音響ライフログへのアノテーションのための話者と場所の自動分類
Automatic Classification of Speaker and Location for Annotating Audio Life-Log

山野 貴一郎
Kiichiro Yamano
法政大学大学院情報科学研究科情報科学専攻
E-mail:kiichiro.yamano.rr@gs-cis.hosei.ac.jp

Abstract

Life-log applications which are the log of personal life experiences recorded on cameras, microphones, GPS devices, etc. are studied. A purpose of this study is automatic classification of locations and speakers involved in audio life-log recorded by a wearable microphone. In this study, 80 hours audio life-log data was recorded with four IC recorders. Then, a method for classifying locations where the audio life-log was recorded is suggested. This method is dividing audio life-log into fix-length segments and clustering an acoustic feature of the segments. An average precision was 49.2% and an average recall was 65.5% in case of clustering audio life-log segments assuming use of acoustic and location information. The accuracies were better than case of using only acoustic information. Moreover, a voice detection method modeling speech and non speech with mixture Gaussian distribution is proposed for speaker identification. False accept rate was 6.8% and false reject rate was 72.42%. However, the accuracies varied with situations. Therefore, multimodal processing and robustness against many situations are necessary for processing the audio life-log appropriately.

1 まえがき

個人の生活や体験を様々なセンサを用いて記録し，利用するための研究が行われている [1]．利用されているセンサはカメラ，マイク，GPS，加速度計，脳波計など様々である．また，電子メール，ウェブの閲覧履歴，購買履歴などの情報も利用されている．このような個人に関する記録をライフログと呼ぶ．映像や音声のライフログは他の情報を利用し備忘録や自動の日記作成への応用が期待されている．また，ウェブの閲覧履歴や購買履歴は商品のレコメンドシステムやマーケティング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．従って．効率的な利用のためには要約や検索の必要があり，これまで様々な試みがなされてきた．

本研究ではウェアラブルなマイクで日常生活の音を常時記録した音響ライフログと GPS によって取得した位置情報について扱う．音響ライフログは様々な音を含んでおり，多くの情報が得られる．例えば，音声からはその時の会話の内容や話者情報などが得られる．騒音や音楽からは雑踏や店にいるなど周辺の情報が得られる．また，環境音からはユーザの行動が推測できる場合がある．例えば，キーボード打鍵音やクリック音などの PC の操作音からはユーザの行動がわかる．しかし，音響ライフログはほぼ環境音や音声などが含まれていない非常に静かな部分や，含まれている音が聴取しただけでは不明な部分がある．つまり冗長な部分が多く，収録をしたデータをそのまま提示しただけでは所望の情報を探すのが困難である．このような問題に対し，従来研究では時系列での情報の提示が行われてきた [5]．

本研究では，音響ライフログを場所で分類するため，音響情報のスペクトルを用いてデータのセグメンテーション，クラスタリングを行う．さらに，従来の時系列の情報提示だけではなく，GPS 情報を併用して地図上で時系列情報をブラウジングする手法を提案する．また，音響ライフログには音声が多く含まれる部分と，全く含まれない部分が存在するため，データに含まれる音声区間を検出する必要がある．そこで本論文では音声データと非音声データの混合ガウスモデル (GMM) を定義し，音響ライフログに含まれる音声を検出する手法について提案をする．

2 ライフログデータの分類

2.1 先行研究

ライフログを有効に扱うための処理についてこれまで多くの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，ウェアラブルカメラで常時収録されたライフログ映像の検索を行うシステムが提案されている．

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 [8] 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．62.5 時間のデータに対しオフィス，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 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．また，文献 [3] 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 1 日単位のデータに分割し，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には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 ヶ月間記録した 271163 枚の画像を使って実験を行っている．

ライフログの音響情報の利用についても提案されている．文献 [4] 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可能性があるため，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．文献 [5] 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いる．データの音スペクトルに着目しセグメンテーションと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 の環境や場所の分類を行っている．

2.2 音響ライフログと位置情報の利用

音響ライフログに含まれる音響情報には様々な利用法がある．例えば，文献 [6] では音声，キーボード打鍵音，紙をめく

Supervisor: Prof. Katunobu Itou

る音の出現頻度を利用して，オフィス環境でのデスクワークとミーティングの分類をしている．文献 [7] では駅ホームにおける電車の発着音，通過音を用いて，ライフログ映像のシーン分割を行った．

