スペクトル情報を用いたライフログ映像のシーン検出

山野貴一郎　　伊藤克亘
法政大学情報科学部

E-mail: n04k1035@cis.k.hosei.ac.jp

ライフログ映像を効率よく扱うためにはライフログ映像をシーンごとに検出しなければならない．シーン検出は，映像の色相の変化に着目し行われることが多いが，それだけでは精度が不十分である．本論文では映像のみでは困難な駅での電車待ちシーンを対象に，スペクトル情報を用いてシーン検出する手法を提案する．特徴抽出には，パワースペクトル，およびその包絡を検討した．実際に収録したデータを用いて，シーンを検出するためのショット識別実験を行った結果，平均識別率はパターン間距離 (パワースペクトル平均) が 37%，パターン間距離 (パワースペクトル包絡) が 27%，パワースペクトル包絡の確率モデルが 71% であった．

1　まえがき

計算機やビデオカメラの小型化，ハードディスクの容量増加により，体験を常時記録することが可能になりつつある [1]．このように記録された人の体験や生活の記録をライフログという．

人の生活を常に記録するという性質上ライフログは長時間に渡り記録されるものであり，データ量が膨大で，かつ冗長であるという特徴がある．よって，後から参照したいと思った出来事を選び出すためのインデキシングが必要である [2]．

インデキシングを行うためには，ライフログ映像をユーザにとって意味のあるシーンごとに分けなければならない．ライフログは常時記録されるので，ユーザの負担にならないためにもインデキシングやシーンの検出は自動で行われるのが望ましい．

ライフログ映像のシーンの検出は，文献 [3][4] のように映像情報に基づき行われることが多い．映像情報を用いたシーン検出には文献 [3] のように，映像の色相情報の変化を用いる手法がよく行われている．

色相の変化に基づいたシーン検出は，屋外から屋内への移動，部屋から廊下への移動など，移動によるシーンの変化に対し有効であるが，カメラの前を横切って移動する妨害物がある場合に，シーンを誤って検出してしまうことがある．人や車が横切る程度の短時間の妨害物であれば，エラーとして処理することができるが，電車の通過，停車などの時間が長い妨害物の場合は，意味のないシーンの検出を起こす可能性がある．

このように映像情報だけでは，シーンの分類が十分に行えない場合に，加速度センサー，ＧＰＳ，脳波など他の情報と組み合わせてライフログ映像のシーン検出を行うことがある [5]．

また，文献 [6] では，音響情報を用いてオフィス環境でのデスクワークとミーティングなどの行動を分類している．検出する必要のある行動 (紙をめくる音，キーボードの打鍵音，人の声) をスペクトル情報を用いて検出している．検出には，平均スペクトルと学習データの 2 乗誤差の平均と標準偏差を利用している．さらに倍音構造の特徴を用いて人の話し声を区別している．検出した音声の出現頻度で行動をモデル化し，研究室内でのデスクワークとミーティングを分類している．

そこで，本論文では駅構内や電車内において，音響情報を利用することにより，電車待ちのシーンの検出を行う手法を提案する．

2　シーン，ショットの検出手法

2.1　シーン，ショットの定義

本研究ではライフログ映像におけるシーンをユーザにとって一つの意味を持つ映像の区切りのこととする．

電車待ちを扱う本研究でのシーンとはホームで電車を待ち始めてから電車に乗り込むまでとする．ホームで電車を待っている際には電車の通過や発着がある．これらは，画像情報のみでは，それまでと異なるシーンとして検出される可能性がある．しかし，本研究で想定する状況では，ユーザにとってインデキシングする意味を持たないと考え，電車の通過や発着をショット

として扱うこととする．

ショットとはシーンを細分化したもので，ショットが集まることでシーンが構成される．

本論文ではショットを以下の 7 つに分けて考える．

- 電車前方のホームでの発車時（以下，発車 F）
- 電車後方のホームでの発車時（以下，発車 R）
- 電車前方のホームでの停車時（以下，停車 F）
- 電車後方のホームでの停車時（以下，停車 R）
- 電車通過時（以下，通過）
- 通過，発車，停車がない時のホームでの待機中（以下，待機）
- 車内

