Vol.2017-MUS-115 No.28
2017/6/17

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

古谷 崇拓[1,a)]　千葉 祐弥[1]　能勢 隆[1]　伊藤 彰則[1]

概要：スマートフォンなどで個人の日々の生活を記録するライフログが注目を集めている．収集されたライフログデータの活用方法として，ユーザの行動記録の自動生成がある．従来の研究では，設定した行動クラスの識別手法に焦点を当てるのが一般的であり，取得された認識対象クラスの類似性や検出窓幅の妥当性など，基礎的な分析や検討が十分ではなかった．そこで，本稿では音響情報を利用したユーザの行動の認識を対象として，特徴量が考慮するべき時間幅の検討と，認識対象クラスの音響的な類似性に関して分析を行った．実験データは 1 名の被験者により収集された約 22 時間分の日常音データであり，行動対象クラスは収録時の行動に基づいて人手で決定した．音響特徴量として MFCC (Mel-Frequency Cepstral Coeficients) を採用し，窓幅を変更しながら実際に識別実験を行うことで，特徴量抽出の窓幅について分析した．結果として，窓幅を大きくするほど識別性能が向上し，窓幅 8.0 [sec] としたとき最大の識別性能が得られた．また，階層的クラスタリングによって音響特徴量空間上の認識対象クラスの距離を可視化した．その結果，今回収集した行動クラスの音響データは，人の声が含まれるグループと背景雑音が比較的小さいグループ，背景雑音が大きいグループの 3 つに分類できることが確認され，特に音声クラスにおける識別結果の混同が大きいことが示唆された．

1. はじめに

近年のスマートフォンの普及により，日々の活動の記録やその活用が容易になった．自身の活動の記録を補助するアプリケーションの一つとして，ライフログがある．ライフログとは，使用者がいつ，どんな行動に，どれくらいの時間を費やしたのかを記録するものである．ライフログデータを有効に活用することで，その人にとって重要度の高いタスクにより多くの時間を割り振るといった，日々の生活の見直しの効率化が期待できる．また，過去に起こったイベントと，記憶との効果的なつながりをつくることで，記憶能力の向上にも有用であることが報告されている [1]．

一方，現在一般的に利用可能なライフログアプリは，いずれも手動でその日の自分の行動を入力しなければならず，利用者の心理的，時間的な負担になっている．そのため，ウェアラブルデバイスで収録した音や画像などのライフログデータから人間の行動を推定することで，活動記録を自動生成できることが望ましい．ユーザの行動クラスの推定に関しては，動画像情報を用いる手法 [2]，音情報と加速度情報を用いる手法 [3] など，多くの試みがなされてきた．文献 [2] の動画像情報を用いる手法では，SenseCam と呼ばれるウェアラブルカメラを用い，50 秒に一回撮影された画像を用いて，“Eating” や “Drinking” などの計 23 の着用者の行動クラスを対象とし，識別器に HMM(Hidden Markov Model) を用いることで，全行動クラスのうち 16 のクラスで平均の F 値 0.9 の識別率を達成している．文献 [3] の音情報と加速度情報を用いる手法では，Nishida らによって収集された被験者の日常生活を加速度信号と環境音を通して収録されたコーパス [4] を用いて，“Cleaning” や “Cooking” など計 9 の行動クラスの識別において，識別器に DNN (Deep Neural Network) を用い，窓幅，シフト幅ともに 1 [sec] とした音響特徴量の，前後 5 フレームを含めた 11 フレームでフレーム処理して逐次認識し，60 秒単位の識別率で 86.3%を達成している．文献 [3] の実験では，識別器に DNN を用いる他に，比較対象として SVM (Support Vector Machine) も用いられている．その結果，前述の 9 の行動クラスの他に，認識対象外行動の Other クラスを加えた計 10 の行動クラスの識別実験において，60 [sec] 単位の認識率で，DNN が SVM を約 5%上回るという結果を得ている．しかしながら，従来の研究では，設定した行動クラスの識別手法に焦点を当てており，認識対象とする行動クラスの背景雑音の有無などの音響的性質や音響特徴量抽出の際の窓幅の妥当性など，基礎的な分析や検討