音響ライフログで収録された音響情報の中で，多くの応用に役立つ情報としては音声が挙げられる．音声からは会話の内容や話者情報などを含んでいる．このような音声記録は備忘録としての用途が期待できる．しかし，収録した音響ライフログデータは音声を含んでいない部分も多い．実際に収録したデータの 3 時間分を 1 分のセグメントに分割し，聴取したところ 180 セグメント中に音声を含むものは 91 セグメントであり，半分は音声が含まれていなかった．また，状況によっては数時間にわたり音声が含まれていない場合もある．本研究で収録したデータでは，自宅や研究室で 1 人で PC 作業などをしている場合にほぼ音声が含まれていなかった．また，音声が含まれる部分が抽出できても，それが長時間になると所望の部分を探すのは困難である．そこで，検索や整理をするため音声にインデクスやタグ等のアノテーションを付与する必要がある．音声に与えられるアノテーションとしては，時間，話者，場所，会話の内容などが考えられる．会話の内容を音声認識によりアノテーションとする手法が考えられるが，音声認識の誤りや辞書に登録されていない単語を認識できないなどの問題がある．特に未知語が会話のトピックとなる場合，有効な検索が行えなくなる．そこで本研究では時間，場所，話者情報での音声データの提示を行う．

位置情報は GPS により取得が可能であるが，建物内の部屋の情報までは取得できない．そこで，同じ部屋内では背景音が似ていることを利用し，クラスタリングにより詳細な場所の分類を行う．従来の研究では 1 分のセグメントに分割をして特徴量を抽出し場所のクラスタリングを行っている．本研究でも 1 分のセグメントでのクラスタリングを行う．

また，上記のように音声が含まれていないデータがあるので，音声の検出も必須である．本研究では音声認識システムで利用されている，混合ガウスモデルによる音声検出を利用する．音声を検出することで話者情報もクラスタリングにより，同じ話者の会話データを分類できる可能性がある．話者や部屋の情報はクラスタリングを行った段階では，同じ話者や場所でのクラスタができているだけである．このようなクラスタへのインデクスはユーザにより付けられることを想定している．

以下の章ではデータ収録や音響ライフログを有効に扱うための手法について述べる．3 章では本研究の音響ライフログと位置情報の収録について解説する．音響ライフログはマイクを装着し常時収録を行うため，収録方法にも工夫が必要である．

3 データ収録

音響ライフログは 4 種類の IC レコーダーと 2 種類のピンマイクとバイノーラルマイクを用いて収録した．収録したデータはおよそ 80 時間である．レコーダーとマイクは日によって異なるものを使用した．異なるレコーダーやマイクを用いることで，デバイス間で録音時の音量が統一できなくなるため，処理が難しくなる場合がある．しかし，ライフログの収録期間は非常に長く，レコーダーの製品寿命よりも長くなると考えられる．従って，デバイスの違いによる影響に頑健な処理が必要である．バイノーラルマイクはイヤホン型のマイクで，本来は両耳に装着をして利用するマイクである．しかし，両耳に装着をした状態での長時間の収録はユーザの負担となるため，図 1 のように肩から提げて収録した．サンプリング周波数は 48kHz，量子化ビット数はレコーダーによって異なり，24 ビットもしくは 16 ビットで収録をした．以上のデバイス，条件で日常生活の音を収録した．主な収録音を表 1 にまとめた．

本研究では処理をした音響ライフログは地図上に提示するため，音響情報の収録と同時に位置情報の記録も行っている．位置情報は GPS を用いて 5 秒間隔で取得している．時間は IC レコーダーに収録開始時刻が記録されている．

図 1. 収録時のバイノーラルマイクの装着方法

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，音楽，踏切の警告音

3.1 位置情報の利用

音響ライフログは GPS によって取得した緯度と経度を用いて地図上に表示する．図 2 に GPS から得た情報の例を示す．図 2 の各線はユーザと凡例で示した 6 か所の距離を時系列で表している．ユーザは最初は大学におり，そのあと徒歩で帰宅している．自宅までには 3 か所のコンビニとスーパーを通過する．このような場所の緯度と経度は GoogleMaps から取得できる．距離は GPS で取得したユーザの緯度, 経度と GoogleMaps によって取得した場所の緯度，経度を座標と考えユークリッド距離を求めた．図では屋内にいるなど GPS が位置情報を取得できない場合直線で補間してある．