収集したデータを聴取して，人間が聞くだけで区別できるかどうかを基準にショットの区分を決定した．ホームで発車音や停車音を聞いた場合，電車の前方で聞いた場合と後方で聞いた場合で音が異なるので，データの解析を行う場合には別のショットとして扱うべきだと考えられる．

図 1 にシーン，ショットの例を示す．電車待ちというシーンは上記の車内以外ショットで構成される．つまり，図 1 に示したように待機や停車 F などが集まって電車待ちというシーンになり，電車に乗り込んだときにシーンが変わる．具体的には車内のショットが検出された場合に，そのショットが検出される前までを電車待ちとしてシーン検出するため車内も 1 つのショットとして扱う．

図 1　シーンとショットの例

2.2　音響情報を用いたシーン同定

2.2.1　提案手法

音響情報を利用したシーン同定の流れについて図 2 に示す．

まず，入力音を短時間スペクトルに分ける．時系列で前から波形を切り出し短時間スペクトル (42.7ms) にしていくが，時間をシフトするときに前の切り出した波形と半分オーバーラップするようにシフトする．次に，個々の短時間スペクトルにフィルタバンク分析を行う．フィルタバンク分析を行うことで，入力音の各帯域における特徴を抽出する．そして，入力音と事前に調査した各ショットのプロトタイプを比較する．もし，ショットが車内であったら，そこまでが 1 つのシーンと見なす．ショットが車内以外の場合は，ショットとみなしシーンの誤検出を防ぐ．

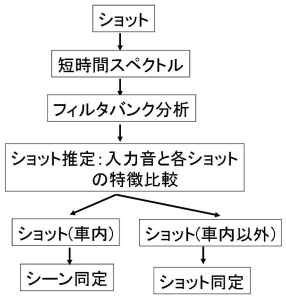

図 2　シーン同定の流れ

2.2.2　フィルタバンク分析

フィルタバンク分析とは周波数軸上に配置された複数のフィルタ群の出力に基づき行うスペクトル分析のことである [7]．

本研究ではメル周波数軸上に三角窓を配置し，短時間スペクトルにフィルタバンク分析を行う．つまり，メル周波数上でフィルタ密度（フィルタバンクの中心周波数間隔）を一定にした三角窓 BPF でパワースペクトルを切り出し，帯域ごとのパワーの総和を算出する．

三角窓は窓の半分がオーバーラップするように配置する．また，メル周波数軸上でフィルタバンク分析を行うのは，人の感覚尺度に近いので，人が聞いて違いがわかる音の検出に向いているからである．

以上のようにして，各帯域の和であるパワースペクトル包絡が求められる．フィルタバンク分析を行う前のパワースペクトル平均とフィルタバンク分析後のパワースペクトル包絡の比較を図 3 に示す．図は縦軸が対数パワーで，横軸が周波数 (Hz) である．図 3 よりフィルタバンク分析を行うことで，パワースペクトルの概形が分かりやすくなることが読み取れる．

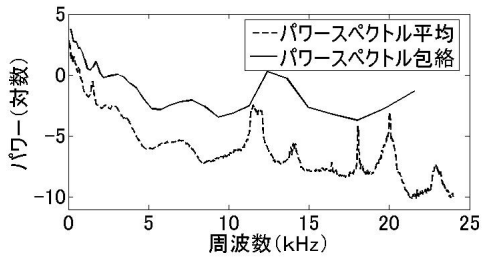

図 3　平均パワースペクトルと平均パワースペクトル包絡

パワースペクトル包絡はパワースペクトルよりも周波数上の特徴が集約されている．そのため，後述するパターン間距離によるショット識別ではよりショットごとの特徴が明確になり，距離の差が出やすいと考えた．

また，フィルタバンク出力の帯域ごとの分布を確率モデルに用いた．

2.2.3 メル周波数

メル周波数とは音の高低に関する人間の感覚尺度であり，その値は実際の周波数の対数に大略対応する．メル周波数は式 (1) で求められる [7].

