CJ-001

音ライフログに向けた NMF に基づく特徴量による環境音識別
Enviromental Sound Classification based on NMF toward Sound Life-log

秋庭裕 † Yutaka Akiba
大川茂樹 † Shigeki Okawa

1. 背景

個人の生活・体験・状態を記録することやその記録のことをライフログと呼ぶ．近年になってストレージの大容量化により，大容量のデータを様々な記録メディアで蓄えることが現実的になった．

1.1. 音ライフログ

本研究では，生活における様々な情報の中でも音に着目した音ライフログを取り扱う．音ライフログは，マイクロフォンで環境音や音声を連続して録音することで，様々な生活の中で起きる音響イベントをライフログとして保存するものである．近年ではスマートフォンが普及したことにより，内蔵マイクを利用して音ライフログを容易に行えるようになった．また，録音する行為自体は非常に手軽であるため，音ライフログは実用的であると考える．一方で，音ライフログを実用的にするための課題は，以下の 3 点が挙げられる．

1. 所望の情報が記録された区間を探し出す必要がある．
2. 所望の情報にアクセスするためには，聞き返す必要がある．
3. 言語的なクエリによる検索ができない．

このような問題点を解決するためには，まず音データから情報をマイニングし再利用しやすい形で保存しておく必要がある．次にユーザが意識せずともライフログが記録され利用できる必要がある．そこで，図 1 に示すような音ライフログのエコシステムを提案する．

本研究では，どの時刻に，どの音クラスの音が含まれているのかということが音ライフログに関する重要な問題であると位置づける．この問題に対して，各音クラスにおいて頻出するスペクトルパターンを NMF によるクラスタリングによって抽出し特徴量とすることで音クラスの識別を行う．これは，図 1 における Life-log System での処理に相当する．

こういった問題に対する研究報告も多くある．まず MFCC など意図的な特徴量による環境音識別の報告がある [1]．一方で，意図的な特徴量では信号の特定の特徴のみ強調されてしまい，柔軟性に欠ける面がある．そこで，鳥羽らの CNN を用いた環境音分類 [2] やÇakır らの CRNN による鳥のさえずり検知 [3] のように深層学習を用いた研究がある．他には，Bisot らは DCASE 2016 にて TDL [4] の非負拡張である Nonnegative TDL を提案し，教師あり特徴学習による音響シーン識別を行っている [5]．本研究でも，NMF による特徴学習によって柔軟性のある特徴量を用いて環境音識別を行う．

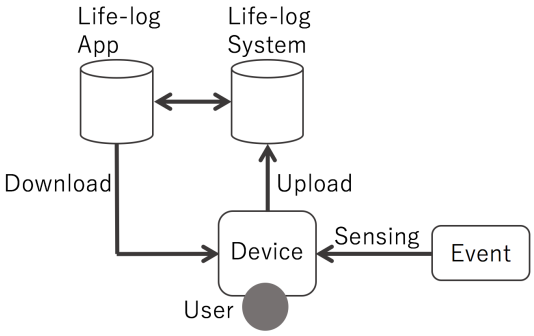

図 1　音ライフログのエコシステム

2. 特徴量抽出

音ライフログとして録音されたデータには様々な音クラスが含まれているだけではなく重畳している場合も多い．しかし，この重畳の組み合わせを教師データとして与えて識別器を学習させるのは現実的ではない．そこで，重畳している音を分離し特徴を抽出することで，重畳している個々の音を識別することを試みる．

2.1. 非負値行列因子分解

非負値行列因子分解 (Nonnegative matrix factorization; NMF) [6, 7] は，非負値データをその因子行列に分解する手法として音源分離など様々な分野で注目されている．音データのスペクトログラム表現を非負値行列として NMF に入力することでスペクトログラムに含まれている共起する成分が基底として取り出される．すなわち，データ構造に基づいた分解が行われるため，データの特徴を教師なし学習している効果が得られる．ここで，行列 $\boldsymbol{V} \in \mathbb{R}_{+}^{P\times N}$ を行列 $\boldsymbol{W} \in \mathbb{R}_{+}^{P\times K}$ と行列 $\boldsymbol{H} \in \mathbb{R}_{+}^{K\times N}$ の積として因子分解する問題は，以下の行列 $\boldsymbol{V}$ と行列積 $\boldsymbol{WH}$ の距離最小化問題を解けば良い．

$$\boldsymbol{W}, \boldsymbol{H} = \underset{\boldsymbol{W},\boldsymbol{H}}{\operatorname{argmin}} D_{\beta}(\boldsymbol{V}|\boldsymbol{WH}) \qquad (1)$$

