

# 音響情報を用いたライフログデータのクラスタリング

山野貴一郎 伊藤克亘  
法政大学大学院情報科学研究科

E-mail: kiichiro.yamano.rr@gs-cis.hosei.ac.jp

個人の生活をカメラ、マイク、GPS などを用いて記録し利用するための試みがなされている。このような個人の生活の記録をライフログと呼ぶ。ライフログはデータ量が膨大かつ冗長で、効率的な利用のためには検索のためのインデキシングや要約が必須である。本論文では音響ライフログ情報のクラスタリング手法について述べる。初めに音響ライフログ情報から冗長部分を除去するために音響情報を約 5 秒のセグメントに分割し、スペクトルと振幅を利用して冗長な部分である無音のセグメントの除去をした。さらに残ったセグメントに対し k-means クラスタリングをした。特徴量としてスペクトルや振幅の差分値を用いたところ、正規化スペクトル包絡で比較的良好な結果が得られた。

## 1 まえがき

個人の生活や体験を様々なセンサ (カメラ、マイク、GPS など) を用いて記録し、利用するための研究が行われている [1]。このような個人の記録をライフログと呼び、備忘録や自動の日記作成、業務内容の共有などへの利用が期待されている。しかし、ライフログは常時記録をしているためデータ量が膨大かつ冗長であり、そのままでは利用が難しい。よって、効率的な利用のためには要約や検索の必要があり、様々な試みがなされている。

本研究では特に日常生活の音を記録した音響ライフログ情報について扱う。音響ライフログ情報には様々な情報 (音声、音楽、生活音など) が含まれているが、情報の含まれていない冗長な部分も多くある。そこで音響ライフログ情報を固定長のセグメントに分割し、スペクトル包絡と振幅の差分を利用して冗長な部分を省いた。また、音響ライフログ情報は得られる情報の予測が困難なため、事前に分類を決められない。したがって、スペクトル包絡などの特徴量を用いて k-means クラスタリングを行い、効果的な特徴量を調査した。

## 2 ライフログのデータ処理

### 2.1 従来研究

ライフログデータの検索や要約に関する様々な研究が行われている。文献 [2] では脳波、加速度、位置情報などのセンサ情報とインターネットの履歴、e-mail などから検索キーを抽出しライフログ映像を検索を行うシステムが提案されている。また、Life Pod[3] という携帯電話を利用したライフログシステムも提案されて

いる。Life Pod では携帯電話から得られた画像や位置情報、ユーザが入力したメモの管理を行っている。さらに、RFID タグを使うことで周囲の物の情報の取得もできるとしている。

ライフログ情報のクラスタリングやセグメンテーションについても議論されている。例えば、文献 [4] では個人が常時記録した映像の色ヒストグラム情報によるクラスタリングを行っている。データは MPEG-4 で 62.5 時間収録されオフィス、階段、廊下、エレベーターなど 34 種類のラベルを付与している。時間的に近いデータが同じクラスタに含まれやすくなるようにされた TCK-means クラスタリングを、このデータに適用し、k-means 法を改善した結果を報告している。日常のイベント毎にセグメンテーションを行う手法も提案されている。文献 [5] では 1 日 1785 枚の画像からなるライフログ映像を扱っている。その映像をデバイスの電源が 2 時間以上切られている箇所で分けることで 1 日単位のデータに分割し、その上でイベント毎にセグメンテーションを行っている。セグメンテーションは MPEG-7 のメタデータより得られた色情報やエッジ情報を用いており、隣接する画像の非類似度が高いところをイベント境界としている。この文献では 5 名のユーザが 1 か月間記録した 271163 枚の画像を使って実験を行っている。また、文献 [11] ではオフィス環境でのデスクワークと会議を分類している。デスクワークでは紙をめくる音やキーボードの打鍵音などが多く、会議では音声が大きくことに着目し、これらの音をスペクトル情報を利用して検出することで、行動の分類を行っている。