[1] 東北大学 大学院工学研究科
Graduate School of Engineering, Tohoku University
[a)] furuya@spcom.ecei.tohoku.ac.jp

が十分ではなかった．また，特に文献 [3] において，音情報から行動クラスの識別をするという観点からは，識別対象の行動クラス数が十分ではなかった．そこで，本研究では，音響特徴量抽出の際の適切な窓幅の検討と，行動クラス全体の認識率向上の考察を得るため，認識対象クラスの音響的類似性に関して分析を行う．識別に用いる情報としては，多くのウェアラブル端末で利用が可能な音響情報に着目する．音響データによる行動識別では，カメラと比較して死角の影響を受けにくいこと，会話に含まれる幅広い情報を扱うことができるといった利点がある．

本研究ではまず，一人の被験者による日常音データの収集を行い，人手で付与した行動クラスのラベルを用いて識別実験を行った．識別実験では文献 [3] の手法を参考に，識別器としてフルコネクションの DNN を利用し，1 フレーム毎の認識を実施した．この際，特徴量抽出の窓幅を変化させることで，適切な窓幅の検討を行った．その後，各行動クラスから抽出された音響特徴量の，各クラスの平均値に対して階層的クラスタリングを行うことで，認識対象クラスの類似性を可視化し，識別に有用な特徴量について議論を行う．

2. 関連研究

本節では，音響情報による行動識別に限って説明する．日常音の分類に関しては，典型的なタスクとして，音響イベント分類，SAD (Speech Activity Detection) などがある．音響イベント分類は，連続した環境音から，人が歩く音などの対象とする音響イベントを分類するタスクであり，SAD は同じく連続した環境音から人が話している区間を検出するタスクである．Ohishi らや Peng らによって，音響信号を用いた音響イベントの検出が行われた [5, 6]．同じく音響イベント分類に関して，Temko らは実環境で録音された連続音の中から，ドアを閉める音や咳をする音などの特定の音響イベントの発生時間と音の種類を，SVM を用いて識別した [7]．また，Zhuang らは同じデータセットを用いて，音響イベント検出に有用な特徴量を用いて SVM-GMM-supervector [8] を学習することで，識別精度が 10%以上向上することを示している [9]．一方，SAD に関しても多くの研究がなされており，ノイズに頑強な教師無しの SAD として，Combo-SAD がある [10]．また，GABOR や Combo など様々な特徴量を組み合わせることで，ノイズが多く混じった会話音に SAD を適用し，ノイズに対する頑強さを向上する手法 [11] も提案されている．

一方，日常音を利用した人間の行動認識に関して，Ziaei らは SAD やキーワードマッチングなどを併用することで，部局ミーティングや研究ミーティングなどの状況の識別を行った [12]．この研究では，計 30 日分の仕事の様子をポータブルデバイスで録音した ProfLifeLog コーパスを利用した [13]．また，林らは高齢者の生活行動見守りシステムの構築を目標として，大学生一名の 3 日間のマンションでの生活の様子を収録したデータを用いて，音データと加速度信号からユーザの行動の識別を行った [3]．

識別手法に関しては，音響情報の分類においても DNN の利用が盛んであり，例えば，環境音分類においてはスマートフォンで収録された音声から得られたスペクトログラムの 2 次元画像を入力として使用し，CNN (Convolutional Neural Network) で識別を行う研究がある [14]．ユーザの行動認識に関しても，DNN による識別を行うことで高い識別性能が得られることが報告されている [3]．

上記の研究のうち日常音の識別を行っているのは [3, 12–14] であり，かつウェアラブルデバイス着用者の行動の識別を行っているのは [3] であるが，これらの研究では，主に識別手法に関して焦点が当てられており，認識対象クラスの音響的な類似性や，音響特徴量の検討も十分に行われていない．そのため本研究では，ウェアラブルデバイス着用者の行動を識別するのに必要な日常音の収集を行い，かつ識別実験において特徴量抽出の窓幅を変化させることで，適切な窓幅についても検討する．また，各認識対象クラスの類似性を階層的クラスタリングによって可視化し，識別に有用な特徴量について議論する．ここで，認識対象のクラスはユーザの生活行動見直しといったアプリケーションに適したものを採用した．