ここで，一般的に $\boldsymbol{W}$ は基底行列，$\boldsymbol{H}$ はアクティベーション行列と呼ばれる．また，D_{β} は β ダイバージェンスである．本研究では，KL ダイバージェンスを用いたため $\beta = 1$ である．

3. 音クラス識別システム

提案する音クラス識別システムの概要を図 2 に示す．本システムでは，まず純粋に 1 つの音クラスのみが含

† 千葉工業大学工学研究科未来ロボティクス専攻

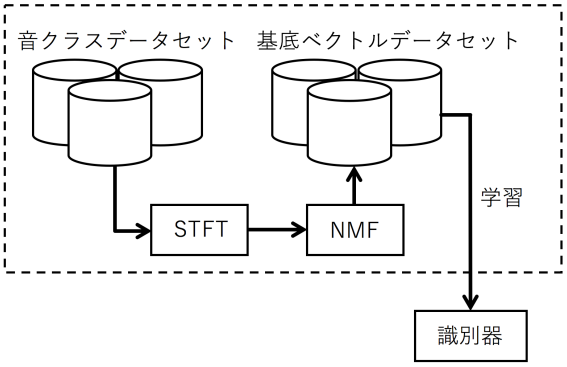

図 2　提案する音クラス識別システムの概要

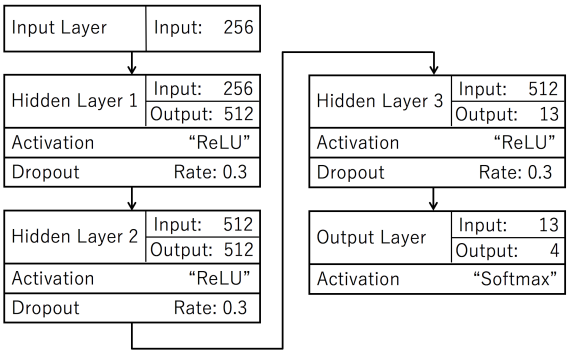

図 3　MLP モデルの構成

まれる音データを集めて音クラスデータセットを作成する．次に，短時間フーリエ変換 (STFT) によってスペクトログラム表現に変換し NMF に入力することで基底ベクトルを抽出する．この操作によって，基底ベクトルデータセットが作成される．基底ベクトルは NMF の基底数分だけ抽出される．そのため，基底ベクトルデータセットは音クラスデータセットの基底数倍のサイズになる．この基底ベクトルデータセットを用いることで識別器を学習させる．また，識別器には 4 層の多層パーセプトロン (Multi-Layer Perceptron; MLP) を用いた．本研究で設定した MLP モデルの構成を図 3 に示す．

3.1. 学習データとテストデータ

本研究での実験では音クラスは，{ 音声，車の走行音，電車の走行音，流水音 } の計 4 クラスとした．(ここうまく)

また，本研究で用いた音データに関して音クラス毎に合計した時間長とファイル数を表 1 に示す．それぞれの音データファイルは，異なる状況や環境で収録されたものであり，なるべく単音のみ収録されているものを選んだ．音声以外の音データは，それぞれ YouTube や Orange Free Sounds などからダウンロードして得たものである．この際，ライセンスが CC BY-NC または CC BY であるものを選定した．音声データに関しては，PASL-DSR 連続音声データベースを用いた．

表 1　本研究で用いた音データの時間長とファイル数

音クラス	時間長 [s]	ファイル数 [個]
車の走行音	02:44:02	30
音声	01:28:47	1104
電車の走行音	02:36:13	29
流水音	03:18:29	45

