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

Life-log Video Shot Discrimination using Binaural Microphone

山野貴一郎 † 伊藤克亘 ‡
† 法政大学大学院情報科学研究科
‡ 法政大学情報科学部

Kiichiro YAMANO† Katunobu ITOU‡
†Graduate School of Computer and Information Sciences, Hosei University
‡Faculty of Computer and Information Sciences, Hosei University .

アブストラクト ライフログ映像を効率よく扱うためには映像へのインデキシングのためのシーンの検出が必要である．シーンは映像情報によって検出されることが多いが，映像情報のみでは不十分な場合がある．本論文ではそのようなシーンである駅構内における電車待ちシーン検出に必要なショット識別をスペクトル包絡，Δパワー，音源の移動などの特徴量のモデル化により行った．平均識別率はスペクトル包絡を用いた手法では 67.8%(7 ショット，フィルタ次数 39)，94.8%(3 ショット，フィルタ次数 12)，スペクトル包絡に Δ パワーを加えてモデル化した手法では，73.5%(7 ショット，フィルタ次数 39)，91.7%（3 ショット，フィルタ次数 39），移動音源のモデル化をした手法では，31.7%(7 ショット)，67.2%(3 ショット) であった．

1 まえがき

個人の体験や生活を常時記録し利用するという研究が行われている [1]–[3]．記録された個人の生活や体験の記録をライフログという．ライフログは映像，音声，位置情報，文書など様々な形式で記録され，備忘録，日記，防犯などへの利用が期待されている．映像は最も出来事の再現性が高いが，常時記録された映像はデータ量が膨大で冗長である．このような映像を効率よく閲覧，使用するためには，映像を検索，整理するためのインデキシングが必要である．そのためにはインデクスとなるシーンを検出しなければならない．

映像のシーン検出は色相情報に着目し行われることが多い．例えば文献 [4] では放送用スポーツ映像にアノテーションを付けるためのシーン分割を，色相ヒストグラムを用いたブロックマッチング法で行っている．色相情報の利用はテレビ番組のような編集された映像には有効であるが，ライフログ映像では不規則にカメラの前を人や物が横切る場合があるので，色相情報だけでは不必要なシーンを検出する場合がある．例えば駅構内において映像情報のみを用いてシーン検出を行った場合，図 1 の上図のように電車の停車などがあるたびに，映像情報が大きく変化して別のシーンとして検出されてしまう．この場合は図 1 の下図のように電車の発着などはショットとして識別をし，電車待ちを 1 つのシーンとして検出するのが望ましい．ショットとはシーンの構成要素であり，ショットが集まることでシーンが構成される．

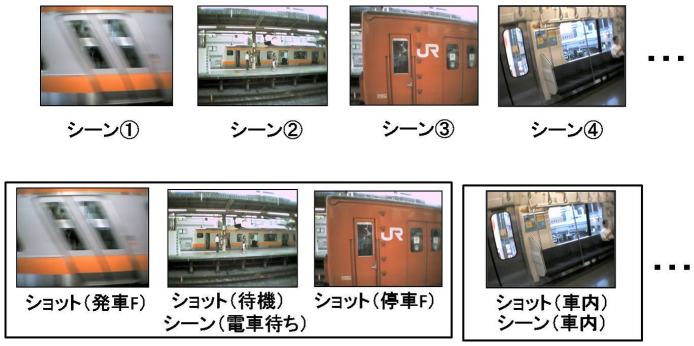

図 1. シーンとショットの例．上段は色相変化でシーンを検出した場合で不必要なシーンが検出されている．下段は正しいシーン検出の例 (電車待ち，車内がそれぞれ 1 つのシーンとなっている．)

このようなシーンは色相情報だけでは検出が困難である．しかし，音響情報を併用することでシーンを正しく検出できる可能性がある．音響情報が有効な例として本論文では駅ホームでの電車待ちシーンを正しく検出するためのショット識別について述べる．識別はバイノーラルマイクで収録したデータを用いてショットをモデル化することで行った．音響情報のモデルは，スペクトル包絡を用いたもの，スペクトル包絡と Δ パワーを用いたもの，音源移動を用いたものの 3 つのモデルを提案する．

2 音響情報を用いたライフログ映像インデキシング

ライフログ映像の音データには，様々な環境音 (背景雑音) や音声が収録されている．これらの音データには様々な情報が含まれており，特に音声からはその時に話したり聞いたりしたことの内容だけでなく，その時の感情や

