ニューラルネットワークによる実環境下での環境音認識
An environmental sounds recognition under real environment by neural network

山本 修生
Naoki Yamamoto
法政大学情報科学部ディジタルメディア学科
E-mail:naoki.yamamoto.9u@stu.hosei.ac.jp

Abstract

In this paper, we propose a method to recognize environmental sounds under real environment by computers. Therefore, at pre-experimentation, environmental sounds are recognized in an anechoic chamber. Afterward, environmental sounds are recognized under real environmental sounds. As a method, we used a neural network that has be prove successful in past studies on anechoic chamber data. There is stationary noise under the real environment sound. So we experiment using the raw data and the sound source separation data on the 4ch microphone array. For input, we experiment using MFCC(Mel-frequency cepstral coefficient), LPC(linear predictive coding) and power and spectrum at power peak as feature value for comparative. In addition, as proposed method, we propose a combination of MFCC and power pattern as feature value.As a result, in 83 kinds of anechoic chamber data was 94.3%, 80.1%, 74.2%, 94.7%, respectively. In 20 kinds of real environmental raw data was 98.9%, 95.6%, 91.4%, 99.0%. In sound source separation data was 94.9%, 87.6%, 78.8%, 95.7%. Even under the real environmental sound, MFCC input succeeded in giving a recognition rate of past study by considering human auditory characteristics. From this result, real environmental sound recognition can be expected practical use by removing noise.

1 まえがき

環境音の識別は「コンピューターが人や物事の状況や変化を認識する」というコンテクストアウェアコンピューティングを実現する上で重要な要素であると考えられる．ここで言う環境音とは街中や生活音などの非音声音全般を示すこととする．人間は音を用いて多くの情報を取り入れ，現状の把握に用いている．コンピュータにおいても同じことであり，スイッチ型のセンサーや，カメラだけでは死角におけるモノや人の動向を把握することは不可能であり，音の認識が必要になる．実用的な環境音認識器が作成出来れば，監視システムやライフログ，自立型のロボットの組み込み機能としての応用が期待できる．具体例として例えば，高齢化により，一人暮らしのお年寄りが増えているが，一人暮らしであるために，転倒し怪我を負ったり，体調を崩したりした場合でも一人で救急車を呼べず，しばらくして死体で発見されるようなケースがある．このような事故を防ぐには毎日誰かがこのお年寄りを訪問するなどの対策が考えられるが，在宅の高齢者を毎日長時間監視することは難しい．単純に生活の状況を把握するには監視カメラを設置することによって実現できるが，監視カメラのみの場合死角ができるというような問題が発生する．そこで，環境音を識別することによってこの問題を解決することが期待できる．

本稿では，実環境下での環境音識別器の作成を目指す．特徴量と識別手法を調べ，高い認識率を出す事，それに併せて環境音の区間を正確に検出することを目標とする．また，環境音の中でも生活音に絞って識別を行うことを考える．マイクをローカルで一つのコンピューターに接続し，音の解析を行うシステムの構築を行う．従来研究では難しかった単音の識別や，実環境下では発生するであろう雑音環境下の場合にも対応できるような方法を検討していき，さらに生活音の解析を正確に行うための手法や特徴量の検討を行っていく．

2 環境音認識システム

本研究の関連研究として，環境音の周波数変動が少ないという特徴から，音の時系列変動と周波数特徴を別々に抽出して組み合わせた 1 次元のベクトルを特徴量として，ニューラルネットワークを利用し，分類部，識別部の 2 段構成の手法を提案している [2] というものがある．分類部で，後述する 3 種類の音に分類し，識別部で具体的な音を特定する．具体的には，分類部では長時間のパワーパターン (1,48) を特徴量とし，識別部では (1,16) のパワーの時間パターンと (1,32) のパワーピーク時の瞬間的なスペクトルの組み合わせを利用している．従来手法である隠れマルコフモデルや時間遅れニューラルネットワークの場合は入力がスペクトログラムになるため，入力データが 2 次元となり，大きくなるが，この手法では 1 次元の入力で済むので処理が速いとしている．ニューラルネットワークは分類部，識別部共に隠れ層は 3 層でニューロンは 32 個である．