$$\mathrm{mel}(f) = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (1)$$

f はメル周波数に変換する前の通常の周波数である．

2.3 各ショットのプロトタイプ抽出

シーンの検出を行うために，入力された音と各ショットのプロトタイプの比較を行う．そのためには各ショットのプロトタイプを事前に抽出しておかなければならない．比較はフィルタバンク分析により得られる各出力のパワーの帯域和によりなされる．よって，いくつかの同一ショットから得られたパワーの帯域和の平均値を求めることで，各ショットのプロトタイプとする．

求めた各ショットのパワーの特徴を図 4，5 に示す．

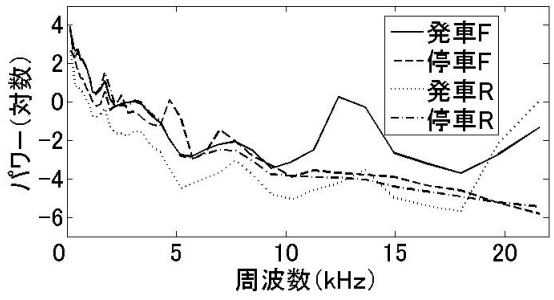

図 4　ショットの平均パワースペクトル包絡 (1)

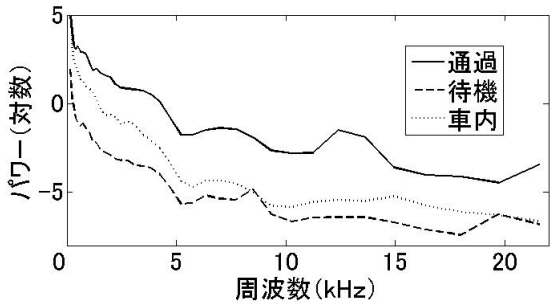

図 5　ショットの平均パワースペクトル包絡 (2)

また本論文では，入力音と各ショットのパワースペクトル平均のプロトタイプ (図 6，7) での比較も行う．

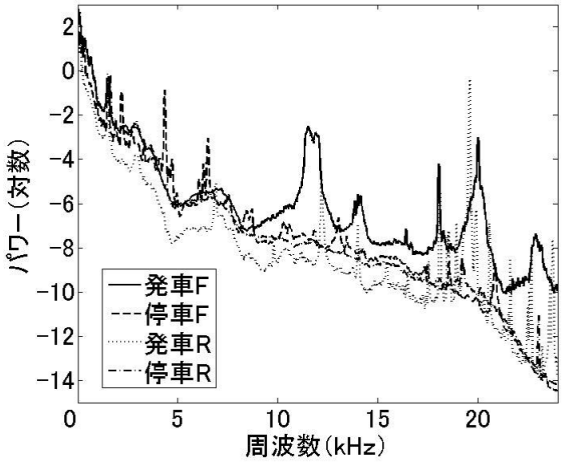

図 6　ショットの平均パワースペクトル (1)

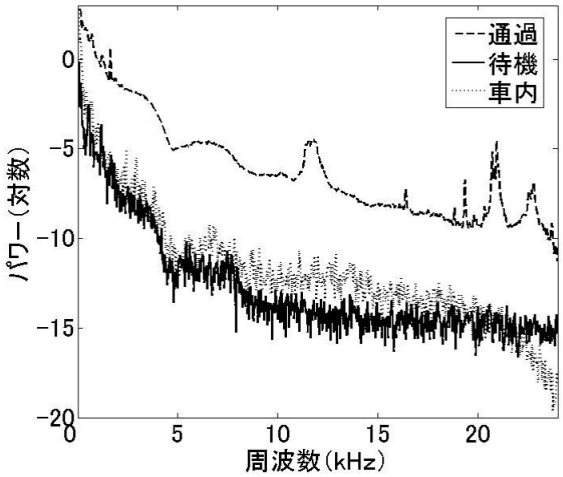

図 7　ショットの平均パワースペクトル (2)