3. 実験に用いる日常音データの収録

実験用データは男性 1 名が収録した．データの収録にはウェアラブルカメラ (Panasonic, HX-A100) を利用した．被験者は付属のイヤーフックを用いてウェアラブルカメラとマイクロフォンを耳元に装着し，食事や家事などの日常的な活動を行った．被験者は 2，3 時間の連続撮影を 10 日間続け，約 22 時間の音声画像データを収録した．動画像データは MPEG4 形式で保存し，音声データは 2 チャンネル，サンプリング周波数 48 kHz，16 bit 量子化で収録した．録音場所は，大学の研究室やカフェテリア，被験者の自宅などであった．識別実験には，動画像データから切り出した音響データを，16 kHz にダウンサンプリングしたものを用いる．

データの収録後，被験者自身が行動クラスのアノテーションを行った．アノテーションは動画像に対して実施し，被験者は収録した動画像データを観察しながら行動クラスを付与した．このとき，アノテーションのラベルは先行研究 [2, 3] を参考に，日々の生活行動見直しのためのアプリケーションに適したクラスを設定した．アノテーションによって付与された行動クラスのラベルは 48 種類であった．行動クラスのラベルの一部を表 1 に示す．ここで，動画像のアノテーションには ELAN [15] を利用した．

表 1　窓幅 8 [sec] のときの認識対象とする日常音クラス
Table 1　Target classes of everyday sound when window size is 8 [sec].

Index	Class name	Notation	Index	Class name	Notation
1	desk work	DSW	13	play with smartphone	PLS
2	Q&A in presentation (listen)	QAL	14	Q&A in presentation (give)	QAG
3	talking	TLK	15	rest	RST
4	presentation (listen)	PRL	16	announcement	ANC
5	cooking	CKG	17	shopping	SHP
6	meeting	MTG	18	presentation (give)	PRG
7	transfer inside a building	TIB	19	wait for something	WTS
8	eating lunch	ETL	20	washing something	WGS
9	English lesson	EGL	21	teaching	TCG
10	transfer outside	TOT	22	preparation of presentation (listen)	PPL
11	sleeping	SLP	23	study	STD
12	speaking	SPK	24	office work	OFW

4. 日常音に基づく行動クラスの識別

4.1 特徴量抽出

本研究では，文献 [3] を参考に，音響特徴量を用いた行動クラスの識別を行う．識別実験の条件を表 2 にまとめた．行動クラスの識別は 60 [sec] の音響信号を 1 サンプルとしたサンプル単位で実施する．収録した日常音データはクラス毎に分割し，分割した日常音データから指定した窓幅とフレームシフトで特徴量を抽出した．日常音データをクラス毎に分割する際，窓幅に満たないクラスの場合はその後続のクラスを含めて，窓幅と同じ時間になるように分割している．そのとき，窓幅に足りないクラスが音声ファイルの終端にあって，その後続のクラスの音を含めることができない場合は，そのデータは除いている．特徴量としては，パワーを含む下位 12 次元の MFCC とその Δ，Δ^2 係数の，計 39 次元の特徴量を利用する．学習時には，抽出された特徴量をフレームごとに識別器に入力する．一方，認識時にはフレームごとに行動クラスの推定を行ない，サンプル内の各フレームの推定結果の多数決によって当該サンプルの行動クラスを決定する．

4.2 識別器の学習

前述の通り，識別器には DNN を用いた．ネットワークの構造は，表 3 のように設定した．中間層数とノード数は先行研究 [3] に準拠した．フレームごとに識別を行うため，入力ノード数は音響特徴量の次元数と等しく，出力ノード数は認識対象とする行動クラス数と等しい．

識別結果は 4-fold Cross Validation で評価した．各クラスのデータはサンプル単位でランダムに 4 分割し，1 つの fold をテストデータとし，それ以外のデータを学習データとした．60 [sec] ごとに 1 サンプルとしているため，継続時間が合計で 240 [sec] に満たないクラスは認識対象から除外した．表 1 に窓幅 8.0 [sec] のときの認識対象クラスをまとめた．表 1 において，Class Name は認識対象行動クラス名，Notation はそれぞれのクラスを表す記号である．

表 2　識別実験の条件
Table 2　Experimantal condition.