また，環境音は単音，繰り返す音 (定期，不定期)，長期間に渡る音の 3 種類に分類することが出来るとして，この 3 つの音に対してそれぞれのニューラルネットワークを用意し，識別部で利用している．単音というのはドアが閉まる音であったり，ボールが落ちた音．繰り返す音というのはタイピングを続ける音であったり，電話のベルの音．長期間に渡る音というのはスプレーの音や，ドライヤーの音である．この方法で RWCP（新情報処理開発機構）データベースの非音声音 [4] 45 種類の環境音のデータの識別率はおよそ 92 %であった．この研究や，その他の研究では，短い音の識別が難しいという課題がある．

しかし，この研究では実際の環境下を想定しておらず，雑音や，音の重なりが考慮されていないという問題が存在する．実際の環境下では認識対象とならないような小さすぎる音や空調の音等の雑音が入ったり，いくつもの環境音が重なることは明白であり，実用的なシステムの開発にはこれらを考慮し，多数のマイクを利用し，音を分離させたり，最も強いもののみを識

Supervisor: Prof. Katunobu Itou

別するなどの対応をする必要があるといえる．

本研究ではこれらを考慮しながら実験を進めていく．

3　識別手法

3.1　識別機

関連研究 [2] では NN を使用することのメリットは，汎化能力であり，訓練後のニューラルネットワークの認識部分は簡単で，隠れマルコフモデルよりも高速である点であると述べられている．汎化能力とは未学習の類似データに対する識別能力であり，同種の音であっても変動の大きい環境音認識において重要であると考えらる．また，簡単で高速であるという点は実用化するという課題において非常に重要な要素である．識別能力が高くとも実行に時間がかかってしまえば実用化には至らない．また，ニューラルネットワークは近年注目されている技術であり，パターン認識において優れた性能を発揮することができるという特徴がある．以上の点から本研究における識別方法として採用した．

3.2　特徴量

こちらも，現在までに音声認識で用いられている方法や，研究報告 [1] で紹介されている特徴量としては基本周波数，パワー，MFCC(メル周波数ケプストラム係数)，スペクトル等が挙げられる．

上記が主な音響イベントとしての特徴量であるが [3] のように物体の衝突や摩擦から 1kHz 以上の多数の共鳴成分が観察されることに着目し，音の波形の極値数を特徴量として利用するような研究もなされている．他に興味深いものとして，スペクトログラムを画像処理したものや，音響指紋技術によって得られる特徴量を利用している手法も存在する．従来研究では，音声認識，環境音認識共に，特徴量として MFCC そのものやそれを応用したものが主に使用されている．その他特徴量に関しても音声認識や話者認識，感情認識，音楽判定に使われ実績のある特徴量である．主にはこれらの値を参考にし，複数の値を組み合わせることなどによって使用する．特に MFCC は多くの実験で用いられ，実績がある [?]．本研究で対象とする音は人間に聞き取れる音であることから，人間の聴覚特性を考慮することで高い識別率を出すことができるのではないかと考え MFCC を用いることとした．また，その聴覚特性を考慮した場合と比較するために LPC 係数を用いた．その他に従来研究 [2] の比較用として，パワーパターンとパワーピーク時のスペクトルを用いた．更に MFCC とパワーパターンを組み合わせた特徴量を用いた．

3.3　実環境音下での問題への対処

実環境音下では雑音や，音の重なりがあることが予想され，周囲雑音があれば認識率が低下する事は免れない．人間の聴覚においては，雑音環境下でも特定の音を選択的に聞き取ることができる．これはカクテルパーティー効果と呼ばれるが，研究報告書 [5] によるとカクテルパーティー効果が生じる原因としては，それぞれの音源に対して両耳聴によって知覚される音像の空間的位置（方向と距離）の違い，音の大きさ，ピッチ，音色など音源の特性そのものの違いが関係していると見られていると述べられている．そこで本研究では人間の知覚に近づけるために，マイクロホンアレイを用意し，音源を分離するなどによって周囲の雑音に対して処理を行い認識率の低下を避ける方法 [7, 8] の検討を行った．