パワーの平均を算出するために解析した学習データは，3 章で述べる方法で収録した内の 4 時間分の中から手動で各ショットを切り出したものである．

2.4 パターン間距離によるショット識別

以上のようにして得られた各ショットのプロトタイプと入力の短時間スペクトルとのパターン間距離を式 (2) で定義する．$SC_i(f)$ はショット i の特徴である．$I_n(f)$ は入力音の n 番目のフレームから短時間スペクトルを求め，フィルタバンク分析を行い，各帯域の和を算出したものである．

$$\mathrm{shot}(n) = \underset{i}{\mathrm{argmin}} \left[\int |SC_i(f) - I_n(f)| df \right] \quad (2)$$

識別する区間の全フレームに対しショット候補を推定し，最も多いショット候補を当該区間のショットとして推定する．

2.5 確率モデルによるショット推定

データの短時間スペクトルのフィルタバンク出力は帯域ごとの対数正規分布として推定される．そこで，各ショットをフィルタ数と同じ次元をもつ対数正規分布の確率モデルとして，入力されたショット音に対する尤度を求めショットを推定する．確率モデルは学習データの短時間スペクトルのフィルタバンク出力の対数をとった値の分布から平均, 共分散を求め推定した．尤度は式 (3) で求める．

$$\mathrm{shot} = \underset{i}{\mathrm{argmax}} \left[p(x|SC_i) \right] \quad (3)$$

$p(x|SC_i)$ が尤度である．x が入力音を短時間スペクトルに分けて, 個々にフィルタバンク分析を行い，出力の対数をとった対数パワースペクトル包絡の平均である．また，SC_i は各ショットを i で区別したものである．各ショットのモデルに対し尤度を計算，尤度が最も大きいものをショットとして推定する．

3 ショット識別実験

3.1 データ収録

各ショットのプロトタイプ，確率モデルを求めるための学習データ，それらの手法を実験するためのテストデータの収録について述べる．

データは 2007 年 9 月 28 日から同年 12 月 18 日に JR 三鷹駅 (5，6 番線ホーム)，吉祥寺駅 (3，4 番線ホーム) 及びその 2 駅間の車内で録音したものである．

収録を行った時間帯は 10 時〜16 時の間で，11 時〜13 時位に収録したものが中心である．

データとなる駅ホームでの電車の音はホームの両端で, 向かいのホームの方を向き録音する．データはバイノーラルマイク（adphox BME-200）を装着して PCM 録音機 (EDIROL R-09) を用いて収録した．録音条件はサンプリング周波数 48kHz，離散化ビット数 24 ビットである．

上記の方法で，まず三鷹駅のホーム前方で 15 分録音した後に東京方面の電車に乗り，吉祥寺駅のホーム後方で降車した後に再び 15 分録音し八王子方面行きの電車に乗る．再び三鷹駅で降車し，今度は三鷹駅のホーム後方，吉祥寺駅のホーム前方での収録を行う．総収録時間は 11 時間である．

昼ごろの時間帯だと，三鷹駅では東京方面の電車が 1 時間に 15 本程度，吉祥寺駅では八王子方面の電車が 1 時間 10 本程度の頻度で停車する．吉祥寺駅の方が本数が少ないが，通過電車があるので結局同じ本数が発着，通過している．

収録した 2 駅に発着する車両の種類は，7 種類であるが，本研究では向かいのホームのことは考慮していないので，実験に使用したデータに現れるのは 5 種類である．また，そのうち 3 種類は特急電車で，本数が少ないので主に収録されているのは，中央線の快速電車 (2 種類) である．

車両数は中央線が 10 両で，特急が 6〜11 両である．

このようにして収録した本研究のデータのパワースペクトルは低い周波数のパワーが大きく，高い周波数のパワーが小さい．このため通常の周波数上でフィルタ密度を一定にしてに和を求めると，高い周波数では値が小さくなりすぎてしまう．一方，本論文の手法ではメル周波数軸上でフィルタ密度を一定にして和を求めているため，低い周波数は狭い帯域の和，高い周波数では広い帯域の和が求まり，通常の周波数軸上で行うより和が小さくなりすぎることはない (図 3)．

3.2 ショット識別実験

学習データとは別のテストデータを入力し，ショットの識別ができるかを実験した．用いた手法は上記したパターン間距離 (パワースペクトル包絡，平均) と確率モデルの 3 つである．