また、個人が常時記録した音響情報の要約や検索に

関する研究も行われている．文献 [6] ではユーザの記憶を支援するためのシステムとして，位置情報や会話データに音声認識を行った結果を利用している．しかし，会話に対する音声認識の結果は誤りを含むことが多いので，認識をした単語の信頼度も併せて提示することで，ユーザの想起を支援するシステムが提案されている．また，文献 [7] では収録時のユーザの負担を最小にするため，センサは無指向性マイクと GPS のみを利用し，62 時間のデータを収録して音響情報のスペクトル情報に着目しクラスタリングを行うことで，図書館，レストラン，授業，会議などの 16 の場所や環境に関するセグメントの分類を行っている．また，セグメントの平均時間が 26 分とし，最短の場合でも 15 分程度のセグメントをクラスタリングするために，1 フレームを 1 分として特徴量の抽出を行っている．

ビデオのインデキシングをの手法として TV 番組など編集された映像を対象とするものが提案されている (例えば，[8, 9, 10])．編集された映像ではシーンの切り替えが暗転するなどして明確である．しかし，ライフログ映像では当然そのような効果はなく，編集された映像とは異なりカメラの前を人や物が横切ったりすることも多いので処理が難しい．また，放送用の音響情報では雑音などが少なく，音声などはっきりと収録されている．しかし，音響ライフログ情報では，周囲の環境によって様々な雑音が収録される．さらにマイクからの距離により音量にばらつきがある．音響ライフログ情報を扱う場合にはこれらのことを考慮しなければならない．

音響ライフログ情報には様々な音が含まれており，得られる情報も多い．例えば，音声からは会話の内容や話者などの情報がわかり，その他の音からは，場所や行動，周辺の状況 (騒がしい，静かななど) が得られることもある．このような情報を詳細に分類するためには，上記の研究から考えて，場所や含まれている音の種類によってクラスタリングを行うべきと思われる．また，一回のクラスタリングで利用しやすく分類を行うのは難しいので，数回に分けてクラスタリングを行うべきと考えられる．例えば，最初のクラスタリングでは場所によって分類し，次に場所ごとに含まれている音に着目しクラスタリングを行うという手法である．音の種類 (音声，音楽，雑音など) によって適切な処理を行い，検索キーやインデクスを付けるということもできる可能性がある．

## 2.2 プライバシー

映像や音響情報を常時記録するためにはプライバシーに関する問題を解決しなければならない．従来研究でもプライバシーについて言及されており，音響情

報を用いて自動日記を作成するための研究を行っている文献 [7] では，会議を収録する場合には参加者を話者認識によって記録として残すことは許されるが，発言内容まで残すことは許されないかもしれないと述べられている．また，プライバシーに配慮して音響情報を常時記録するためには，技術的に信頼できる解決法が必要であるとし，話者の情報のみを残すためには，音声区間を 1 秒に分割し，さらに 50 ミリ秒の区間でランダムに入れ替えてしまうことで，スペクトルを変化させずに発言内容を消去できるとしている．収録の許可を得られた話者に関してはこのような処理はいらないので，精度の高い話者認識が必要であるとも述べている．

文献 [6] では，収録にはどのような許可が必要か，誰がデータにアクセスするのか，常時記録を行うデバイスを利用するためにはどのような社会的規則が必要か，どのような法的保護が利用できるか，ということが解決されなければ一般的な利用は期待できないとしている．さらに上記の問題が解決し一般的に利用できるようになったとしても，裁判所などから押収された場合に個人情報保護するための方法も考えなければならないと述べている．

また，モダリティによって，用途に向き不向きがある，収録しやすい環境が異なる，効果的な処理が異なるなどの問題があり，プライバシーの観点からも，映像や音響情報等の様々なデータいつでも収録できるとは限らないので，音響ライフログ情報のみの処理も重要であろう．