実験データ	全行動クラス合わせて計約 22 時間
学習形式	4-fold Cross Validation
音響特徴量	(MFCC12 次元+power)+Δ+$\Delta\Delta$
サンプリング周波数	16 kHz

表 3　DNN の構造
Table 3　Structure of neural network.

入力ノード数	39
中間層数	5
中間層のノード数	2048
出力ノード数	24
活性化関数	ReLU
学習エポック数	500

ここで，収集されたデータは，クラスごとに出現頻度が大きく異なる．したがって，実験では，データを学習データとテストデータに分けた後，最も学習データ量が多いクラスに合わせてアップサンプリングを行うことで，学習データにおける各クラスの出現頻度を均一化した．識別性能は各クラスの識別率に対して出現頻度を重みとした重み付き平均 (WA, Weighted Average) と単純平均 (UA, Unweighted Average) の 2 つで評価した．サンプル単位の識別結果を SAMPLE，フレーム単位の識別結果を FRAME と記述する．

4.3 特徴量抽出における窓幅の検証

まず，特徴量抽出の窓幅を変えて識別実験を行うことで，窓幅の影響を調査した．結果はフレームごとの識別実験によって評価した．窓幅を 0.5，1.0，...，16.0 [sec] と変化させたときの UA (FRAME) の推移を図 1 に示す．本実験ではシフト幅を窓幅の半分に固定した．特徴量抽出において，

表 4　各窓幅における出力ノード数
Table 4　Number of output nodes to window size.

窓幅 [sec]	0.5	1	2	4	8	16
ノード数	27	27	26	25	24	24

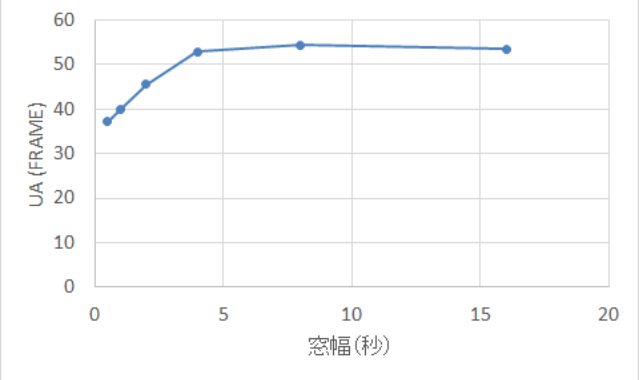

図 1　特徴量抽出における窓幅に対する識別率の変化
Fig. 1　Discrimination result to window size.

窓幅の長さに足りないデータは除き，かつ識別器の学習において 4 分割ができなかったクラスに関しては除外しているため，窓幅を変化させる際は条件によって出力ノード数が異なる．表 4 に，各窓幅のときの出力ノード数を示す．

結果より，窓幅を大きくすることで，フレームごとの識別性能が向上することがわかる．また，窓幅を 8.0 [sec] としたとき最も性能が高く，54.3% の識別性能が得られた．窓幅を大きくすることによって認識率が増加する理由として，各クラスを特徴付ける音響情報が 1 つのフレームに含まれやすくなることが挙げられる．一方，窓幅を 16.0 [sec] とすると UA (FRAME) は減少したが，これは今回の実験では，日常音データをクラス毎に分割する際，窓幅に足りないクラスの音声区間は窓幅と同じになるように，後続するクラスを含める形で対応しているため，複数のクラスの音が含まれるフレームの割合が増加し，その結果誤認識されるフレームが増加したためだと思われる．

また，窓幅 8.0 [sec]，フレームシフト 4.0 [sec] のときのサンプル毎の識別結果を表 5 に示す．各サンプルにおける識別結果は各フレームの識別結果の多数決で決定されるが，フレームごとの識別精度についても同様に掲載した．結果より，サンプル単位の単純平均においては，24 の認識対象クラスに対して平均で 63.0%の性能が得られた．表 5 について，クラス毎のサンプル単位の認識率に着目すると，“eating lunch” や “presentation (give)”，“washing something” は 100.0% 正しく識別されているのに対して，“talking” と “rest” は全く認識されていないことが分かる．この原因として，背景雑音の多様性が挙げられる．“eating lunch” や “presentation (give)”，“washing something” はほぼ同じ場所 (“eating lunch” は大学のカフェテリア，“presentation (give)” は研究室のゼミ室，“washing something” は被験者自宅の流し台) で収録されているのに対して，“talking” と “rest” は様々な状況において収録されている．この背景雑音の多様性の差が，このような識別率の差に大きく関わっていると考えられる．