これらの手法を用いて, 収録したデータから手動で切り出したショットを入力しショット識別実験を行った．

ショットのプロトタイプ，確率モデルを決めるための学習データの量を表 1 に示す．中段がデータ数で下段がそのデータの平均時間 (秒) である．但し，待機の時間は状況により大きく変わるので，手動で切り出す段階で 10 秒程度にして切り出したため 10 秒となっている．

表 1　学習データ数 (中段) と平均時間 (秒)(下段)

発車 F	停車 F	発車 R	停車 R	通過	待機	車内
18	13	24	26	21	40	16
27	13	14	25	16	10	113

実験時の条件を表 2 にまとめた．

表 2　実験条件

短時間スペクトル	
データ長	2048 点 (42.7ms)
時間シフト	1024(21.3ms)
FFT 長	2048 点
フィルタバンク	
三角窓長	メル周波数軸上で 200
周波数シフト	メル周波数軸上で 100
フィルタ次数	38 次元

テストデータとして用いたショットの数を表 3 に示す.

表 3　テストデータ数

発車 F	停車 F	発車 R	停車 R	通過	待機	車内
13	12	12	13	10	20	8

3.3　ショット識別実験結果

識別率を表 4 に示す．表 4 は左列がパワースペクトル包絡を用いた場合で，中央列がパワースペクトルの平均を用いた場合，右列が確率モデルを用いた場合である．発着は発車 F，停車 F，発車 R，停車 R，通過を一つのショットとして考えたものである．

この識別率は入力したショットの正解の割合である．

表 4　識別率

ショット	識別率 (%) (PS 包絡)	識別率 (%) (PS 平均)	識別率 (%) (確率モデル)
発車 F	0	7.7	61.5
停車 F	8.3	8.3	41.6
発車 R	8.3	33.3	41.7
停車 R	7.7	0	76.9
通過	20.0	20.0	90.0
発着	70	73.3	98.3
待機	100	100	90.0
車内	50	100	100

3.4　識別実験考察

3.4.1　パターン間距離

パターン間距離を用いた方法では平均的に低い識別率だった.

パターン間距離でのショット識別はパワーの平均だけに依ってショットを識別している．それに対して電車の発車や停車に関するショットは，電車がホームに入ってくるときやホームから出て行くときの速度などによりパワーが異なる．このため，プロトタイプの形が類似しているショットを誤識別したと考えられる.

しかし，図 4，6 を見ると，発車 F はパワースペクトルにピークがあり他のショットのプロトタイプとはあまり似ていない．それにもかかわらず発車 F も低い識別率であった.

誤識別したテストデータからパワースペクトル平均，包絡を算出したものを見てみると，ピークがずれていたり，ピークがないものだった．電車からの音は主にモーター音や風切音などであり，モーターの回転数や電車の速度よってこれらの音の高さは異なる．このため，パワースペクトルを平均する事で得られた，ピークの平均を用いただけでは不十分だったと考えられる.

3.4.2　確率モデル

ショットをモデル化した場合には，停車 F と発車 F が特に低い識別率であった.

停車 F の誤識別されたショットは主に発車 R として誤識別された．図 8，9 に示したように，停車 F と発車 R はパワーの平均が近く分散が異なる．このためパワーの小さい停車 R のショットが入力された場合に分散が小さい発車 R の尤度のほうが高くなってしまい，誤識別が起きたと考えられる.