## 3 データ収録

実験に使用したライフログの音響データは 7 日分 (約 41 時間) である．データはバイノーラルマイク (adphox BME-200) と PCM 録音機 (EDIROL R09, YAMAHA POKETRAK CX) を用いて収録をした．7 日分のうち 4 日分 (約 23 時間) は EDIROL R09, 3 日分 (約 18 時間) は YAMAHA POKETRAK CX を用いての収録を行った．複数の録音機を使うことでそれぞれの機器でマイクゲインが異なるため，その点を考慮しなければならない．

バイノーラルマイクはイヤホン型マイクで，両耳に装着し録音をする物であるが，両耳に装着しての長時間の収録はユーザの負担となるので，バイノーラルマイクは肩から提げ，マイクを胸の辺りに設置して使用する (図 1)．その際，マイクの間隔が一定となるように針金で固定した．その状態で日常生活の音を収録した．収録条件はサンプリング周波数 48kHz で，量子化ビット数は録音機によって異なる 24 ビットもしくは 16 ビッ

トで収録した．表 1 に主な収録場所とそこで収録された音をまとめた．



図 1 パイノーラルマイク

表 1 主な収録場所と収録された音

場所	主な収録音
研究室	音声, PC 操作の音, 紙をめくる音, ファンの騒音
教室	音声, ファンの騒音
廊下	足音, 音声
大学構内(屋外)	工事, 排気ダクト, 音声
自宅	TV, 音楽
レンタルビデオ店	音楽, 音声
ファストフード店	音声
コンビニエンスストア	音楽, 音声
スーパーマーケット	音声, 音楽
路上	車, 音声

## 4 不要なセグメントの除去

### 4.1 セグメント

以上のように収録したデータには多くの音が含まれており, 様々な情報が得られる. 一方で, データには音がほとんど含まれていない部分や, 含まれていても何の音かが判別できない冗長な部分も多くある. 冗長部からは情報が得られないのでインデキシングやクラスタリングの必要性はない. そこで, 冗長部の除去や必要な部分のクラスタリングのためにデータを短時間のセグメントに分割した. 文献 [7] では識別を行う場所や環境の平均時間を算出しており, 平均 26 分と述べられている. また, 識別を行いたい最短の出来事が 15 分程度であるとし, 1 分のフレームで処理している. しかし, 識別したい場所や環境は長時間のものや短時間のものがあり, セグメント長は用途によっても異なる. 場所や状況といったある程度長い出来事を分類するには長めのセグメントでも適当かもしれないが, 音声, 音

楽, その他の環境音などの分類を行いたい場合には, 短時間のセグメントの方が有効であろう. そこで, 本論文ではセグメントをなるべく短くした場合のクラスタリングの様子を調査をする. セグメント長は音声が含まれている場合は話者や話している内容が認識できる長さ, それ以外の音ではその音が何の音かが判別できる長さとした. データを聴取し, そのような長さとしては 5 秒程度が最低必要であると判断し, 後のクラスタリングで行う処理のことを考慮してセグメント長を約 4.99 秒とした.

### 4.2 スペクトル包絡

不要なセグメントの除去には次の手順で求めた正規化スペクトル包絡を用いた. 各セグメントに 4096 点 FFT を 2048 点ずつシフトさせながら行い, 短時間スペクトルを求め, フィルタバンク分析をしてスペクトル包絡を得た. フィルタバンク分析ではメル周波数軸上で固定長の帯域幅の三角窓をシフトさせながら波形を切り出し, その帯域の和を求めた. 三角窓の幅は 600 でシフトは 300 とした.

この処理により短時間スペクトルの特徴が 12 点に集約されたスペクトル包絡が得られる. セグメント毎にスペクトル包絡は複数得られるので, それらを平均してセグメントの特徴とした. 正規化はスペクトル包絡から包絡の各点の平均を減算することで行った.