5. 認識対象クラスの音響的な類似性の分析

5.1 分析条件

識別性能の向上に有用な特徴量の検証を目標として，4 節で利用した特徴量に基づいてクラス間の音響的な類似性を分析した．まず，クラスタリングによって特徴量空間上の認識対象クラスの距離を可視化した．音響特徴量としては MFCC を採用した．ここで，Section 4 で採用した特徴量ベクトルはフレーム処理によって抽出されるため，時系列信号として取得される．そのため，クラスタリングにおいては各認識対象クラスで抽出された MFCC の平均ベクトルを代表値として利用した．したがって，

$$\hat{f}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} f_{cn} \tag{1}$$

である．ここで，f_{cn} は認識対象クラス c の n 番目のフレームの MFCC，N_c はクラス c のフレーム数，$\hat{f}_c$ は認識対象クラス c の MFCC のフレーム平均を示す．クラスタリングは決定木学習によって行い，決定木の学習には Ward 法の R による実装を利用した．

5.2 分析結果

学習された決定木を図 2 に示す．図より，認識クラスの集合は根ノードで大きく 2 つに分割される．一つ目の集合には “talking” や “presentation” といった被験者本人や第三者の発話が多く含まれる行動クラスが分類されており (以下，この集合を Speaking Group と記述する)，二つ目の集合ではその他のクラスが分類されていることがわかる．その他の音声データは，その後 “cooking” や “transfer outside” などの比較的背景雑音が大きい行動クラスのグループ (Noisy Group) と，“sleeping” や “desk work” などの静寂な環境で出現する行動クラスのグループ (Silent Group) に分割される．上記の 3 グループ毎の認識率の違いを見るため，表 5 の各クラスのサンプル毎の認識率をグループごとに分け，グループ毎にその単純平均を求めた．その結果を表 6 に示す．これより，Speaking Group の UA (SAMPLE) が他 2 グループと比べて低いことが分かる．

また，表 7 は Speaking Group に着目した場合の Confusion Matrix である．ただし，Speaking Group 以外への誤認識もあるため，表 7 の行方向の認識結果の合計が 100.0%でないクラスも存在する．ここで，認識対象クラスのインデックスは表 1 に対応する．表 7 より，Speaking Group に含まれるクラスは，ほとんどのクラスにおいて同じ Speaking Group のクラスに誤認識されていることが

表 5　識別実験の結果
Table 5　Result of discrimiation experiment.

class Name	継続時間（秒）	出現頻度	FRAME(%)	SAMPLE(%)
desk work	18470	32	40.2	54.0
Q&A in presentation (listen)	13645	22	42.7	52.9
talking	10658	56	0.6	0.0
presentation (listen)	7948	14	82.6	95.7
cooking	7265	28	55.3	68.5
meeting	6707	9	28.1	31.9
transfer inside a building	4702	57	13.7	15.8
eating lunch	3635	7	87.1	100.0
English lesson	3085	3	61.1	92.3
transfer outside	2737	17	61.5	81.8
sleeping	2526	4	87.8	95.5
speaking	2075	7	70.6	97.2
play with smartphone	1208	11	80.0	95.0
Q&A in presentation (give)	1079	2	28.6	31.3
rest	1037	17	7.6	0.0
announcement	947	7	41.1	56.3
shopping	919	3	76.3	93.8
presentation (give)	647	1	94.6	100.0
wait for something	597	10	67.0	75.0
washing something	371	5	82.1	100.0
teaching	357	2	33.9	25.0
preparation of presentation (listen)	311	7	35.7	75.0
study	261	1	33.9	50.0
office work	253	3	16.1	25.0
WA			51.2	63.0
UA			51.2	55.6

表 6　3 行動クラスグループそれぞれの UA (SAMPLE)
Table 6　UA (SAMPLE) in 3 groups of activity classes.