4　実験

4.1　環境音データベースを用いた実験

実環境を考慮する前に背景雑音なしでの実験を行う．これは雑音環境下での実験と比較を行うためである．ここでは，参考論文と同様に RWCP データベース [4] を用いて実験を行う．RWCP データベースは実環境における音声・音響信号処理の研究を対象とした評価用データベースで，無響室で測定された雑音が無いデータである．ここにはドライヤーの音や目覚まし時計の音など計 105 種類の非音声音が含まれている．このデータベースを用いることで実際の環境音に近い音で実験できると考え本実験に用いることとした．RWCP データベースの raw データを wav 形式に変換し，これを matlab で読み込み実験を行った．

4.1.1　ニューラルネットワーク

ニューラルネットワークはフィードフォワードネットワークの逆誤差伝播法を用いる．荷重の更新方法は共役勾配法を用いた．また，ニューラルネットワークの隠れ層のニューロンの個数は 10 個とした．学習の終了条件は検証データにおいて 6 回失敗した場合である．それぞれの特徴量，データセットでネットワークを用意した．入力層は特徴量数．出力層は識別数だけ用意した．出力は出力層のニューロンにそれぞれの環境音を割り当て，入力があると，いずれかひとつのニューロンに 1 を出力し，それ以外が 0 になる．

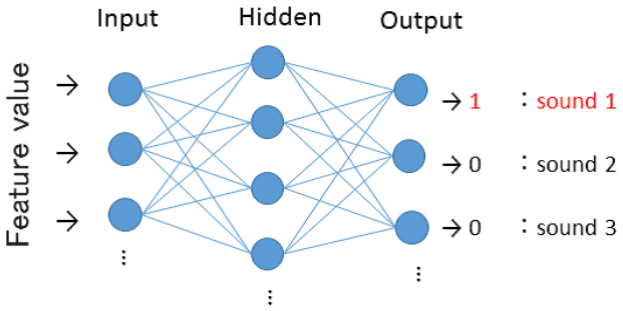

図 1　ネットワークイメージ

4.1.2　区間の決定

今回使用するニューラルネットワークの入力は固定長であるのでその長さを考えなくてはならない．区間を長くとりすぎると短い音において無音区間が多く発生し，データが冗長になるため，識別率が減少することが考えられる．また，区間が短すぎると長期間に渡る音の判別が困難になるという問題が発生することが予測できる．そこで RWCP データベースの無音区間を除いた音の長さを調査した．RWCP データベースの雑音レベルは実測値で 17.3dBA 44.3dBC であることから 0.001 以下の音を棄却した．その結果は図 1 のようになった．平均は 0.618 s，最高長は 3.56 s，最低長は 0.03 s，1 s 以上のものは 17 個で 1 s 以下のものが 84 %となり，1.5 s 以上のものは 5 個でそれ以下のものは全体の 95 %という結果になった．最も短い音は木板（小）と木板（小）を手で持って打ち合わせる音で 0.0921 s．最も長い音はコイン 1 個を合板に落とす音で 3.5648 s であった．

従来研究において，短い音の識別が難しいと言われている．このことから特徴量を抽出する区間が長すぎると入力が冗長になり, 識別率に影響を及ぼす可能性がある．よって RWCP-DB の半数を完全に内包することのできる平均値 0.6 s，84% を内包できる 1 s，95% を内包できる 1.5 s を基準に MFCC の特徴量として後述するデータセット 2 である RWCP-DB 内の音 14 種類を用いて比較実験を行った．

この結果から，最も識別率の高かった，0.6 s 区間を本実験に用いることとした．

4.1.3　実験

実験 1，2，3，4 で 3 種類のデータセットを用い，合計 8 種類の実験を行った．全ての RWCP データベース内の音を分類する前に，特定の音を選び識別機の作成を行った．選択した音

の基準としては 100 個のデータがあり，人間の耳で判別ができるものとした．3 種類のデータセットを用いた理由として，同種の音同士での識別率，異なる音同士での識別率，多種になった場合の識別率を測定するという目的である．