### 4.3 除去手法

収録を行ったデータ全てをセグメントに分割したところセグメント数は 30157 であった. その中からランダムに 100 セグメントを選び, 聴取により各セグメントの場所と含まれている音のラベリングと要不要の分類をした. その結果, 31 セグメントが情報の含まれていない不要なセグメントと判明した. 不要なセグメントには空調などの低騒音のみかほぼ何も含まれていない 18 セグメントと, 含まれる音が何の音か判別できない 13 セグメントがある. 前者はほぼ無音に近く, ある程度特性の予測ができるので除去が可能と思われるが, 後者は他の音との違いを識別するのは難しい. また, 必要なセグメントを誤って除去してしまうのはできる限り避けなければならない. よって, 比較的容易と考えられるほぼ無音のセグメントの除去を行う. 無音のセグメントの特徴としてスペクトルに変化がないことが挙げられるので, 正規化スペクトル包絡の差分を利用した. 次の 3 つの条件を全て満たしたものを除去した..

1. 正規化スペクトル包絡の差分の総和が閾値以下
2. 正規化スペクトル包絡の差分の最大値が閾値以下

### 3. 正規化した振幅の差分値が閾値以下

1 と 2 では 2 フレーム間の正規化スペクトル包絡の各点の差分値の絶対値を利用している．3 では振幅差分の絶対値の総和を求めているが，それだけではマイクゲインによって値が変わってしまうので，振幅の平均で除算をすることで正規化する (式 (1))． $A_n$  は  $n$  点目の振幅， $A_{average}$  は振幅の絶対値を平均した値である．

$$A_{delta} = \frac{\sum_n |A_n - A_{n-1}|}{nA_{average}} \quad (1)$$

1 と 3 では音の変化が少ないサンプルを求め，2 では瞬時に変化する音の抽出を行っている．瞬時に変化する音の例としては，PC 操作の音 (クリック音，キーボード打鍵音) が挙げられる．これらの音もコンテキストを理解するために必要な音である．閾値は前述の 100 セグメントを処理した場合に必要なセグメントが除去されない値を採用し，1 は 3 以下，2 は 13 以下，3 は 0.1 以下とした．この場合では 10 セグメントが除去されそのうち 9 セグメントが無音のセグメントで，1 セグメントが無音ではない不要なセグメントであった．

#### 4.4 不要セグメントの除去実験

前節とは別の 100 セグメントをランダムに選び，同じ閾値を用いて不要セグメントの除去を行った．この 100 セグメントには不要な 31 セグメントが含まれ，そのうち 12 セグメントが無音のセグメントであった．除去を行ったところ 8 セグメントが除去され 6 セグメントが無音で 2 セグメントがその他の不要なセグメントだった．以上のことから本論文の手法である程度は無音セグメントを除去できると考えられる．

## 5 k-means 法によるセグメントのクラスタリング

前節の実験で利用した 100 セグメントに除去処理を行い，残りの 92 セグメントに対して k-means クラスタリングを行った．本論文ではクラスタ数を 2~10 と変化させ実験を行った．

### 5.1 特徴量

ライフログには多様な音が含まれている．したがって，適当な処理を行うには様々な特徴量を考慮しなければならない．本論文では以下の 7 つの特徴量を用いて実験を行った．

- 平均スペクトル包絡
- 正規化平均スペクトル包絡
- パワー
- 振幅の差分の最大値

- 振幅平均により正規化した振幅の差分値
- 正規化スペクトル包絡の差分の最大値
- 正規化スペクトル包絡の差分の総和

平均スペクトル包絡はセグメント内のすべてのフレームから得たスペクトル包絡の平均である．セグメントに現れた音響情報の周波数の特徴とパワーの特徴を含んでいる．スペクトル包絡の正規化は各点の平均値を減算することにより行った．スペクトル包絡は同じ環境音でのクラスタリングができ，場所の分類ができると考えた．また，正規化を行うことで，マイクゲインの違いに関係なくクラスタリングができると思われる．パワーは平均スペクトル包絡の各点の総和である．パワーはスペクトル包絡をより簡潔にしたものなので，音の種類に関係なく単純に大小でのクラスタリングができる．振幅の差分値は時間波形の前後の点の差分の絶対値を求め，その最大値を利用した．下から 3 つは除去に用いた特徴量と同じ処理をして求めた．正規化した振幅の差分値の総和と正規化スペクトル包絡の差分の総和はセグメントに多くの音が含まれている場合に大きくなる．よって，様々な音が含まれやすい環境 (屋外) と含まれにくい環境 (室内) でのクラスタリングに役立つと考えた．振幅の差分と正規化スペクトル包絡の差分の最大値は瞬間的に変化する音の検出をしている．よって，環境音と音声やその他の音が含まれているセグメントの分類に役立つと思われる．