図 2 においてユーザは最も距離が小さい場所の近くにいると推測される．そのように考えると，実際のユーザの動きと図 2 の動きは一致しており，大体の位置情報は取得できている．このような位置情報は後述する音響情報を用いたクラスタリングに役立つ可能性がある．音響情報を用いたクラスタリングでは研究室や教室など室内の分類をしている．前処理として大体の場所を取得できれば，分類すべき室内の場所が限定され，クラスタリング精度の向上が期待できる．

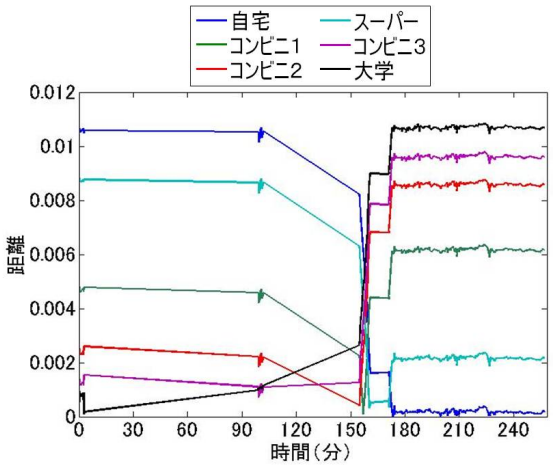

図 2. ユーザと周辺施設との距離

4 ベイズ情報量規準による可変長セグメンテーション

従来研究ではベイズ情報量規準 (BIC) を用いて，類似した音が含まれるセグメントを作る手法が提案されている [5, 9]．この章では BIC による音響ライフログのセグメンテーションを実験を行った．また，場所の移動の検出について評価，考察した．

4.1 BIC

BIC とはモデルの評価基準である．n 個のデータ $X = x_1, x_2, ..., x_N$ に関する r 個のモデル候補を $M = M_1, M_2, ..., M_r$ とする．このとき $L(X, M)$ をモデル M の最大尤度とし，$\#(M)$ をモデル M のパラメータ数とすると M の BIC は次式のように表される．

$$BIC(M) = \log L(X, M) - \lambda \frac{1}{2} \#(M) \log(N) \qquad (1)$$

右辺で減算している項はモデルのパラメータ数が増えると増加するペナルティである．学習データを入力とする尤度，つまり最大尤度は一般にパラメータ数が増えると値が大きくなるが，ペナルティによりオーバートレーニングを考慮したモデルの評価ができる．したがって BIC の値を最大にするモデルが最適と考えられる．

4.2 BIC を利用したセグメンテーション

BIC を利用して文献 [5] では音響ライフログのセグメンテーションを行っている．また，文献 [9] ではテレビ番組を対象として，話者，環境，チャンネルの変化の検出を行っている．

これらの文献では長時間の音データを固定長のフレームに分割し，そのフレーム内で 3 つのモデルを想定する (図 3)．1 つはフレーム全体をモデルで，他の 2 つは時刻 0〜t のモデルと t+1〜N までのモデルである (フレーム長は N)．もし時刻 t にセグメントの境界がある場合，モデル M_1, M_2 の最大尤度が高くなる．つまり，共分散や分散が小さくなる．したがって，式 (2) が最大値となる時刻 t を求めることでセグメント境界を得られる．

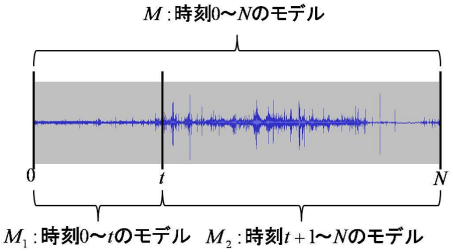

図 3. フレーム内のモデルの分割

$$R(t) = N \log |\Sigma| - N_1 \log |\Sigma_1| - N_2 \log |\Sigma_2| \qquad (2)$$

N, N_1, N_2 と $\Sigma, \Sigma_1, \Sigma_2$ はそれぞれ M, M_1, M_2 のデータ数と共分散である．しかし共分散は導出できない場合があるので，実際には共分散の対角成分 (分散) を用いた．これより時刻 t の BIC は式 (3) になる．