3.2. 音クラス識別実験

提案する音クラス識別システムによって音クラスをどの程度の性能で識別可能か評価実験を行った．この性能評価尺度には，Precision，Recall，F 値の 3 つの値を用いた．Precision，Recall，F 値は，表 2 においてそれぞれ次のように計算される．

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F-score} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

つまり，Precision は正例として予測されたサンプルのなかでどれだけ真の結果と一致していたかを表しており，Recall は真の結果が正例であるサンプルをどれだけ正例として予測できたかを表している．F 値は，それらの調和平均であり，本研究では F 値を元に汎化性能を評価する．また，この評価実験の目的には，後に説明する混合音識別実験にて用いる識別モデルを選定することも含まれている．そこで，まずは 5 分割交差検証により汎化性能の高い識別器を選定した．

提案システムでは，NMF によって基底ベクトルを特徴量として抽出するが，このとき音データは 1 秒毎に入力する．また，NMF の基底数は 10 とした．つまり，1 秒ごとに基底ベクトルが 10 個出力され，それぞれに元の音クラスを教師データとして付与するということである．ゆえに，たとえば車の走行音クラスの総時間長が 2 時間 44 分 2 秒であることから，基底ベクトルは 98420 $(= (2 \times 60^2 + 44 \times 60 + 2) \times 10)$ 個抽出される．また，交差検証はベクトル単位ではなくファイル単位で行うものとする．これは同じファイルに収録されている音から抽出される基底ベクトルには関連があり，同じファイルから抽出された特徴量を交差検証しても汎化性能を評価するには適していないと考えられるためである．

3.3. 識別実験結果とその考察

5 分割交差検証により，5 つの結果を得て各項目の平均をとった結果を図 4 に示す．この図を見るとわかるように，電車の走行音クラスに関する評価値が軒並み低い結果となった．しかし，その他の音クラスに関しては車の走行音の F 値が 0.6 を下回っているものの音声と

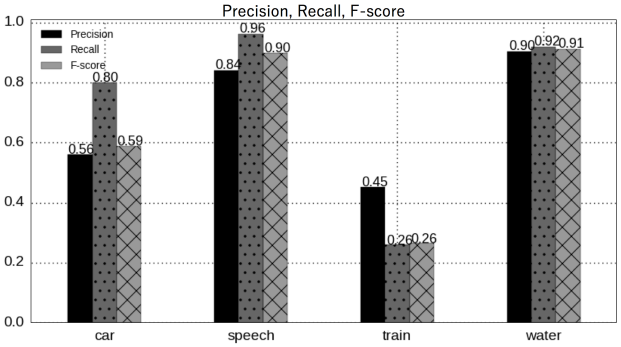

図 4　5 分割交差検証によって得た各評価値の平均

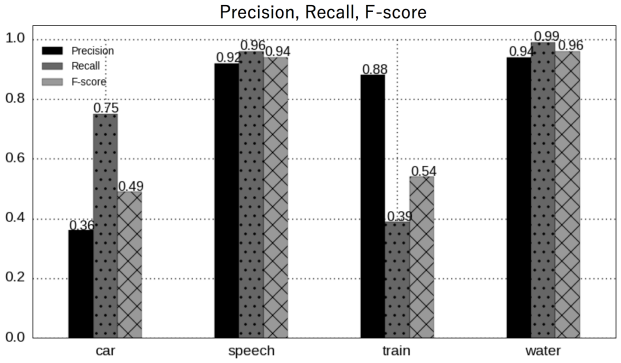

図 5　混合音識別実験に用いた識別器の評価値

流水音は F 値が 0.8 を超える識別精度となった．

電車と車の走行音クラスの結果の原因は，電車や車の走行音には様々なパターンがあり学習データとして選ばれたデータではテストデータを予測しづらいケースが多かったためと考えられ，基底ベクトルを特徴量とした場合学習データの網羅性が重要だと考えられる．

最後に，図 5 は，5 分割交差検証においてもっとも F 値の平均が高かった識別器の評価値である．この識別器を今後の実験で用いるものとする．

4. 音ライフログシステム実験

前節まで，各音クラスの識別について検討した結果を説明してきたが，冒頭でも述べた通り音ライフログのためには音が重畳している区間に対応する必要がある．そこで，混合音の入力と結果の出力を工夫して図 6 のような混合音に対応した音ライフログシステムを提案する．