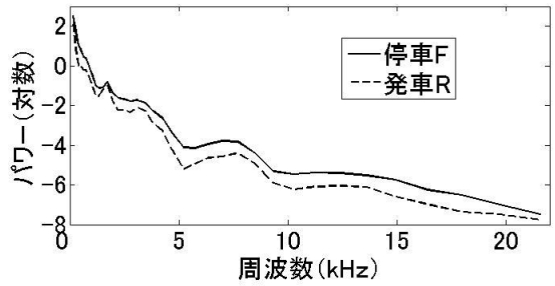

図 8　停車 F と発車 R のパワーの平均

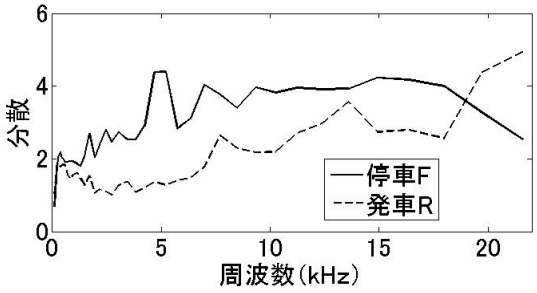

図 9　停車 F と発車 R の分散

発車 R の誤識別されたショットは主に発車 F として誤識別された．発車 F と発車 R はパワ-スペクトル包絡と分散が類似している (図 10，11)．パワーはやや発車 R のほうが小さいので，ある程度大きい音量の発車 R のテストデータが入力された場合に，誤識別が起きたと考えられる．

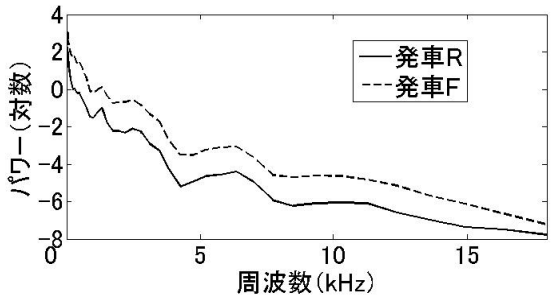

図 10　発車 R と発車 F のパワーの平均

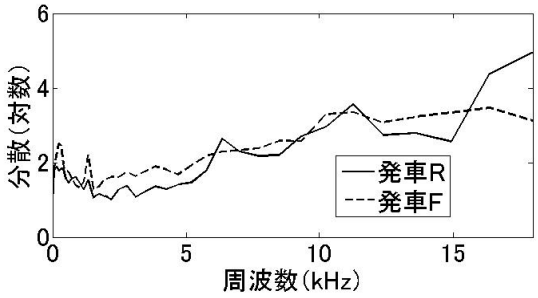

図 11　発車 R と発車 F の分散

3.4.3　発着の識別率

表 4 において，発着というショットが追加されている．これは，発車 F，発車 R，停車 F，停車 R，通過のショットを合わせたもので，5 つのショットのどれが識別されても発着として考えた．発着を用いることでも不要なシーンの検出が防げる．

識別実験では個別のショットを識別するよりも発着の方がよい結果が出ており，特に確率モデルでは 98% と高い識別率であった．これより電車の到来とその他のショットの識別はできていると思われる．

3.4.4　3 つの手法の比較

3 つの手法の平均識別率はパターン間距離 (パワースペクトル平均) が 37%，パターン間距離 (パワースペクトル包絡) が 27%，確率モデルが 71% で確率モデルが最も高かった．

また，発着を使うと仮定すると各手法の平均識別率はパターン間距離 (パワースペクトル平均) が 87%，パターン間距離 (パワースペクトル包絡) が 73%，確率モデルが 96% で確率モデルが最も高かった．

このことから本論文の実験では確率モデルが最もよくショットを識別できたといえる．

4　ショット検出実験

4.1　ショット検出方法

識別実験で最もよい結果が得られた確率モデルで，2 つ以上のショットが続いた場合にどのような検出結果が得られるかを調査するための実験を行った．

テストデータはホームで収録したデータ 2 時間分から収録者のいるホームに発着した電車の音データだけを対象とした．このようなデータから例えばショットが「待機→停車 F →待機」というようなショットの変化があるものを, 手動で切り出しテストデータとした．

テストデータには主に以下のようなショットの変化が現れた．括弧内の数字は実験に用いたデータ数である．

- 待機→発車 F →待機 (11)
- 待機→発車 R →待機 (5)
- 待機→停車 F →待機 (8)
- 待機→停車 R →待機 (8)
- 待機→通過→待機 (5)
- 待機→車内→待機 (4)

この実験では通過や発車が重なった場合は除外しているため，待機からの変化が多く，待機以外からのショットからショットへの変化は少なかった．

これらのテストデータから一定の時間幅で時系列で前からデータを切り取り，切り取られたデータごとにショットを判別する．

時間幅は 4 秒，6 秒，8 秒である．また，4，6，8 秒で検出を行うが, 時間幅のシフトは時間幅の半分とし，オーバーラップをさせてショット検出を行う方法でも実験を行った．