	Speaking Group	Noisy Group	Silent Group
UA(SAMPLE)	57.0	65.1	77.7

分かる．従って，本研究で設定した認識対象クラスをより精度よく識別するためには，Speaking Group に含まれる行動クラスを分別できる特徴量の導入が必要であると考えられる．例えば “presentation (give)” や “presentation (listen)” などでは，発話している話者が異なるため，話者情報を表現する特徴量の導入が有用であると考えられる．また，Ziaei ら [12] で行われた状況の分類のように，言語情報の導入なども “meeting” や “speaking” などの分類に効果的であると考えられるため，今後は特徴量の選択に関しても検討を行う予定である．

6. おわりに

本研究は，生活行動の見直しを補助するアプリケーションの作成を目標として，ウェアラブルデバイスで収録された日常音データから着用者の行動推定を行うための基礎的な分析を行った．被験者 1 名によって収録された 22 時

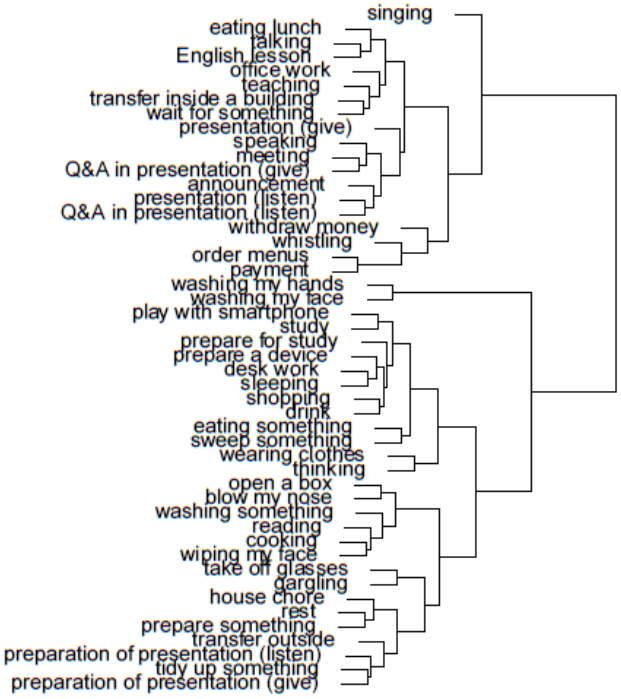

図 2　行動クラスの階層的クラスタリング結果
Fig. 2　Result of hierarchical clustering of activity classes.

表 7　Speaking Group におけるサンプル毎の混同行列 (%)
Table 7　Confusion matrix by sample in speaking group (%).

		認識結果														合計
		QAL	PRL	ETL	EGL	ANC	OFW	WTS	TCG	TLK	MTG	TIB	SPK	QAG	PRG	
真	QAL	52.9	30.4	0.0	0.0	13.3	0.0	0.0	1.3	0.0	0.8	0.0	0.8	0.0	0.0	99.5
	PRL	2.9	95.7	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.3
	ETL	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
	EGL	5.9	0.0	0.0	94.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
	ANC	25.0	6.3	0.0	6.3	56.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	93.9
	OFW	0.0	0.0	25.0	0.0	0.0	25.0	0.0	25.0	0.0	0.0	0.0	25.0	0.0	0.0	100.0
	WTS	0.0	0.0	0.0	0.0	0.0	0.0	75.0	0.0	0.0	0.0	12.5	0.0	0.0	0.0	87.5
	TCG	0.0	0.0	0.0	25.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	25.0	0.0	0.0	75.0
	TLK	1.1	0.5	19.6	37.0	1.6	2.1	1.6	4.9	0.0	0.0	0.0	3.8	6.0	11.4	89.6
	MTG	1.7	10.3	0.0	2.6	7.8	0.0	0.0	0.9	0.0	31.9	0.0	39.7	0.0	0.0	94.9
	TIB	1.3	0.0	18.4	3.9	3.9	0.0	19.7	3.9	0.0	0.0	15.8	0.0	1.3	1.3	69.5
	SPK	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	97.2	0.0	0.0	100.0
	QAG	0.0	6.3	0.0	6.3	18.8	0.0	0.0	0.0	0.0	0.0	0.0	6.3	31.3	31.3	100.0
	PRG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0