実験 1 では，MFCC(メル周波数ケプストラム係数) を用いた．本研究では，RWCP データベースの平均近くの値である 600ms を区間とするためパワーピークから前後 300ms を用いたフレームシフトを 50ms とした 13 次の MFCC を 25ms でシフトしたものを一列に並べた 299 次元の配列を用いた．周波数帯域の選択のために，RWCP データベース 14 種類の環境音で周波数範囲を変更した実験を行った．識別率のエラー数を調べ，最も識別率が高かった 0-8000Hz を用いた．

実験 2 では，LPC 係数を用いた．こちらも，実験 1 と同様な区間，フレーム長，フレームシフトで 13 個の LPC 係数一列に並べた 299 次元の配列を用いた．

実験 3 では，パワーパターンとパワーピーク時のスペクトル [2] を用いた．パワーパターンは実験 1，2 と同様な区間でフレーム長 20ms，フレームシフト 10ms で 60 点，パワーピーク時のスペクトルは 128 点で求め，スペクトル 128 点 + パワー 60 点を結合した 188 次元の配列とした．

実験 4 で，MFCC とパワーパターンを組み合わせた特徴量を用いた．実験 1 の MFCC299 次元と実験 3 でのパワーパターン部分 60 点を組み合わせた 359 次元の配列を用いた．

データセット 1 では RWCP データベース内で似ている木板（小）を手で持ち木棒で叩く音大, 中, 小の 3 種類を使用した．データセット 2 では RWCP データベースで音の種類として分類されている 14 種の中から各 1 種類を使用した．データセット 3 では RWCP データベースで 100 サンプルある音 82 種類を使用した．

データの使用内訳は各 100 データのうち 70% を学習用のデータ，15% を検証用データ，15% をテストデータとして用いた．

4.1.4 結果

結果を表 1 に示す．

表 1　RWCP-DB での実験結果

	データセット 1	データセット 2	データセット 3
実験 1	99.7%	99.4%	94.3%
実験 2	87.1%	93.6%	80.1%
実験 3	99.3%	94.2%	74.2%
実験 4	99.6%	99.3%	94.7%

データセット 1 では，誤判定の例として，木板 (小) B (チェリー材) を叩く音が木板 (中) B を叩く音に誤判定されていた．全体の結果は概ね 100% となったが．全ての特徴量で例の通りの誤判定があり，全体の識別率が低い特徴量ほど多く誤判定された．

データセット 2 では，誤判定の例として，プラケースを合板上に落下させる音がガススプレーを噴射する音に誤判定されていた．実験 1，4 では 99%, 実験 2，3 では 94 %程度の認識率であった．実験 1，4 では例の誤判定と，割りばしを割る音が誤判定された．実験 3，4 では更に摩擦系音源の紙ヤスリで木片を擦る音，破裂破壊系音源の割り箸を手で折る音，弾性音系音源のキャップを閉める音等の誤判定がみられた．

データセット 3 では，誤判定の例として，最も多かったのが本・雑誌を紙の上に落とす音が木板 (大) A (チーク材) を叩く音にも誤判定されていた．実験 1，4 では 94% 程度の識別率となり，衝突系音源 (木質) の 8 種類の音で相互誤判定が多い結果となった．実験 2 では 82.2%, 実験 3 では 73.6% となり，更に動作系音源 (摩擦系)5 種類の間で相互誤判定が多い結果となった．この実験では特徴量による違いが大きくみられ，MFCC,LPC, パワーピーク時のスペクトルの順でそれぞれ 10 %程度の差が開いた．

4.1.5 考察

実験 2 と実験 3 のデータセット 1 の結果において木板（小）を手で持ち木棒で叩く音が木板（大）を手で持ち木棒で叩く音に多く誤判定されていたが，これは実際に音を聞くと特にこの 2 つが似ていること，周波数の特徴が酷似していることが原因であると考察した．

実験 1 のデータセット 2 の結果において識別に失敗している音は割り箸を割る音であったが，音を観察したところ可聴部の音の長さがおよそ 0.05 s と短い．更に切り出される音の長さの平均が 0.48 s , 標準偏差 0.18 s に対して外れ値とみなされる 1.04 s 以上の音があることが原因であると考察した．

実験 3 のデータセット 2 の結果において失敗している例として紙ヤスリで木片を擦る音，割り箸を手で折る音，キャップを閉める音があったが，この 3 つの音に共通することは特定の周波数が強く出ない音であるということが言える．よってパワーピーク時のスペクトルだけでは判定が難しい．

