。Android 端末上で動作するクライアントプログラムは Java、サーバ上で動作するプログラムは CGI で実装している。

## 3. 環境音認識システムの性能評価

収集した音響信号サンプルの一部を用いて環境音認識の予備実験を行った。提案システムは、音響信号等の比較的大容量のデータを扱うため、サーバ・クライアント方式を採用した。このため、提案システムを利用するにはネッ

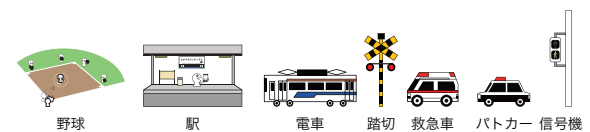


図 3 情報提示に使用する画像の例

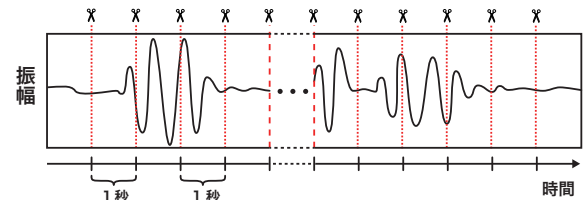


図 4 収録した環境音を 1 秒毎に区分

ワークに接続している必要がある。しかし、場所によっては、通信環境が常に安定しているとは限らない。また、音響信号の収録時間が長くなると、システムに与える負荷が高くなるだけでなく、情報提供までに要する時間も長くなる問題が一般に生じる。このような状況を想定し、本研究では、図 4 に示すように、収録信号を 1 秒間毎に分割した上で、サーバ上の認識システムの入力とすることを検討した。本研究では、分割された固定長の学習データを大量に用いることで、比較的単純な認識アルゴリズムでも認識性能を確保することを目指す。

また、収集した環境音には、途中でターゲット音の信号が存在しない区間が含まれる。例えば、ブザーや信号機の誘導音などの、断続的に発せられる音を収録したデータには非ターゲットの区間が多く含まれる傾向がある。そこで、図 5 に示すように、ターゲット区間と非ターゲット区間に分ける時間ラベルを手作業で付与し、ターゲット区間に対する学習処理を行った。

### 3.1 予備評価実験

収集した音響音の一部を用いて、認識の予備実験を行った。

今回の予備実験に用いたデータは、事前に収集した環境音からの抜粋であり、10 のクラスから構成される (表 2 参照)。各クラスに含まれる音の具体的な内容は、以下のよう

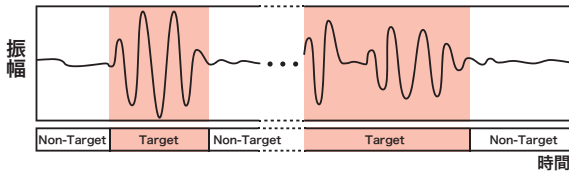


図 5 ターゲット区間と非ターゲット区間を分ける時間ラベル

表 1 実験条件

|           |                                                  |
|-----------|--------------------------------------------------|
| サンプリング周波数 | 16kHz                                            |
| 量子化 bit 数 | 16bit                                            |
| 音響特徴量     | 12 dim. MFCC<br>+ $\Delta MFCC$ + $\Delta Power$ |
| クラス数      | 10                                               |
| 評価方法      | leave-one-out 交差検証                               |

になっている。「救急車」、「パトカー」、「消防車」はそれぞれのサイレンの音である。「ブザー」は安全ブザー、防犯ブザーの 2 種類を使用した。「踏切音」は踏切警報機が鳴動しているときの音である。「信号機」は青信号の時に鳴る、視覚障がい者交通信号付加装置の誘導音である。「ホイッスル」は電車出発を通知する車掌の笛音である。「電車内」は電車に乗っている状態で走行音を録音したものである。「駅のホーム」は駅のホームで収録した音で、電車の通過、到着を知らせるチャイム、到着時のブレーキ音などが含まれる。「電車のドア」は電車のドアの開閉警告音、開閉状態を知らせるサイン音が含まれる。実験に用いたサンプルの数は、区分前で合計 132 個、区分後は 1,860 個となった。