間の日常音データに対して 48 の認識対象クラスのラベルを付与し，識別実験を行った．識別実験では特徴量抽出の窓幅を変更することで最適な窓幅の検討を行った．結果より，窓幅を 8.0 [sec] としたときに最も高い識別性能が得られ，このときサンプルベースでの識別率は平均で 63.0%であった．また，認識対象クラスのクラスタリングでは，3 つの集合に分割され，特に音声に関するクラスに関して識別性能が低下することが示された．今後は音声クラスの識別に有用な特徴量の検討を行う予定である．

参考文献

[1] Sellen, A. J., Fogg, A., Aitken, M., Hodges, S., Rother, C. and Wood, K.: Do life-logging technologies support memory for the past?: an experimental study using sensecam, *Proc. SIGCHI*, ACM, pp. 81–90 (2007).
[2] Wang, P. and Smeaton, A. F.: Using visual lifelogs to automatically characterize everyday activities, *Information Sciences*, Vol. 230, pp. 147–161 (2013).
[3] 林知樹，西田昌史，北岡教英，武田一哉：DNN による環境音と加速度信号を用いた日常生活行動認識，日本音響学会春季講演論文集，pp. 83–86 (2015).
[4] Nishida, M., Kitaoka, N. and Takeda, K.: Development and preliminary analysis of sensor signal database of continuous daily living activity over the long term, *Proc. APSIPA*, pp. 1–6 (2014).
[5] Ohishi, Y., Mochihashi, D., Matsui, T., Nakano, M., Kameoka, H., Izumitani, T. and Kashino, K.: Bayesian semi-supervised audio event transcription based on Markov Indian buffet process, *Proc. ICASSP*, pp. 3163–3167 (2013).
[6] Peng, Y.-T., Lin, C.-Y., Sun, M.-T. and Tsai, K.-C.: Healthcare audio event classification using hidden markov models and hierarchical hidden markov models, *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 1218–1221 (2009).
[7] Temko, A. and Nadeu, C.: Acoustic event detection in meeting-room environments, *Pattern Recognition Letters*, Vol. 30, No. 14, pp. 1281–1288 (2009).
[8] Campbell, W. M., Sturim, D. E., Reynolds, D. A. and Solomonoff, A.: SVM based speaker verification using a GMM supervector kernel and NAP variability compensation, *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 1, pp. I–I (2006).
[9] Zhuang, X., Zhou, X., Hasegawa-Johnson, M. A. and Huang, T. S.: Real-world acoustic event detection, *Pattern Recognition Letters*, Vol. 31, No. 12, pp. 1543–1551 (2010).
[10] Sadjadi, S. O. and Hansen, J. H.: Unsupervised speech activity detection using voicing measures and perceptual spectral flux, *IEEE Signal Processing Letters*, Vol. 20, No. 3, pp. 197–200 (2013).
[11] Graciarena, M., Alwan, A., Ellis, D., Franco, H., Ferrer, L., Hansen, J. H., Janin, A., Lee, B. S., Lei, Y., Mitra, V. et al.: All for one: feature combination for highly channel-degraded speech activity detection., *Proc. INTERSPEECH*, pp. 709–713 (2013).
[12] Ziaei, A., Sangwan, A., Kaushik, L. and Hansen, J. H.: Prof-Life-Log: Analysis and classification of activities in daily audio streams, *Proc. ICASSP*, pp. 4719–4723 (2015).
[13] Sangwan, A., Ziaei, A. and Hansen, J. H.: ProfLifeLog: Environmental analysis and keyword recognition for naturalistic daily audio streams, *Proc. ICASSP*, pp. 4941–4944 (2012).
[14] 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，日本音響学会春季講演論文集，pp. 139–142 (2017).
[15] Trilsbeek, P.: ELAN, The Language Archive (online), available from ⟨https://tla.mpi.nl/tools/tla-tools/elan/⟩ (accessed 2017-05-29).