### 5.2 クラスタリング結果

クラスタ数が 4 の場合の結果を表 2 に示す．括弧内はセグメント数である．

### 5.3 考察

正規化していない特徴量ではマイクゲインによってクラスタリングされる傾向にあった．つまり，音量が反映されているので，収録時のゲインの大小でクラスタリングされてしまった．正規化をした場合にはマイクゲインによってクラスタリングされることはない．したがって，異なった録音機を使った場合に場所や音によるクラスタリングを行うには，正規化した特徴量が音量によらない特徴量を用いるべきと考えられる．

正規化を行った特徴量ではスペクトル包絡が比較的良好な結果が得られた．クラスタ数が 4 の場合は教室，研究室，自宅というクラスタができていた．しかし，理想的には研究室のクラスタが 1 つになり，別のクラスタができるのが望ましい．正規化スペクトル包絡を用いてクラスタ数を 3 にした場合は研究室，教室，自宅の 3 つのクラスタができており，5 以上の場合は研究室のクラスタがより細かく分割された．よって，正規化スペク

表 2 クラスタ数が 4 の場合の各クラスタに含まれるセグメント

特徴量	含まれるセグメント
スペクトル包絡	教室 (7), 大学屋外 (2), 廊下 (1) 自宅 (1), コンビニ (1)
	自宅 (12), 研究室 (1)
	研究室 (32), 教室 (1) 廊下 (1), 自宅 (1)
	研究室 (24), 教室 (4) 廊下 (1), 自宅 (1)
正規化 スペクトル包絡	教室 (12), 研究室 (3) 大学屋外 (1) 自宅 (16), 研究室 (1)
	研究室 (37), 廊下 (3), コンビニ (1), 大学屋外 (1)
	研究室 (16), 自宅 (1)
	研究室 (6), 廊下 (2)
正規化 スペクトル包絡の 平均差分	研究室 (19), 自宅 (12), 教室 (6), 廊下 (1)
	研究室 (12), 教室 (4), 自宅 (4), 大学屋外 (1)
	研究室 (20), 教室 (2), 自宅 (1) 大学屋外 (1), コンビニ (1)
	研究室 (6), 廊下 (2)
正規化 スペクトル包絡の 最大差分	研究室 (18), 教室 (3) 自宅 (2), 大学屋外 (2) 廊下 (1), コンビニ (1)
	研究室 (17), 教室 (5), 自宅 (5), 廊下 (1)
	研究室 (12)
	研究室 (12), 自宅 (10), 教室 (4), 廊下 (1)
パワー	研究室 (23), 廊下 (1), 自宅 (8)
	研究室 (9), 教室 (5), 自宅 (4), 廊下 (1)
	教室 (7), 自宅 (3), 大学屋外 (2) 廊下 (1), コンビニ (1)
	研究室 (25), 自宅 (2)
振幅の差分	研究室 (4), 教室 (2), 自宅 (1)
	研究室 (6), 教室 (3), 廊下 (2), 自宅 (1), 大学屋外 (1)
	研究室 (37), 自宅 (14), 教室 (6) 廊下 (1), 屋外 (1), コンビニ (1)
	研究室 (10), 教室 (1)
正規化振幅の差分	自宅 (8), 研究室 (1)
	自宅 (8)
	研究室 (45), 教室 (10), 廊下 (3) 大学屋外 (2), コンビニ (1)
	研究室 (11), 教室 (2), 自宅 (1)