まず，単クラスの識別実験と同様に混合音を 1 秒ごとに区切り，スペクトログラム表現に変換して NMF によって基底ベクトルを抽出する．複数の音クラスの数は

表 2　混合行列

		真の結果	
		正例	負例
予測	正例	True Positive	False Positive
	負例	False Negative	True Negative

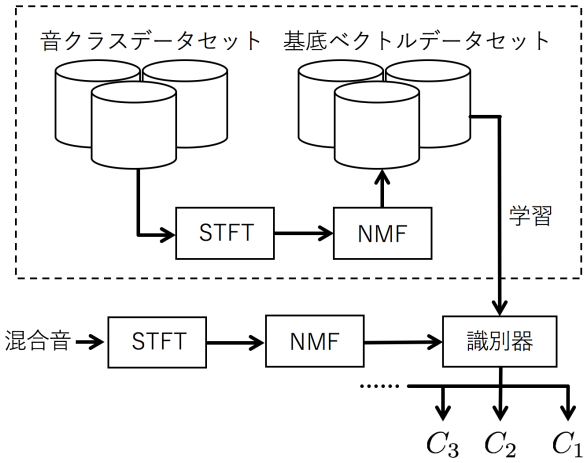

図 6　提案する音ライフログシステムの概要

NMF の基底数に依存しており，本研究では 5 としたため各区間で基底ベクトルは 5 つ抽出される．その各基底ベクトルそれぞれを識別器は評価するため，予測音クラスは 5 つ出力される．そのため，重複するクラスは 1 つにまとめる．たとえば，$[C_1, C_1, C_2, C_1, C_2]$ という結果だった場合には，ユニークなクラスを取り出して $\{C_1, C_2\}$ という集合を評価とする．このようにして，図 6 に示すようなライフログシステムに拡張する．このシステムを用いて混合音を入力した際の評価実験を行った．

4.1. 実験用混合音データ

混合音データは，5 分割交差検証においてテストデータとして隔離したデータを用いて作成した．そのデータセットにおいて各クラスの全音データを連結し，1 つの音データとしてから 180 秒ずつに分割する．これにより，各クラスそれぞれで 180 秒の音信号を要素とする集合 S_c, S_s, S_t, S_w を得る．次に，これら 4 つの集合から 2 つを選ぶため 6 $(= {}_4\mathrm{C}_2)$ 組に関して，デカルト積を考えその和集合を取ることで考え得る混合音の全パターンの集合 M を得る．そして，この集合 M に従って，2 つの音を重複させることで混合音を作成した．基底数は 5 としているため，重複している音クラスを識別できる最大の数も 5 である．本研究では簡単のため混合する音の最大数は 2 とした．

また，作成した混合音は前半後半の 90 秒は 1 つの音クラスのみであり，中間 90 秒間だけ 2 つの音クラスが重畳しているように作成した．すなわち，全体で 4 分 30 秒の時間長である音データである．

4.2. 評価方法

混合音にあらかじめ正解データを与えておき，その正解データと一致しているかどうかで評価した．1 秒ごとに評価が行われるため，その度に正答しているかどうかを評価する．正答した秒数を PINT (Positive Interval),

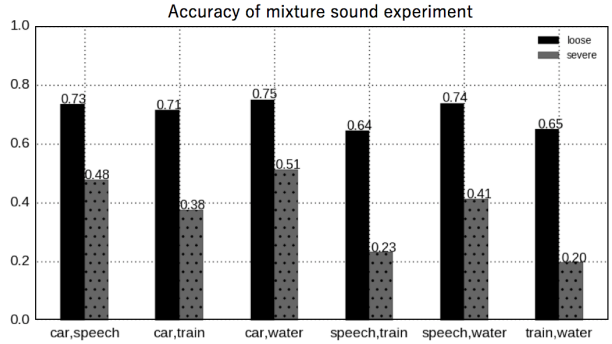

図 7　混合音識別実験の結果

不正解であった秒数を NINT (Negative Interval) とするとき，以下のように Accuracy を定義する．

$$\text{Accuracy} = \frac{\text{PINT}}{\text{PINT} + \text{NINT}} \tag{5}$$

また，正答には Loose と Severe の 2 つの定義を設けた．ここで，予測クラス集合 C_{pred} と正解クラス集合 C_{corr} を考える．このとき，以下のような 2 つの条件を考える．

$$C_{\text{corr}} = C_{\text{pred}} \tag{6}$$
$$C_{\text{corr}} \in C_{\text{pred}} \tag{7}$$