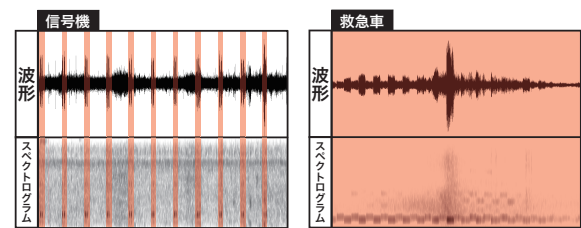
予備実験では、話者認識で広く用いられる HMM (隠れマルコフモデル) による認識手法を評価した。実験条件を表 1 に示す。この際、収録信号は、量子化ビット数 16bit、サンプリング周波数 48kHz で録音を行っているが、実験前にサンプリング周波数を 16kHz にダウンサンプリングし、窓シフト長 10ms で音響分析を適用することで入力とした。音響特徴量の抽出には、HTK[6] を用いた。収録音声を一秒で分割し、音声認識と同様の、12 次元の MFCC (メル周波数ケプストラム係数) と  $\Delta MFCC$ 、 $\Delta Power$  を音響特徴量として抽出した。MFCC は、過去にも雑音認識の特徴量に採用され、音声以外を入力信号とする認識処理において、一定の精度が確認されている [4]。

本実験では、分割前の 1 サンプルを評価データ、残りの全てを認識器の学習用データとし、完全にオープンな状態で交差検証を行った。学習用と評価用を 1 サンプルずつ入れ替えて実験を繰り返した。

HMM を音響モデルとした音声認識器を利用した方法は、式 (1) により結果クラス  $s_r$  を出力する問題として定式化できる。式中の  $p(o|\lambda_s)$  は、入力した特徴量  $o$  がクラス  $s$  の音響モデル  $\lambda_s$  と適合する尤もらしさの度合いを示す尤度である。 $p(o|\lambda_s)$  は、音響モデルによって算出される。

表 2 実験試料と評価実験結果

| クラス       | サンプル数 | 分割数  | 正解率 [%] |
|-----------|-------|------|---------|
| 救急車のサイレン  | 16    | 374  | 72      |
| ブザー音      | 3     | 21   | 5       |
| 踏切音       | 24    | 575  | 70      |
| ホイッスル     | 8     | 20   | 60      |
| パトカーのサイレン | 2     | 31   | 0       |
| 信号機の電子音   | 9     | 161  | 27      |
| 消防車のサイレン  | 5     | 193  | 18      |
| 電車のドア     | 35    | 95   | 56      |
| 電車内の音     | 19    | 242  | 86      |
| 電車外の音     | 11    | 148  | 68      |
| 合計        | 132   | 1860 | —       |



(※ 赤の区間はターゲット区間であることを示す)

図 6 「信号機」と「救急車」でのターゲット区間の違い

$$s_r = \arg \max_s p(o|\lambda_s) \quad (1)$$

今回、認識プログラムとしては、オープンソースの音声認識器 Julius 4.2.1[5] を利用した。音響モデルには、3 状態の GMM (混合ガウス分布モデル) からなる HMM を用いており、その構築には HTK を用いた。

実験結果として、表 2 に正解率を示す。「救急車のサイレン」、「踏切音」、「電車内の音」に関しては 70% 以上の高い正解率を得ることができた。「電車内の音」に関しては、収録環境が室内であったため、風や雑音の影響を受けにくく、収録条件が他の環境音よりも良かったことが高い正解率の理由に挙げられる。一方で、「ブザー」、「パトカー」、「信号機」、「消防車」は、他のクラスに識別される事例が多数となった。これらは、いくつかの原因が考えられる。まず、「ブザー」や「信号機」などは、ターゲット信号の継続時間が短く、音響モデルの学習に際して、十分なサンプルが得られなかったことが原因の 1 つに挙げられる。例えば図 6 に示すように、「救急車」のような連続した音はサンプルとして実質的に利用できるデータ量も多くなるが、「信号機」のような断続的な音は、サンプルとして利用できる数が少なくなる。また、サンプル数の偏りも結果に大きく影響している。「パトカー」、「ブザー」、「消防車」の正解率が低いのは、他のクラスに比べてサンプル数が極端に少ないことが原因の 1 つであると考えられる。

今回の予備評価からは、認識システムの改善と、学習データの拡充が必要であることを確認できた。また、シス

テムに与える負荷などを考慮して入力信号の区分長を1秒間と短く設定したが、認識処理の入力とする上で問題が生じないことを確認した。

#### 4. 環境音の収集

本研究では、前節で問題となった、実環境下の環境音サンプルの拡充のために、環境音収集用アプリケーション ROCKON を作成した。Android 端末を用い、実環境で周囲の雑音も含めて収録することで、利用実態に即した検討を可能とするためである。

収録の条件として学外と学内の2つの場面を想定し、それぞれ異なる方法で環境音の収集を行った。学外では、電車の走行音や救急車のサイレン等を対象に収録を行った。学内では、同じ地点で変化する環境音を観察するために、定点で24時間の収録を行った。