$$BIC(t) = R(t) - \lambda P \qquad (3)$$

但し，P は d を特徴量の次元数として式 (4) で求めたペナルティである．

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N \qquad (4)$$

結局，BIC によって求めたセグメント境界の推定時刻 $\hat{t}$ は，式 (3) が最大となる時刻 t である．

実際にあるフレームに BIC を適用した結果を図 4 に示す．最大値になっている時間がフレーム境界の候補である．

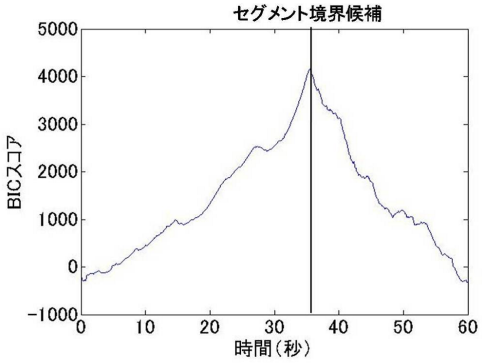

図 4. フレーム内の BIC スコアの変化

4.3 特徴量

特徴量は正規化平均スペクトル包絡を用いた．スペクトル包絡は短時間スペクトルにメル周波数軸上でフィルタバンク分析をして求める [10]．

短時間スペクトルは 85.3ms のハニング窓を 42.7ms ずつシフトさせて切り出した波形に FFT をして求めた．フィルタバンク分析では短時間スペクトルをメル周波数軸上で 600 の幅の三角窓を 300 ずつシフトして切り出し，各帯域のパワーの和を求めた．このような処理を行うことで短時間スペクトルが 12 点に集約され，図 5 のような概形が得られる．スペクトル包絡は細かいスペクトルの違いに対して頑健な処理が行える．1 分のセグメントからスペクトル包絡は複数得られるのでそれらの平均を求め，正規化を行いセグメントの特徴量とした．正規化は平均スペクトル包絡の 12 点のパワー平均を包絡全体から減算することで行った．

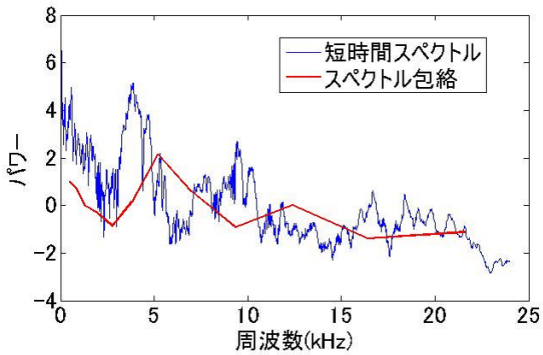

図 5. スペクトル包絡

4.4 実験

2 日分 (約 9 時間 30 分) の音響ライフログデータを BIC スコアによってセグメンテーションした．フレーム長は 1 分でフレームシフトは 30 秒である．フレームがオーバーラップしているため 1 つのセグメント境界に対して複数のピークが検出されるので，オーバーラップ部分を加算した．

全てのピークをセグメント境界とすると，スコアが低いピークが検出されたり，1 秒以下の非常に短いセグメントが生成されたりする．そこで，BIC スコアが 2500 以上のみをセグメント境界候補とすることにした．また，これだけではピーク付近に発生する小さなピークを検出してしまうので，ピークの前後 5 秒に発生する小さなピークは無視した．

評価はデータの場所が変化する点に手動で付与したセグメント境界と BIC スコアによって求めたセグメント境界を比較して行う．具体的には手動で付与した境界と最も近い BIC によって求めた境界のずれが 5 秒以内なら正解とした．ただし，1 か所にいる場合でも音声などで音響的な変化が生じた場合高

い BIC スコアになりセグメント境界となることがある．しかし，同じ場所でセグメンテーションに関してはクラスタリングまで含めた評価が必要なので，この実験では部屋から部屋への移動などが正確に検出できるかのみに着目して評価をした．

4.5 実験結果

前節の条件で実験したところ 663 のセグメント境界が求められた．手動で付与したセグメント境界は 35 個所である．その 35 個所に最も近い境界との誤差を求めたところ誤差の平均が 19.6 秒，最短誤差が 0.1 秒，最長誤差が 70.9 秒で，誤差が 5 秒以内の境界は 10 で正解率は 28.6% であった．誤差の分布は図 6 のようになった．