トル包絡を用いる場合は、この実験の範囲ではクラスタ数が 3 が最も良い結果を得られると考えられる。しかし、廊下や屋外、コンビニエンスストアなどは適当なクラスタリングができていないので、正規化スペクトル包絡では特徴量として不十分である。したがって、他の特徴量を組み合わせて使うか別の特徴量を用いなければならない。

また、正規化スペクトル包絡の平均、最大値や正規化振幅の差分では場所によるクラスタリングはできなかった。これらは音の時間変化の特徴量であり、場所でのクラスタリングは難しかったと考えられる。時間変化を使うことで音響情報でのクラスタリングができるのかを調べるため、正規化を行った特徴量のクラスタ数が 4 のときの各クラスタに含まれる音を表 3 にまとめた。括弧内は出現したセグメント数である。表より、時間変化や正規化した平均スペクトル包絡は音響情報のクラスタリングに適していないことがわかる。我々が過去に行っていた電車の発着と電車内、ホームで電車を待っている 3 つの状況を音響情報によって識別をする研究では、スペクトル包絡が 12 点でもある程度の識別が可能であった [12]。また、識別を行いたい状況(ショット)の平均時間は 35 秒程度であった。しかし、本研究での音響ライフログ情報は状況が限られていないため含まれる音の種類が膨大である。そのため 12 点のスペクトル包絡では特徴量として粗く、音響情報のクラスタリングができなかったと考えられる。

## 6 従来研究との比較

文献 [7] でも手動で付けたラベルとクラスタリング結果を比較している。ラベルは図書館、大学、会議、レストランなど 16 ラベルで、スペクトルエントロピーを用いてクラスタリングを行い 15 クラスタに分けている。本論文とは特徴量や手法が異なるが、結果を比較すると文献の手法は 4 つのクラスタが再現率、精度ともに 70% 以上の良好なクラスタリング結果が出ている。しかし、他のクラスタでは再現率、精度の一方のみが高いか、両方が低い結果になっている。一方本論文での手法ではクラスタ数を多くすると、同じ場所のクラスタが複数に分かれてしまう傾向があった。その原因として本論文の手法は約 5 秒という短時間の特徴を用いているため含まれている音によって、異なるクラスタができてしまったためと考えられる。よって、場所でクラスタを作るためには長時間の特徴を利用するのが良いと思われる。本論文で用いた特徴量やセグメント長は粗いクラスタリングののちに各クラスタに対して詳細なクラスタリングを行うときに有効であるかもしれない。

表 3 クラスタ数が 4 の場合の各クラスタに含まれる音響情報

特徴量	含まれる音響情報
正規化 スペクトル包絡	音声 (14), 騒音 (1), 物音 (1) TV(9), 音楽 (7), 物音 (3) 作業 (21), 音声 (12), 物音 (5), 騒音 (1), 足音 (1)
	作業 (8), 物音 (6), 音声 (2), 無音 (2), 音声 (5), 作業 (3)
	物音 (11), 作業 (8), 音声 (6) 音楽 (6), TV(5), 足音 (1) 作業 (5), 音声 (4), 物音 (3) 無音 (3), TV(3), 音楽 (1) 作業 (13), 音声 (13), 音楽 (1) 騒音 (1), 物音 (1), TV(1)
正規化 スペクトル包絡の 平均差分	作業 (13), 音声 (11), 騒音 (2) 音楽 (2), TV(1) 物音 (1), 無音 (1)
	音声 (11), 作業 (7), 物音 (6) TV(3), 音楽 (2), 無音 (1) 作業 (8), 音声 (4)
	物音 (8), TV(5), 音声 (4) 音楽 (4) 無音 (3) 作業 (1), 足音 (1)
	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
正規化 スペクトル包絡の 最大差分	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
正規化振幅の差分	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)
	音楽 (6), 物音 (2), TV(1) TV(8), 音楽 (1), 物音 (1) 音声 (25), 作業 (22), 物音 (9) 無音 (4), 騒音 (2) 音楽 (1), 足音 (1) 作業 (6), 音声 (5) 物音 (3), 無音 (1)

