音響情報を用いたライフログデータのインデキシング*

山野貴一郎 (法政大学大学院情報科学研究科), 伊藤克亘 (法政大学)

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，ライフログはデータ量が膨大であり，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．

ライフログの音響情報においても要約や検索に関する研究が行われている．文献 [2] 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，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している．また，文献 [3] 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，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，会議などの場所や環境の分類を行っている．

以上のようにライフログ音響情報には様々な情報が含まれているが，情報の含まれていない冗長な部分も多くある．そこで本論文ではライフログ音響情報を固定長のセグメントに分割し，振幅を利用して冗長な部分を省いた．また，ライフログ音響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，適当なクラスタ数や効果的な特徴量を調査した．

2　データ収録

実験に使用したライフログの音響データは 4 日分 (約 23 時間) である．データはバイノーラルマイク (adphox BME-200) を用いて，サンプリング周波数 48kHz，量子化ビット数 24 ビットで収録した．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．その状態で日常生活の音を収録した．表 1 に主な収録場所とそこで収録された音をまとめた．

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，排気ダスト，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
路上	車，音声

3　不要なセグメントの除去

3.1　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 [3] 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ってくると考えられる．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の調査をする．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．

3.2　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 16717 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に含まれている音のラベリングと必要，不要の分類をした．その結果，50 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音がほぼ含まれていないものと，含まれる音が何の音か判別できないものがある．前者は振幅を利用すること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．そこで，振幅の差分を利用してほぼ無音のセグメントの除去を行った．PC 作業の音 (クリック音やキーボードを叩く音) のような一瞬のみに大きな差分値を持つ必要なセグメントがあるので，除去はセグメント内の振幅の差分の最大値が閾値以下のときに行う．

3.3　不要なセグメントの除去実験

すべてのセグメントから不要なセグメントを除去する実験を行った．閾値は前述の 100 個のセグメントの場合に必要なセグメントが除去されない値を求め 0.001 とした．除去を行ったところ 1113 のセグメントが除去

* **An Indexing of Life-log Data Using Audio Information by YAMANO, K. (Grad. School of CIS, Hosei Univ.) et. al.**

対象となった．この中から 100 セグメントをランダムに抽出し聴取によって必要なセグメントが含まれていないか調べたところ，7 つが音声や音楽を含む必要なセグメントであった．しかし，この前後の必要なセグメントは除去されていないので，より広い時間幅でインデキシングを行えば影響はないと思われる．

4 k-means 法によるセグメントのクラスタリング

前節初めでランダムに選んだ 100 セグメントから振幅の差分によって不要なセグメントを除去し，残りの 92 セグメントのスペクトル包絡，正規化したスペクトル包絡，パワー，前節と同様の手法で求めた差分の最大値に対して k-means クラスタリングを行った．

4.1 スペクトル包絡

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 [4] 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．また，スペクトル包絡の各点の総和を求めパワーとした．

4.2 k-means クラスタリング

本論文ではクラスタ数を 2〜9 と変えて，各セグメントの特徴量を k-means 法によってクラスタリングをした．クラスタ数が 4 の場合の各クラスタに含まれる主なセグメントの場所と音を表 2 にまとめた．

4.3 考察

本論文ではセグメントのラベルとして場所と含まれる音を用いた．クラスタリング結果より，スペクトル包絡を用いた場合は場所をある程度分類できている．1 つだけ教室や屋外を含むクラスタがあるが，クラスタ数を 5 以上にしても，同じようなクラスタが複数できるだけで，より詳細なクラスタリングは行えなかった．その他の手法では場所，含まれる音いずれの観点からも適当なクラスタリングはできなかった．パワーや差分では音色の違いはわからないので，含まれる音のクラスタリングは難しい．場所に関しても同じ場所で常に一定の音が収録されるということはないため，パワーや差分では特徴量として不十分であると考えられる．正規化スペクトル包絡ではスペクトルの構造のみに着目してクラスタリングを行っているため，類似した音が集まったクラスタができると思われたが，実際にはクラスタ内の音に統一性はなかった．しかし，フィルタバンクの次数を増やすことで，結果が変わる可能性はある．また，単独では不十分であった特徴量でも，複数組み合わせることで適当なクラスタリングが行えるかもしれない．

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	自宅 (無音，音楽，TV)
	研究室 (作業音・無音)
	研究室 (物音)，教室 (物音・音声) 廊下 (足音)，屋外 (騒音)
	店 (音声，音楽)
正規化 スペクトル包絡	自宅 (無音，音楽，TV)， 研究室 (作業音・物音)，廊下 (足音)
	研究室 (作業音・無音)，廊下 (足音)
	研究室 (物音)
	教室 (音声)，研究室 (物音)， 屋外 (騒音)，店 (音声・音楽)
パワー	自宅 (無音，音楽，TV)， 研究室 (作業音・物音) 廊下 (足音)
	研究室 (作業音・物音・無音)， 自宅 (無音・音楽)
	教室 (音声)，屋外 (騒音)， 研究室 (物音)，廊下 (足音)
	店 (音声・音楽)
差分	自宅 (TV)，店 (音声)， 研究室 (作業音・物音)
	教室 (音声)，自宅 (音楽・無音・TV) 研究室 (物音・無音・作業) 屋外 (騒音)，廊下 (足音)
	研究室 (作業音・物音)
	研究室 (物音)，屋外

以上より，本論文の実験の範囲ではスペクトル包絡を用いたクラスタ数が 4 程度のクラスタリングが最も有効であった．しかし，日常遭遇する状況はもっと多いと考えられるので，より多くのデータを収集して実験を行っていく必要がある．更に詳細なクラスタリングを行うには，別の特徴量を用いるか，上記のような大雑把なクラスタリングの後に各クラスタに対して再度詳細なクラスタリングを行うという手法が考えられる．

5 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クラスタリング手法を提案した．結果，スペクトル包絡によって大雑把なクラスタリングが行えたが，これらのデータにインデクスをつけるには，さらなるクラスタリングや別の処理手法が必要である．今後は，インデクスを付けるための手法や別の特徴量を用いたクラスタリング手法の検討をするとともに，より多くデータ収録も行う必要がある．

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] V Sunil, et. al., “An Audio-Based Personal Memory Aid”, *Ubicomp 2004*, Vol.3205, pp.400-417, Oct. 2004

[3] DPW Ellis, et. al., “Minimal-impact audio-based personal archives”, *CARPE'04*, pp.39-47, Oct. 2004

[4] 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