また，生成されたセグメントは平均時間が 56.7 秒，最長が 30 分 40 秒，最短が 5.2 秒であった．セグメント長の分布は図 7 のようになった．

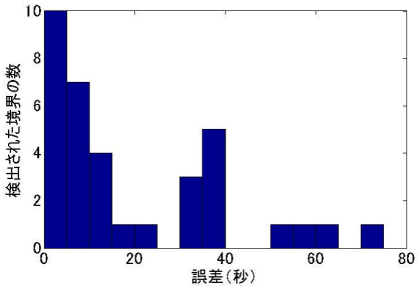

図 6. 手動の境界と BIC スコアで求めた境界との誤差の分布

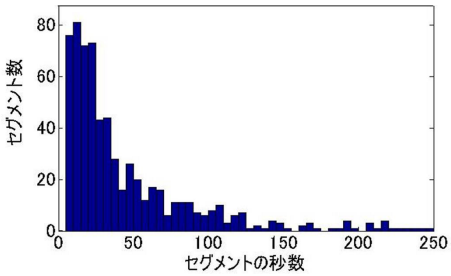

図 7. セグメント長の分布．縦軸がセグメント数，横軸が長さ (秒) である．

4.6 考察

部屋の移動を検出しやすいように低めに閾値を設定し，多くの境界を求めたが誤差の大きい境界が多かった．これは部屋の移動よりも音声やその他の環境音による音響的変化のほうが大きいために起こったと考えられる．実際に高い BIC スコアのときに収録されていた音としては，救急車のサイレンや踏切の警告音などの特徴的な音であった．また，フレームシフトを 30 秒としたが 1 分以内で場所が 2 回切り替わる場面もあったので，フレームシフトはより短いほうが適切な可能性がある．

実際のデータを聴取すると，部屋を移動したり屋外に出た場合に背景雑音が変わるので場所の違いを認識することは可能である．しかし，実験結果ではほとんど境界を識別できていない．誤差を 10 秒まで許したと考えても 50% 程度である．つまり，特徴量に違いが出ていない可能性があり，ほかの特徴量を試す必要があるかもしれない．

5 音響情報を用いた場所のクラスタリング

音響ライフログが収録された場所をクラスタリングにより分類する実験を行った．実験に使用したデータは 2 日分で約 9 時間 30 分である．このときのレコーダーは 1 日が YAMAHA POCKETRAK CX で別の 1 日が EDIROL R-09HR である．データに出現した場所は研究室，廊下，大学構内の屋外，路上，自宅，コンビニエンスストア，スーパーマーケットである．この評価データを 1 分のセグメントに分割し，セグメントから特徴量を抽出しクラスタリングを行った．総セグメント数は 571 セグメントである．

以上のデータを用いて 2 つの実験を行った．1 つは GPS を使わないと仮定しデータに現れた全ての場所をクラスタリングする実験で，もう 1 つは GPS で大学という場所を取得できたと仮定し，大学内のデータだけをクラスタリングするものである．

5.1 クラスタリング手法

クラスタリングに用いた特徴は正規化平均スペクトル包絡で，特徴量を k-means 法によってクラスタリングした．k-means 法では事前にクラスタ数 k を決めて，以下の順序でクラスタリングを行う．

1. ランダムに k 個のセグメントを選びクラスタの中心とする．
2. 選ばれたセグメントの特徴量からの各セグメントの特徴量へのユークリッド距離を求める．
3. 距離が最小のクラスタにセグメントを含める．
4. 各クラスタの重心を求め，それを新しいクラスタ中心とする．
5. クラスタ中心から各セグメントまでの距離を求めて距離が最小のクラスタに含める
6. クラスタ中心が移動しなくなるか．既定のクラスタリング回数を超えるまで 4,5 を繰り返す

クラスタ数 k は全ての評価データをクラスタリングする場合には 7，大学内のデータをクラスタリングする場合には 3 とした．

5.2 クラスタリング結果

クラスタリングの結果を表 2, 3 に示す．評価は各クラスタに手動でラベル付けを行い，再現率と適合率を用いて行った．研究室の適合率と屋外の再現率が高く，廊下の適合率が低かった．

表 2. 大学内のデータのみをクラスタリングした結果．クラスタには手動で場所のラベルを付けた．各行がクラスタに含まれるセグメント数である．例えば，研究室のクラスタには研究室のセグメントが 210，廊下のセグメントが 3，屋外のセグメントが 1 含まれる．

	研究室	廊下	屋外	適合率	再現率
研究室	210	3	1	98.1%	51.5%
廊下	180	6	0	3.2%	50.0%
屋外	18	3	18	46.2%	94.7%