データセット 3 においては全体的に衝突系音源 (木質) の 8 種類，動作系音源 (摩擦系)5 種類の間で相互誤判定が多い結果となった．これらの音は周波数にばらつきがあり，音によっては人間にも識別が難しい程度に周波数が酷似する音がある．このように人間にも判別が難しい音をどのように扱うかを今後考えていく必要がある．その他に誤判定が多かった音は，実験 1,2 については周波数にばらつきが大きいものや，特定の周波数が強く出ていない音であった．実験 3 において誤判定が多かった音は周波数は同じだが音の時系列変動が統一でないものであった．2 つの実験では，誤判定される音の種類に共通性が見られなかった．このことから，2 つの特徴量間では得手不得手があることが推測できる．MFCC にパワーの時系列変動を加えた場合では MFCC のみの場合に比べ，大差がなかったが 14 種類では僅かに識別率が低下したが，82 種類では僅かに識別率が向上した．このことからパワーを加えることで特徴量が冗長になる場合もあるが悪影響は少なく，識別の種類を増やすとパワーの特徴量が有効になってくると考えられる．

4.2 実環境音での実験

次に，実環境音での実験を行った．録音には kinect v2 を用いた．kinect v2 は 4 つのマイクを備え，16000Hz で録音可能な 4ch マイクロホンアレイとして利用することができる．手動で切り出しを行ったデータ，ロボット聴覚ソフトウェア HARK(Honda Research Institute Japan Audition for Robots with Kyoto University)[6] を用いて GSVD-MUSIC(Generalized Singular Value Decomposition - MUltiple SIgnal Classification) 法による音源定位 [7],GHDSS(Geometric High-order Decorrelation Source Separation) 法による音源分離 [8] をおこなったデータを利用し，実験を行った．

ロボット聴覚ソフトウェア HARK[6] はインターネットブラウザ上で GUI を用いて，簡単に音源分離，音源定位，音声認識の処理を記述することのできるソフトウェアで，インターネット上かで公開されている．今回このソフトウェアを使用した理由として，マイク配列に対応しており，環境音に対しても音源分離，音源定位を行うことができるという点が挙げられる．

本実験で用いるデータは概ね静かな居住宅で録音された，雑音を多少含むようなデータである．音を故意に発生させ，1 つの wav ファイルに 100 回録音し，それを音源分離または手動によって切り出した．データセットとして 20 種類の環境音を各

100 サンプル収集した．20 種類の環境音を表 2 に示す．また，今回の実験でも予備実験と同様な識別機，特徴量を用いた．

表 2　環境音録音データ

No.	使用した音
1	ビンをぶつける音
2	ビニール袋をこする音
3	CD ケース開閉
4	ファスナーを開閉する音
5	目覚まし時計の音（電子音）
6	食器棚を開閉する音
7	ドアを開閉する音
8	ドアのロックの開閉音
9	ホッチキスの使用音
10	スマートフォンのロック音
11	鍵が複数ついたキーホルダーを鳴らす音
12	キーボードを打つ音
13	窓のロックの開閉音
14	マジックテープを外す音
15	コップに水を注ぐ音
16	コイン入りのプラスチック貯金箱にコインを入れる
17	マウスのクリック音
18	お菓子の缶の開閉
19	テレビリモコンのスイッチ音
20	ゲームの起動音

4.2.1　結果

実環境音下での実験結果を表 3 に示す．全体的にビニール袋

表 3　実環境音での実験結果

	手動区間検出データ	音源分離データ
実験 1	98.9%	94.9%
実験 2	95.6%	87.6%
実験 3	91.4%	78.8%
実験 4	99.0%	95.7%

をこする音が誤判定が多い結果となった．自身で音を実際に聞いて識別可能な音であっても誤判定がなされていた．実際の環境音であっても MFCC とパワーパターンの組み合わせが最も識別率が高く，それに続いて僅かに MFCC が低くなり，LPC になると数パーセント識別率が落ち，パワーパターンとパワーピーク時のスペクトルの組み合わせで十パーセント程度識別率が低下する結果となった．