## 7 あとがき

ライフログデータにインデキシングを行うための特徴量として、スペクトル包絡や振幅を用いてクラスタリング手法を行った。その結果、正規化スペクトル包絡を用いた場合に比較的良好な結果が得られた。

データの収録は有線のマイクを用いた。しかし、長時間の収録を行う場合にはユーザへの負担を減らすため無線のマイクを用いるべきと思われる。また、無線の受信機などが消費する電力も考慮しなければならない。

収録をしたデータにクラスタリングの事前処理として無音のセグメントの除去を行い、スペクトル包絡などの特徴量を用いて k-means クラスタリングを行った。その結果、正規化を行った特徴量を用いることで、大雑把なクラスタリングができたが、正規化スペクトル包絡では不十分であった。また、従来手法との比較から、場所などでの粗いクラスタリングにはある程度長時間の

特徴を用いることが有効であり、本論文のような短時間のセグメントを用いたクラスタリングは、音の種類によってクラスタリングを行う場合に有効である可能性がある。正規化を行っていない特徴量では、録音機の違いによってクラスタリングされてしまう傾向が見られた。つまり、収録時のマイクゲインによって分類されてしまった。同一の録音機で同じマイクゲインで収録を行えばこの問題は解決できるが、ライフログの収録期間は数十年にわたり、機器の寿命よりも長いので、処理手法は収録機器の影響を受けてはならない。よって、今後は収録機器に関係なく良好なクラスタリングを行うための手法を検討しなければならない。また、最適なセグメント長やインデクスを付けるための手法も検討していかなければならない。

## 参考文献

- [1] J Gemmell et. al., "MyLifeBits: A PERSONAL DATABASE EVERYTHING", *COMMUNICATIONS OF THE ACM*, Vol.49, No.1, pp.88-95, Jan. 2006
- [2] K Aizawa, "Digitizing Personal Experiences: Capture and Retrieval of Life Log", *Proceedings of the 11th International Multimedia Modelling Conference*, pp.10-15, Jan. 2005
- [3] Atsunori Minamikawa et. al, "RFID Supplement for Mobile-Based Life Log System", *Proceedings of SAINTW'07*, pp.50-50, Jan. 2007
- [4] Wei-Hao Lin et. al, "Structuring Continuous Video Recordings of Everyday Life Using Time-Constrained Clustering", *SPIE Symposium on Electronic Imaging*, Jan. 2006
- [5] Aiden R. Doherty et. al, "Automatically Segmenting LifeLog Data into Events", *In WIAMIS 2008*, pp.20-23, May 2008
- [6] V Sunil et. al., "An Audio-Based Personal Memory Aid", *UbiComp 2004*, Vol.3205, pp.400-417, Oct. 2004
- [7] DPW Ellis et. al., "Minimal-impact audio-based personal archives", *CARPE'04*, pp.39-47, Oct. 2004
- [8] 新田他, "放送型スポーツ映像の構造を考慮した重要シーンへの自動アノテーション付け", *信学論 (D-II)*, Vol.J84-D-II, No. 8, pp.1838-1847, Aug. 2001
- [9] 佐野他, "サッカー中継における会場音とスピーチを利用したメタデータ生成", *信学技報*, PRMU, Vol.105, No.415, pp.33-38, Nov., 2005
- [10] 南他, "音情報を用いた映像インデキシングとその応用", *信学論 (D-II)*, Vol.J81-D-II, No. 3, pp.529-537, Mar, 1998
- [11] 志村他, "行動状況により検索可能な体験映像提示手法の検討", *情処論議* 68回, pp.4.81-4.82, 2006
- [12] 山野他, "バイノーラルマイクを用いたライフログ映像のショット識別", 第 23 回信号処理シンポジウム, Nov. 2008