5.3 考察

研究室と屋外のクラスタにおいて場所が取得できたと仮定した方が再現率と適合率が高い結果となった．屋外と研究室はスーパーのセグメントと混同されやすく，場所情報を利用することでクラスタリング精度を向上させる可能性があることがわかる．廊下の再現率は場所を大学に限定した方が高いが，適合率は場所を限定することで低くなった．これは，他のクラスタと誤りやすく，セグメントの数も多い研究室セグメントが，大学に限定することで，むしろ占める割合が大きくなったためである．

本実験で用いたデータを収録した環境ではスーパーやコンビニ，自宅，路上は部屋の分類がないため，GPS で得た場所がそのままクラスタとなる．以上より位置情報を利用することで音響情報のクラスタリングの精度を補完できると考えられる．また，路上については位置は取得可能であるが，移動することが多いためクラスタリングには工夫が必要である．本論文では用いたデータでは 10 分程度の移動なので，それを 1 つのクラスタとしても問題はないが，移動が数時間になった場合のクラスタリング方法も考えなければならない．

研究室のセグメントが分散した原因としては，収録される音が状況により異なることが理由として考えられる．研究室で主に現れる音には表 1 に示したような音が挙げられ，その中で特に音声が含まれているときと含まれていない時で音響的特徴

表 3. 全ての評価データをクラスタリングした結果．クラスタには手動で場所のラベルを付けた．表 2 と同様に各行がクラスタに含まれるセグメント数である．

	研究室	廊下	屋外	自宅	コンビニ	路上	スーパー	適合率	再現率
研究室	113	1	1	4	2	2	3	89.7%	27.0%
廊下	37	4	1	0	1	1	1	8.9%	33.3%
屋外	40	2	8	0	1	13	2	12.1%	42.1%
自宅	0	0	0	78	0	0	0	100%	94.0%
コンビニ	51	0	2	0	2	0	0	3.6%	33.3%
路上	67	1	0	1	0	10	0	12.7%	29.4%
スーパー	100	4	7	0	0	8	3	2.5%	33.3%

の差が大きい可能性がある．会話の場合，数分間にわたり音声が収録されている場合がある．従って，会話をしていない時のデータとはセグメントの特徴量が大きく異なる．特に収録者の音声は音量が大きく，スペクトルの形に大きく影響すると考えられる．音響ライフログの応用によっては，これを利用して会話時と会話をしていないときという状況で，クラスタを分けることが有効であるかもしれない．

また，1 分毎にセグメントのクラスタが頻繁に変化することはほとんどない．従って，文献 [8] のように時間的に近いセグメントを同じクラスタに分類されやすくすることで，研究室のセグメントを 1 つのクラスタにまとめることが可能かもしれない．

本論文の実験では異なるレコーダーを用いた 2 日間のデータを用いたが，正規化によりレコーダーの違いに関わらずクラスタリングができていた．これは表 2 の屋外のセグメントがほぼ 1 つのクラスタに含まれたことからもわかる．しかし，天候や部屋の空調などの条件が異なると音響的な特徴が変化する可能性もある．このような違いによる影響を検証するには，長期のデータを用いた実験が必要である．

6 GMM による音響ライフログの音声区間検出

本研究では音響ライフログの音声有効利用のために，音響ライフログに含まれる音声データ対して話者のタグを付与することを想定している．タグを付与するためには音声区間を検出し話者別に分類する必要がある．本節では音声と非音声区間から特徴量としてメル周波数ケプストラム係数 (MFCC) を抽出し，混合ガウス分布でモデル化して音声区間の検出をする手法を述べる．

6.1 音声区間検出手法

GMM は特徴量の単一の次元を複数のガウス分布で表現するモデルであり，複雑な分布を持つデータをモデル化するのに適したモデルである．音声区間検出をするためには音声と非音声の 2 つのモデルを利用するため，特徴量の分布が複雑になることが考えられる．したがって，複雑な分布を表現できる GMM を用いた．また，音声認識システム Julius でも GMM を利用し音声区間検出を行っている [11].