##### 4.1 環境音収集アプリケーション ROCKON

環境音を効率良く収集するために、学外と学内のそれぞれの収集目的に合わせた2つの環境音収集アプリケーションを作成した。アプリケーションの名前は、学外用を ROCKON、学内用を ROCKON 1s とした。

学外では、録音対象を見つけた時に、できるだけ素早く録音を開始できるように、ROCKON の待機画面には録音開始ボタンとデータ送信ボタンのみを配置し、録音終了時に録音した対象の名称を入力するようにした。また、収集したデータの移行をネットワーク経由で行うことで、大人数での収集を可能とした。ROCKON では、録音を開始した位置と時間の情報と、使用した端末の機種種のログを保存する。これにより、その環境音が、いつ、どこで録音されたかが分かる。また、機種種の仕様の違いによる、収録データの質の違いを調べることが可能である。

ROCKON のユーザインタフェースを図7に示す。操作の流れは以下になっている。

- (1) REC ボタンをタップすると環境音の保存を開始。
- (2) STOP ボタンをタップするか、保存開始から60秒間経過したら保存を停止。
- (3) 画面が遷移した後、収録した対象の名称を入力し、SAVE ボタンをタップしデータを保存。
- (4) SEND ボタンを押すと、これまでに保存した全てのデータをサーバにアップロードする。

学内向けでは、定点で録音を行うため、ROCKON 1s では最初に場所の名称の入力を行った後、録音を開始する。また、一度に大量のデータを集めるために、SDカードの容量限界まで連続で録音可能とした。本研究では固定時間長で区分したデータを学習データとして用いるため、録音の際に自動的に区分するようにした。このアプリケーションでは大量のデータを扱うため、収集したデータの移行はSDカードから直接行う。



図7 ROCKON のユーザインタフェース

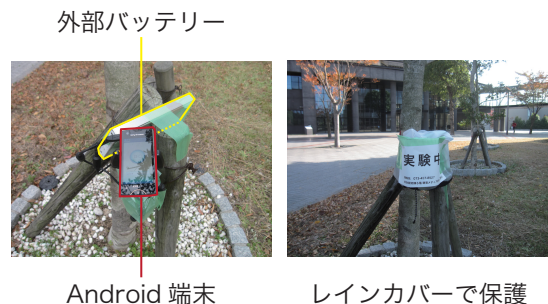


図8 収録時の様子

##### 4.2 環境音の収集

###### 4.2.1 学内での収集

和歌山大学内では、グラウンドや食堂などの各地点で定点で24時間録音を行うことで、時間で変化する環境音を観察した。収録時の様子を図8に示す。付近の植木にAndroid端末と充電用のバッテリーを固定し、雨風から保護するために全体をレインカバーで覆った。グラウンドで録音を行った結果、野球部の練習に伴う音や体育の授業中の周辺音を29時間24分収集することに成功した。1秒間毎に区分された環境音のファイル数は105,070個となった。

###### 4.2.2 学外での収集

学外の収集では、当初、本研究室の学生らで、学外で電車の走行音や救急車のサイレン等の録音対象を見つけた時に、その音が継続している間録音を行った。この時、10クラスに分類できる環境音を収録時間にして31分間集めることが出来た。このデータは、表2で示した実験試料と同じものである。

上記の一方、少人数であらゆる場所の環境音を大量に収

集することは難しいため、クラウドソーシングのアプローチ [7] を採用し、楽天リサーチ社のモニタ誘引サービスを利用し、大規模に環境音を収集した。はじめに、モニタ登録をしている被験者候補に、ROCKON のダウンロードサイトの URL が書かれたメールを送信した。それを受信した被験者が、所有している Android 端末に ROCKON をインストールすることで、本実験に参加することができる。環境音を録音し、データをアップロードした被験者には、報酬として本社よりポイントが付与される仕組みである。

様々な地域の環境音を収集するために、モニタの事前スクリーニングを同社に依頼した。スクリーニングでは、Android 端末を所有していることを確認した後、住所や電話番号等の連絡先を記入を得て、20 歳から 69 歳の男女で、居住地が分散するように被験者を選んだ。2012 年 12 月 27 日にメールを送信し、2013 年 1 月 8 日まで環境音のアップロードを受け付けた結果、日常で耳にする物音やサイレン等の環境音を 841 個集めることに成功した。そのうち 588 個は、録音開始時の位置と時間の情報を記録することに成功した。合計収録時間は 10 時間 5 分となった。

ただし、253 個のデータに関しては、GPS の位置情報を記録することができなかった。これは、アップロードの際にネットワークに何らかのトラブルが発生したため、記録したデータが途中で欠損してしまったことが原因の 1 つとして考えられる。また、録音開始までに位置情報の取得が間に合わなかった事例も確認された。

