3V-08

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*

山本修生（法政大学 情報科学部）, 伊藤克亘（法政大学 情報科学部）

1 まえがき

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

実用的な識別機には速度，精度が求められる．しかし，それらを達成するには様々な問題が存在する [1]．音シーンの認識について 4 つの課題があり，1 つ目の課題として同一音の検出，識別がある．データベースを用いて同一音を識別するという問題ある．2 つ目の課題として音響イベントの検出，識別がある．環境音を音響イベントとして定義し，音響的特徴から同一の音を識別する．3 つ目の課題として音区間の検出，識別がある．環境音の区間を識別する問題である．音声認識分野においても音声とそれ以外の音を識別し，音声信号のある区間を検出する音声区間検出の研究が長年行われているという．環境音においては複数の音が重なることが予想される．4 つ目の課題として音環境の検出，識別がある．場所や出来事を定義しておき，識別する問題である．本研究ではこれらのことを考慮しながら実験を進めていく．

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] で紹介されている手法としては，ニューラルネットワーク (NN)，サポートベクターマシン (SVM)，ガウス混合分布，隠れマルコフモデル (HMM) 等が挙げられる．今回の実験では主に，手法としては主に音声認識においてスマートフォンの音声認識などに用いられ，一定の成果を上げているニューラルネットワークを利用する．

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。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．

特徴量は，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] で紹介されている研究の特徴量として MFCC(メル周波数ケプストラム係数) があり，これを利用することで最も高い識別率を出している研究があることからこれを利用する．

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る．そこで本研究では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を検討した [5]．

3 RWCP-DB での実験

実環境を考慮する前に背景雑音なしでの実験を行った．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．

3.1 事前実験

基本的に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．

3.1.1 事前実験結果

結果として平均は 0.618 秒，最高長は 3.56 秒，最低長は 0.03 秒，1 秒以上のものは 17 個で 1 秒以下のものが 84 %となり，1.5 秒以上のものは 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 秒．最も長い音はコイン 1 個を合板に落とす音で 3.5648 秒であった．

3.1.2 事前実験考察

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 0.6 秒を基準に特徴量として実験を行った．

*An environmental sounds recognition under real environment by neural network.: *Naoki Yamamoto (Hosei Univ.) et al.*

3.2 予備実験

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で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．

実験1では特徴量としては参考論文[3]で用いられており，実績のあるMFCC(メル周波数ケプストラム係数)を用いることとした．本研究ではRWCPデータベースの平均近くの値である600msを区間としたフレーム長を50msとした13次のMFCCを25msでシフトしたものを横に並べた299次元の配列データを用いることを提案する．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い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では特徴量として参考論文[2]で用いられている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．パワーパターンは実験Aと同じくRWCPデータベースの平均付近の600msを用いてフレーム40msシフト20msで28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28点を結合した156次元の配列データとした．

データセットはRWCPデータベースで100サンプルある音83種類を使用した．また，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．

3.3 実験結果

10回の学習結果の平均の実験結果は実験1では92.4%実験2では73.6%となった．この実験ではこの2つの特徴量による違いが大きくみられ，20 %程度の差が開いた

3.4 実験考察

．実験1において誤判定が多かった音は，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2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．

4 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で録音可能なマイクロホンアレイとして利用することができる。そのままのデータ，ロボット聴覚HARKを用いて音源分離をおこなったものを利用し，実験を行った．用いるデータは概ね静かな部屋で録音された，雑音を多少含むようなデータである．データセットとして15種類の環境音を各100サンプル収集し，それを手動で区間検出を行ったもの、音源分離したものの2種類のデータを用いた．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．

4.1 実験結果

実環境音下での実験結果を表2に示す．無響室データとは識別数が異なるため、単純比較はできないものの，2つのデータセットで90%以上の識別率を出すことができた．MFCCを並べた特徴量の場合，実環境音下で音源分離によって区間検出した場合であっても過去研究での無響室での実験[2]より高い識別率を出している．音源分離では手動で区間検出を行った場合と比較し，6 13%程度識別率が低下した．

表 1. 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	99.6%	93.4%
実験2	98.6%	85.9%

4.2 実験考察

4つの実験から、パワーとパワーピーク時のスペクトルを用いる場合よりもMFCCを用いるほうが、環境音識別において高い性能を発揮することができると考えられる。これは環境音が一瞬の周波数スペクトルよりもスペクトルの変化のほうが重要であると推定される．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．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と同じような音であった．

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，それぞれ99%から93%程度の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部屋の正確なインパルス応答を計測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．また，今回の実験では15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．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といえる．

参考文献

[1] 大石康智, ”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–”, 音響学秋季研究, 3-8-1, pp. 1521-1524, Sep. 2014.

[2] Y. Toyoda ,et al., ”Yong LiuEnvironmental Sound Reognition by Multilayered Neural Networks Computer and Information Technology”,IEEE CIT ’04., pp.123-127, Sep. 2004

[3] Andriy Temko, et al., ”ACOUSTIC EVENT DETECTION AND CLASSIFICATION IN SMART-ROOM ENVIRONMENTS: EVALUATION OF CHIL PROJECT SYSTEMS”, Computers in the Human Interaction Loop, ISBN 978-1-84882-053-1, pp.61-73, 2009.

[4] 比屋根 一雄 他,” ＲＷＣＰ実環境音声・音響データベース”, 人工知能誌,pp.2,2002

[5] 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