4.2　ショット検出実験結果

実験結果は図 12 のようにして求めた．例は「待機→発車 F →待機」とショットが変わるデータで，時間幅 4 秒でオーバーラップなしで検出した結果である．横軸が時間で縦軸がその時間から時間幅で指定した区間で検出されたショットが示してある．実線が正しくショット検出されている区間で，破線が誤ってショット検出している区間である．また，矢印とともにグラフ中に書き込んであるショットはその区間で検出され

るべきショットである．

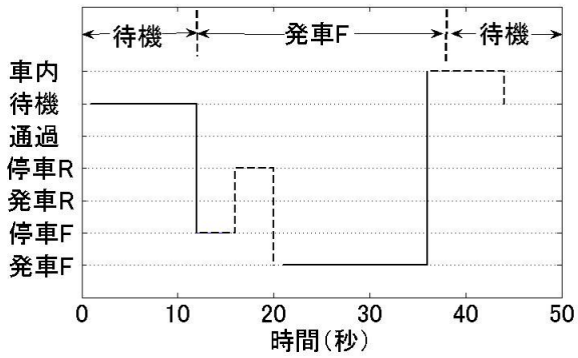

図 12　実験結果の例

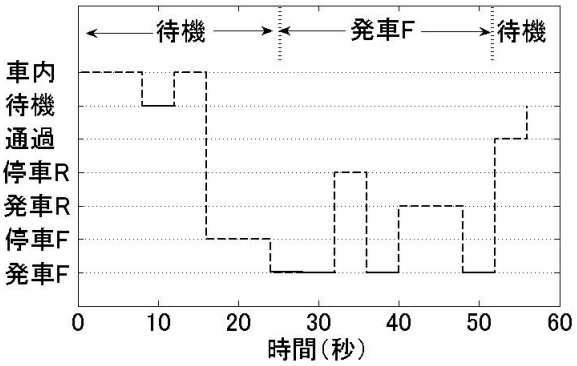

図 13　待機→車内→待機の結果 (4 秒)

4.3　ショット変化への適用

ショット変化を適切に検出できるかということを確かめるために，変化を検出する実験を行ったが，時間幅やオーバーラップの有無に関わらず，ショット変化の際に不要なショットの検出が見られた．

確率モデルは各ショットの全体から推定されている．一方，この実験のように時系列で前からショットを検出していく方法では，ちょうどショット全体が入力されるということはほとんどない．よって，時間幅が 2 つのショットにまたがってしまい，不要なショットの検出を起こしたと考えられる．

また，発着のショットを使うと仮定すると，正しくショット検出ができている例もあった．しかし，構内アナウンスが大きく収録されているデータでは，待機のショットを車内と誤って検出している場合がある．車内のデータにもアナウンスが入っているので，このような誤検出が起きたと考えられる．

対策としては，待機時のアナウンスを確率モデルにして検出するという方法が考えられるが，電車の発着，通過と重なってアナウンスが収録されて場合や，車内アナウンスがどのように検出されるかがわからないので，実験をして検証しなければならない．

4.4　時間幅，オーバーラップによる違い

まず時間幅による違いだが，狭いと不要なショットの検出が増える (図 13，14)．これは，時間幅が狭まりショットの検出回数が増えることで，上記したような 2 つのショットにまたがってしまう可能性も上がることが原因だと思われる．一方，時間幅が広い場合は検出される回数自体が少ないので，不要なショットが減ると思われる．

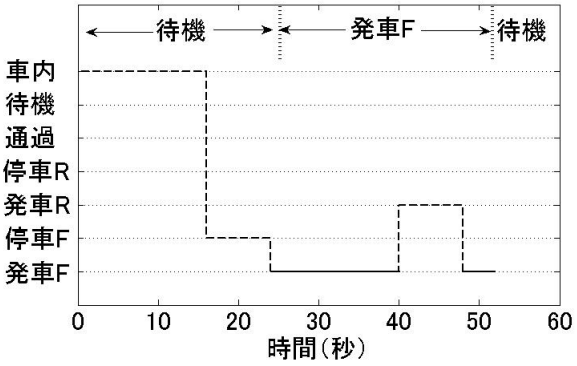

図 14　待機→車内→待機の結果 (8 秒)