式 (6) を満たすとき Loose な正答とし，式 (7) を満たすとき Severe な正答とする．Loose な条件の場合には，予測結果の集合に毎回すべてのクラスが含まれるように出力すれば，Accuracy は 100 % となるため参考評価値である．それに対し，Severe な条件ではランダムにクラスを 5 つ出力し，そのユニークな値を予測結果とするダミー識別器により同様の実験をサンプル 1000 個行うと，0.02 程度の性能となる．本研究では，このダミーテストを評価のベースラインとする．

4.3. 実験結果と考察

混合音に関する音クラス識別実験の結果を図 7 に示す．これは，各組み合わせにおける評価値を平均したものである．その結果，車の走行音と流水音の組み合わせに関して Severe Accuracy が 0.51 という結果を得た．他の組み合わせに関してもダミーテストの 0.02 と比べて優位な差があり，意味のある結果と考える．

電車の走行音が重畳する組み合わせに関して他と比べて評価値が低い傾向が見られる．これは，採用した識別器の性能が原因であると考えられる．その一方で，共に識別精度が高かった音声と流水音の組み合わせに関して，他の組み合わせと比べて評価値に大きな差がないことから単純に識別器の性能を向上させることでは混合音に対応できないことがわかる．

5. まとめ

どの時刻に，どの音クラスが含まれているのかという問題に対して，音クラス識別システムとそれを用いた音ライフログシステムを提案し実験と評価を行った．提案音クラス識別システムは，NMF による教師なし学習により基底ベクトルを抽出し，それを MLP の入力として学習を行い音クラス識別を行う．評価実験では，音声と流水音に関して高い評価値を得ることができ，学習データによって性能が変わることから学習データを密にすることで性能を上げることが可能であると考える．音ライフログシステムは，音クラス識別システムを混合音に対応できるように拡張したものである．この評価実験では，ベースラインとなるダミーテストの評価値と比べることで意味があると考えられる結果を得ることができた．

今後は，本研究で特徴抽出するために用いた NMF や提案した混合音への処理方法に，音響信号の時間方向への構造情報を考慮した処理を加える．

謝辞

本研究は，JSPS 科研費 JP 26330203 の助成を受けた．

参考文献

[1] 古谷崇拓, 千葉祐弥, 能勢隆, 伊藤彰則. “日常音識別による活動記録自動生成のためのデータの収集と分析,” 情処研究報告, Vol.2017-MUS-115, No. 28, (2017).

[2] 鳥羽隼司, 原直, 阿部匡伸. “スマートフォンで収録した環境音データベースを用いた CNN による環境音分類,” 音講論集, pp.139-142, (2017.3).

[3] E. Çakır, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, “Convolutional Recurrent Neural Networks for Bird Audio Detection,” arXiv preprint arXiv:1703.02317v1, 2017.

[4] J. Mairal, F. Bach, and J. Ponce, “Task-driven dictionary learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (4), pp. 791-804, (2012).

[5] V. Bisot, R. Serizel, S. Essid, and G. Richard. “Supervised Nonnegative Matrix Factorization for Acoustic Scene Classification,” Detection and Classification of Acoustic Scenes and Events Challenge, (2016).

[6] D. D. Lee, H. S. Seung. “Learning the parts of objects with nonnegative matrix factorization,” Nature, 401(6755), pp. 788-791, (1999).

[7] D. D. Lee, H. S. Seung. “Algorithms for nonnegative matrix factorization,” Advances in Neural Information Processing Systems 13, pp. 556-562, (2001).