モデルを定義するために用いた特徴量は MFCC である．MFCC は音韻に関する特徴量で音声認識や話者識別などに利用される．MFCC はスペクトルのフィルタバンク出力に，離散コサイン変換をすることで求められる．

音声区間は音声データを音声と非音声のモデルに入力し，音声の尤度が高い区間を音声区間として検出して求める．

6.2 音声区間検出実験

音声と非音声を GMM によって学習し，音響ライフログに含まれる音声の検出実験を行った．音声は JNAS の音素バランス文読み上げデータによって学習した．発話者は男女各 100 名で，1 人につき 50 文を読み上げている．非音声の学習は実際の音響ライフログの非音声区間を 5 時間分利用した．

MFCC を導出に利用した短時間スペクトルは 25ms のフレームを 10ms ずつシフトしながら求めた．フィルタバンク出力は 24 次で，ケプストラムは 13 次まで利用している．しかし，音響ライフログ中の音声のパワーはマイクからの距離の影響などでばらつきがあるので，0 次成分 (パワー) は使用していない．また，特徴量として MFCC の時間方向の 1 次差分である ΔMFCC も用いている．結局，特徴量は MFCC12 次元と ΔMFCC12 次元の 24 次元である．

評価データは学習データとは異なる音響ライフログ 2 時間分である．そのうち 1 時間はゼミを行っている状況で PC の騒音が含まれており，もう 1 時間は雑談で騒音は少ない環境で収録されている．この評価データに手動で音声区間の境界を付与し，自動検出した境界と比較をして評価をする．手動で付与した音声区間は全体で 1578 区間，およそ 61 分である．評価には式 (5)，(6) の False Accept Rate(FAR) と False Reject Rate(FRR) を用いる．N_{FA} は非音声を音声として誤検出した時間長，N_{FR} は音声を非音声として検出した時間長，N_{ns} は非音声の時間長，N_s は音声の時間長である．

$$FAR = N_{FA}/N_{ns} \times 100 \quad (5)$$

$$FRR = N_{FR}/N_s \times 100 \quad (6)$$

6.3 実験結果

実験結果を表 4，5 に示す．表 4 では全てのデータの FAR と FRR で，表 5 ではユーザ (収録者) とユーザ以外の話者別で FRR を求めた．表 5 で FAR は，評価した話者以外の音声が検出されることで増加するため，評価指標として用いていない．状況や話者によって FRR に大きな違いが生じた．全体として音声を棄却しやすい傾向があった．

表 4. 全ての話者の音声検出結果

	FAR(%)	FRR(%)
ゼミ	3.26	80.93
雑談	8.29	52.54
全体	6.8	72.42

表 5. 話者別の FRR

	ユーザ (%)	ユーザ以外 (%)
ゼミ	62.89	81.19
雑談	28.02	63.68
全体	31.45	77.16

6.4 考察

状況によって結果に違いがあったが，いずれも低い検出率となった．音声認識はマイクに向かって発話した音声を対象としているのに対し，音響ライフログに含まれる音声は，マイクに向かって意識的に発話したものではない．ゆえにはっきりと発話されていない，雑音に埋れてしまって不明瞭という性質があり，検出するのが難しいと考えられる．

また，低騒音下においてもマイクから離れた場所から発せられる音声は，SN 比が悪くなり検出が難しいと考えられる．

マイクに近く，騒音も小さい雑談時の場合はユーザ自身の発話で FRR が 28.02% で，ユーザ以外の音声よりも高い精度で

検出できている．しかし，ユーザ自身の発話でも騒音が大きい場合は検出が難しい．

以上より，音響ライフログから音声検出を行う場合には騒音に頑健な処理が必要であることがわかった．

また，複数人の発話が重畳する場合もあったので，検出した音声を話者別に自動分類する場合には，発話オーバーラップの対策も考えなければならない．

7 データの提示手法

7.1 ユーザの要求

ユーザのデータの提示要求としては，

1. 5 月 1 日に研究室で A さんが話していたことを聞きたい
2. 日付は定かではないが，夕方ごろに A さんと話していたことを聞きたい
3. 5 月 1 日に A さんと B さんと 3 人で話していた時の会話を聞きたい

というような様々な要求が考えられる．そのような場合に，本論文で提案をした部屋でのクラスタリングが役立つ．話者のクラスタリングについては，1 分以上のセグメントを用いると 1 つのセグメントに複数の話者の音声が入り込む場合が多い．従って，6 章で提案したような，音声区間検出が必要となる．このような処理によりアノテーションを付けることで，上記の 1 の要求に対しては 5 月 1 日の研究室での A の発話を提示することで要求に応えられる．また，2 の要求に対しては夕方の A の発話を提示することができる．3 に対しては，追加の処理として A，B とユーザの発話が集中している時間帯を探すことで，要求に一致した音声データを提示することができる．