4.2.2　考察

無響室データでの結果を併せて結果を鑑みると，環境音識別分野においては，4 つの特徴量のうちで MFCC とパワーパターンを用いる場合が最も高い性能を発揮することができる．MFCC とパワーパターンとパワーピーク時のスペクトルを用いた場合の結果から特徴量が一瞬の周波数スペクトルを用いる場合よりもスペクトルの変化を用いる場合のほうが重要であると推定される．更に, 音声認識で有用なメル周波数つまり人間の聴覚特性は環境音においてもを考慮すべきであるということもいえる．MFCC とパワーパターンを組み合わせることで MFCC だけの場合より僅かに全体の識別率が向上した．つまり，環境音認識においてパワーパターンは重要な一要素であることが言える．その反面パワーパターンについては考慮すべき場合とそうでない場合がある．また，今回の実験においても誤判定が多かった音は予備実験のデータセット 3 と同じような音であり，これは雑音を含むようなデータであっても特徴量による識別可能音の傾向が同じである．

音源分離において識別率が低下した理由として区間検出が手動と比較し，不正確であったことが挙げられる．よって，雑音が少ないような実環境下では一つの音全体を正確に切り出すことのできる区間検出が望ましい．しかし，音の区間がはっきりしているような音であれば音源定位，音源分離を用いることでも高い識別率を確保することができるだろうと考えられる．

5　まとめ

本稿では RWCP-DB を用いた雑音のない無響室での環境音と雑音のあるような実環境音下で録音した環境音で識別率の比較を行い，無響室データ 83 種類で 94%，実際の環境音 20 種類で手動検出で 98.9%，音源分離で 94.9% と従来研究と比較して高い識別率を出すことができた．これは雑音に対して適切な処理を行うことで環境音識別においてある程度の実用化が見込めると考えられる．音源分離した場合の識別率が低下してしまった問題の解決策として音源定位の為の詳細なパラメーター，例えば部屋の正確なインパルス応答等を指定することによってある程度の解決は見込めるものの，それでは汎用的な識別システムの作成が難しく，この点の解決策を考える必要がある．音源定位では汎用性が少なく，音源分離以外の検出手法を考えなければならない．今回はある程度静かな環境で音を収集しているため，雑音の影響が少なかった可能性がある為，雑音の度合いや種類に応じて比較実験を行い，雑音に対する頑健性を確保できるようにしなければならないと考えられる．

また，今回の実験では 20 種類の環境音を用いたが，実用化するにはさらに多くの識別数が必要となるだろう．インターネット上のフリー素材で生活音を計測したところ少なくとも 100 種類は存在が確認できた．つまり，さらに数を増やすことでどの程度まで識別率が下がるか．また，どのようにしてそれを抑えるかを考えることが今後の課題である．更に，今回は故意に発生させた音を用いたが，実際に自然的に発生する場合のデータを収集したほうが良いと考えられる．その学習データをどのように収集するかも考える必要があるといえるだろう．

参考文献

[1] 大石, “あらゆる音の検出・識別を目指して –音響イベント検出研究の現在と未来–”, 音響学秋季研資, 3-8-1, pp. 1521-1524, Sep. 2014.

[2] Y.Toyoda, “Environmental Sound Reognition by Multilayered Neural Networks Computer and Information Technology”, in *Proc.CIT*, pp.123-127,2004.

[3] 吉田,”生活雑音の可視化に基づくヒト近接イベントの検出割合”, バイオメディカル・ファジィ・システム学会誌 VoLll，pp.53-62,2009.

[4] 比屋根, “ＲＷＣＰ実環境音声・音響データベース”, 人工知能誌, pp.2,2002.

[5] 赤木, “雑音・残響環境下における信号音抽出法に関する研究”, 平成 10-2 年度科学研究費補助金（基盤研究 (C)(2)）研究成果報告書.

[6] “ロボット聴覚オープンソースソフトウェア HARK の紹介 “, 計測制御, pp.1712-1716, 2014.

[7] K.Nakamura, “Real-time Super-resolution Sound Source Localization for Robots”, in *Proc.IROS*, pp.694-699, 2012.

[8] K.Nakadai, “Robot audition for dynamic environments”, in *Proc.ICSPCC*, pp.125-130, 2012.