また，オーバーラップを行うことでも，上記のように狭い範囲での検出と同じように，検出の回数が増えるために，不要なショットの検出が増えた．

以上のことからこの実験のような方法では，不要なショットの出現は避けられないのでショット検出には不十分であり，改善の必要性がある．改善の 1 つ方法としては，映像情報などと組み合わせて使うということが挙げられる．

例えば，検出実験の結果，多くの場合で正解のショットが 1 度は検出されたので，この実験のように全ての音データからショットを検出するのではなく，映像の色相変化などを利用し，ショット検出をする音情報を限定する方法が考えられる．そのためには，映像に色相の変化があった場合に，その前後何秒くらいの音情報からショット検出を行うと正解が出やすいのかを調べ，そこからショット検出を行う必要がある．

5　あとがき

本論文では，ライフログ映像の電車待ちのシーン検出を行うためのショット識別の手法を 2 つ提案した．一つはパターン間距離を用いた手法で，もう一つは確率モデルを用いた手法である．

パターン間距離ではパワースペクトル平均とパワースペクトル包絡を使う 2 つの手法を実験した．

これらの手法を用いてショット識別実験を行った結果，平均識別率はパターン間距離 (パワースペクトル平均) が 37%，パターン間距離 (パワースペクトル包絡) が 27%，確率モデルが 71% であった．

さらに確率モデルによる手法を用いてショット検出実験を行ったが，ショット変化のときに不要なショットの検出が見られた．

以上のように本論文ではパターン間距離と確率モデルを使った手法について述べたが，発車や停車，待機は短時間スペクトルの時間変化を用いることでも識別可能かもしれない．

また，ショット検出実験では評価を図の読み取りによって行ったが，今後はより客観的な評価方法を考える必要がある．

そして，本論文では電車発着，通過が重なったときや向かいのホームの電車の発着についても考慮していないが，それらについても考慮しなければならない．

さらに今後は，映像のシーンを区切るために，本論文での手法をどのように映像のシーン検出に適用していくかも検討する必要がある．

6　謝辞

本論文の研究，執筆にあたって，貴重なご意見を下さった法政大学情報科学研究科高田勝裕氏に深く感謝致します．

参考文献

[1] 志村将吾, 平野靖, 梶田将司, 間瀬健二, “ 体験記録における日記を用いた感情記録インタフェース, ” 情処研報. ヒューマンインタフェース研究会報告, Vol.2005, no.95, pp. 61-68, 2005.

[2] 相澤清晴, 石島健一郎, 椎名誠, “ ウェアラブル映像の構造化と要約： 個人の主観を考慮した要約生成の試み, ” 信学会論文誌. D-II, vol.J86-D-II, no.6, pp.807-815, 2003.

[3] 石上陽一, 飯島俊匡, 川嶋稔夫, 青木由直, “ 日常生活空間における視点映像の階層的セグメンテーション, ”信学技報, vol.99, no.448, pp.165-172, November 1999.

[4] 久保田敏司, 中村裕一, 大田友一, “ 個人行動記録システムにおける注目シーンの検出： 注目シーン検出の高精度化と環境カメラの利用 ” , 信学技報. PRMU, vol.102, no.218, pp. 47-52, 2002.

[5] 相澤清晴, “ ウエアラブルとユビキタスによるライフログデータの取得と処理, ” , 信学講論_情報. システム, vol.2005 年, no.2, pp. ”SS-6”-”SS-7”, 2005.

[6] 志村 将吾, 平野 靖, 梶田 将司, 間瀬 健二, “ 行動状況により検索可能な体験映像提示手法の検討 ” , 情処講論 68 回, pp. 4_81-4_82, 2006.

[7] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹夫, “ MFCC パラメータ, ” , 音声認識システム, 情報処理学会　編集, pp.13-14, オーム社出版局，東京, 2001.