7.2 音響ライフログ提示の例

実際に場所，日付，時間，話者から音響ライフログを提示を行う例を図 8 に示す．図 8 のシステムには Google Maps API[*1]を用いている．はじめに，ユーザは GPS により取得された場所のマーカーを選ぶ．さらに部屋を選ぶと，左部分に日付別にデータがツリー構造で表示される．日付を選ぶと，その日の音響ライフログに登場した話者が表示され，話者を選択するとその話者の発話が時間別に表示される．時間をクリックするとその時の音声が再生される．

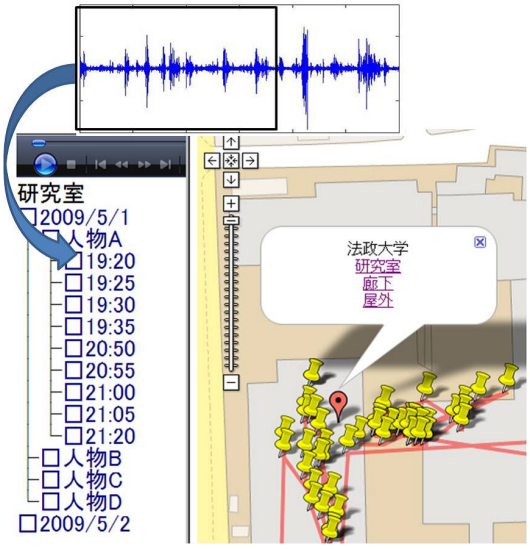

図 8. データの提示

8 あとがき

本論文では音響ライフログの音声情報を有効活用するための，データのセグメンテーションとクラスタリング，音声区間検出の手法を提案した．

まず，ベイズ情報量規準を用いて音響ライフログの可変長セグメンテーションを提案した．手動で付与したセグメント境界と BIC スコアによって求めた BIC スコアの比較を行ったところ平均 19.6 秒の誤差があった．また，場所の変化よりもサイレンや踏切などの環境音や，音声で高い BIC スコアとなりセグメンテーションの障害となった．このような突発的な音を省くには加速度計を利用するのが有効である可能性がある．

また，建物内での位置情報を得るための手法として，音響情報を用いた場所のクラスタリングを提案した．実験として，2 日分の音響ライフログをセグメントに分割し，場所のクラスタリングを行った．実験は GPS により場所を取得したと仮定した場合と，GPS を用いない場合の両方を行い，適合率と再現率で評価をした．その結果，GPS 情報の利用によって大学内のデータの適合率の平均は 36.9% から 49.2% に，再現率の平均は 34.1% から 65.4% に改善され，クラスタリングの精度が向上することが確認できたが，より長期のデータを用いた実験や GPS の精度の検証が必要である．

また，音響ライフログに含まれる音声区間を GMM を用いて検出した．実験の結果 False Accept Rate が 6.8%，False Reject Rate が 72.42% で音声を棄却しやすい傾向にあった．原因としてはマイクから離れた発話や高騒音下での発話によって，SN 比が悪い音声が多いことが考えられる．雑音の種類によっては除去できる可能性があるので，雑音除去の効果を検討する必要がある．

以上より音響ライフログの処理においては，音響情報のみではなく他のセンサを利用するべきである．しかし，常時収録を行うためユーザの負担にならない収録方法を考えなければならない．また，状況によっては一般的に行われている音響処理が有効でない場合もあるので，様々な環境下で有効な手法や状況に応じて処理手法を変えるといった工夫が必要となる．

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] K Aizawa, “Digitizing Personal Experiences: Capture and Retrieval of Life Log”, *Proc. Multimedia Modelling Conf.*, pp.10-15, Jan. 2005.

[3] Aiden R.Doherty et al., “Automatically Segmenting LifeLog Data into Events”, *In WIAMIS 2008*, pp.20-23, May 2008.

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

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

[6] 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”，**情処論講** *68* **回**．pp.4_81-4_82．2006.

[7] 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.

[8] Wei-Hao Lin et al.,“Structuring Continuous Video Recordings of Everyday Life Using Time-Constrained Clustering”, *In IST/SPIE Symposium on Electronic Imaging*, Jan. 2006.

[9] Scott Shaobing Chen et al.,“Speaker. Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion” ,*Proc. DARPA Broadcast News Workshop*, pp.127-132, 1998.

[10] D. O' Shaughnessy, “Speech Communication: Human and Machine”, Reading. MA: Addison Wesley, 1987.

[11] Akinobu LEE, “The Julius book ”, Oct. 2009.

[*1] Google Maps API http://code.google.com/intl/ja/apis/maps/