今回、428 人のモニタからデータのアップロードを得ることができた。事前のスクリーニングの結果、863 人のモニタに協力依頼のメールを送信したため、回収率は 50% となった。

収集した 10 時間 5 分のデータのクラス分類を手作業により行ったところ、93 のクラスに属することが分かった。表 3 は、クラス分類の結果、サンプル数の多い上位 20 クラスの収録データの構成である。上位の音には大きく 4 つに分けることができた。

- 電車内や車、信号の誘導音等の交通に関わる音
- 店内や人混みといった屋外での騒音
- 人の声や TV の音などの室内で収録した音
- 雨音や鳥の鳴き声などの人工物でない音

収集データの数では、交通に関する環境音が大多数を締める結果となった。また、今回の試みによって GPS の位置情報と時間情報に、周辺の環境音をリンクしたデータベースを作成することができるようになった。例えば、「踏切の警告音」が録音されているデータの GPS の位置情報を見ると、収録データが録音された具体的な踏切の場所を確認することができる。また、Google Map やストリートビューを利用することで、写真画像から、より詳しく周囲の状況の把握が可能となる。他にも、「駅のホーム」や「信号の誘導音」などの音源が移動しない条件下においては、

表 3 収集した環境音の上位 20 クラス

| 内容              | 個数 | 内容        | 個数 |
|-----------------|----|-----------|----|
| 電車内の音           | 86 | 雨音        | 15 |
| 車のエンジン音<br>や走行音 | 78 | 食事の音      | 13 |
| 踏切の警告音          | 64 | 人混みの騒音    | 12 |
| 人の声             | 58 | 救急車のサイレン  | 10 |
| 駅のホームでの音        | 51 | 水の流れる音    | 10 |
| テレビの音           | 50 | バスの社内外の音  | 9  |
| 信号の誘導音          | 40 | スーパーの店内の音 | 8  |
| 店内の音楽や放送        | 30 | 歩道の騒音や足音  | 27 |
| 歩道の騒音や足音        | 27 | ファンの回る音   | 7  |

GPS の位置情報は有効に活用できることが分かる。

## 5. まとめと今後の課題

本研究では、環境音を認識し、その結果に応じて情報を提示するモバイルアプリケーションのプロトタイプを開発した。事前に収集した環境音を用いて行った認識システムの評価実験では、認識システムの改善と、学習データの拡充が必要であることを確認した。

環境音を効率よく収集するため、Android 端末で利用できる環境音収集用アプリケーションを作成し、それを用いて学内及び学外で環境音の収集を行った。学内では定点で 24 時間録音を行った。学外では個人での収集に並行して、楽天のモニタ誘引サービスを利用して収集を行った。その結果、学内は 29 時間 24 分、学外は個人では 31 分、モニタ誘引サービスでは 10 時間 5 分の環境音を収集することに成功した。モニタ誘引サービスで収集した環境音は 93 クラスに分類することができた。

今後は、他の認識アルゴリズムとして、雑音の検出や音響信号の認識処理にも有効性が確認されている AdaBoost の利用を検討している [8]。また、位置と時間の情報を環境音認識アルゴリズムに組み込むことで、認識精度の向上が可能であると考え。将来的には、現在はプロトタイプのレベルである提案システムの情報提供機能を充実させる必要がある。

## 参考文献

- [1] <http://www.apple.com/jp/ios/siri/> (2013 年 1 月 11 日)
- [2] 吉村健, “しゃべってコンシェルと言語処理”, 情報処理学会研究報告, Vol.2012-SLP-93-4, 2012.
- [3] 津田ら, “携帯型端末で収録した音サインやサイレンなどの環境音認識の検討”, 日本音響学会 2012 年秋季研究発表会講演論文集, pp.1515–1516, 2012.
- [4] Lee, a., et al., “Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs”, Proc. INTERSPEECH, Vol. I, pp.173–176, 2004.
- [5] Lee, A., et al., “Julius – An Open Source Real-Time Large Vocabulary Recognition Engine”, Proc. Eu-

- rospeech, pp.1691–1694, 2001.
- [6] Young, S.J., et al., "The HTK book version 3.4", Cambridge University Engineering Department, Cambridge, UK, 2006.
  - [7] 秋葉ら, “クラウド時代の新しい音声研究パラダイム”, 情報処理学会研究報告, Vol.2012-SLP-92-4, 2012.
  - [8] 三宅信之, 滝口哲也, 有木康雄, “Multi-class AdaBoost を用いた雑音検出”, 情報処理学会研究報告, Vol.2006-SLP-64-2, 2006.