態度も再現できるので，ライフログとして非常に重要なデータであると考えられる．また，環境音からは少なくとも騒がしい場所，静かな場所など収録した時の周辺の雰囲気がわかり，場合によっては画像データには映っていない，カメラの背後からの情報が得られる．

ライフログ映像の音データはユーザの行動により変化をする．例えば，部屋を移動したり，屋内から屋外への移動をした場合には，背景雑音が変化することが多い．この背景雑音のスペクトルの変化や音量の変化はショットやシーンの識別に役立つと考えられる．しかし，多くの種類の音が収録される環境では同じ環境音や背景雑音が連続して収録されるとは限らない．このような場合はそのシーンで収録される可能性のある音をショットとして識別すべきと思われる．

音響情報を用いた映像インデキシングには文献 [5] のようなテレビ番組を対象とした研究がある．この文献では音楽をパワースペクトルのピークが時間的に安定していることを利用して検出をし，音声を音楽成分を除いた信号から comb フィルタで検出している．そして，検出した音楽，音声をインデックスとして映像に付加している．また，文献 [6] ではサッカー番組にリアルタイムでダイジェスト用のメタデータを付加するために，会場音の短時間パワーでシーンを検出し，アナウンサーの音声から付加するメタデータの検出を行うという手法が提案されている．

また，音源方向や移動音源の推定に関する研究の多くではマイクロホンアレーが用いられ，各マイクロホンなどに生じる位相差などを用いて音源の方向推定などを行う手法が提案されている．例えば，文献 [7] ではマイクロホンアレーの各マイクロホンの相互相関値をもとにクラクションの方向の推定を行っている．文献 [8] ではマイクロホンアレーを用いて，クロススペクトル法や MUSIC 法で等速移動をする移動音源の方向推定を行っている．

一方でバイノーラル信号を用いた音源方向推定の手法も提案されている．例えば，文献 [9] ではダミーヘッドで収録したデータから得られた両時間音圧差の包絡に着目し，各方位の統計モデルを作成して音源の方向推定を行っている．また，文献 [10] では両耳への音の到着時間差によって水平角の音源方向を，頭部伝達関数により垂直角の音源方向を推定し 3 次元の方向推定を行っている．

本研究のように常時記録する目的で使用する場合には，装着の困難さからマイクロホンアレーの使用は実用的ではない．また，事前に頭部伝達関数などの測定もできない．しかし，本研究では上記研究で行われているような正確な音源方向の推定は必要ではなく，おおよその音源の方向がわかればよいので，バイノーラルマイクで収録した両耳の信号の相互相関を用いることで音源の方向推定を行った．

3 ショットのモデル化

3.1 ショットの分類

本論文では電車待ちシーンをホームで電車を待ち始めてから乗車までとする．また，実際の駅ホームで録音したデータを聴取し，音が類似している状況を 1 つのショットとして，電車待ちシーンを次の 6 ショットに分類した．

- ホーム前方での発車時（以下，発車 F）
- ホーム前方での停車時（停車 F）
- ホーム後方での発車時（発車 R）
- ホーム後方での停車時（停車 R）
- 電車通過時（通過）
- 通過などがない時のホームでの待機中（待機）

また，電車内は 1 つのシーンと考えられるが，乗車時のシーン変化を検出するために上記の 6 ショットに加えて識別を行う (以下，車内)．しかし，電車待ちを 1 つのシーンとするには発車，停車，通過に関する 5 つのショットを細かく識別する必要はなく，以下のように 3 つのショットとして識別するだけでも十分である．

- 電車の発着，通過（以下，電車）
- 待機
- 車内

本論文でのモデル化および実験は，ショットを 7 つに分類した場合と 3 つに分類した場合の両方で行った．

3.2 スペクトル包絡を用いたモデル化

各ショットの短時間スペクトルから特徴を抽出しモデル化を行った．短時間スペクトルは 2048 点のフレームを 1024 点ずつシフトさせながら切り出し，各フレームにハニング窓をかけて 2048 点 FFT をして求めた．さらにこの短時間スペクトルから特徴を抽出するためにフィルタバンク分析を行う．フィルタバンク分析はメル周波数軸上で一定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ることで行った．三角窓のシフトは窓長の半分をオーバーラップさせるように行う．これにより短時間スペクトルの特徴が十数点から数十点に集約されたスペクトル包絡が得られる．

上記の処理によって求められた短時間スペクトルのフィルタバンク出力は，帯域毎に対数正規分布をするので，フィルタバンク出力の対数をとることで正規分布として確率密度関数を推定する．この対数をとったフィルタバンク出力から平均スペクトル包絡 (図 2)，共分散を求め確率モデルとする．

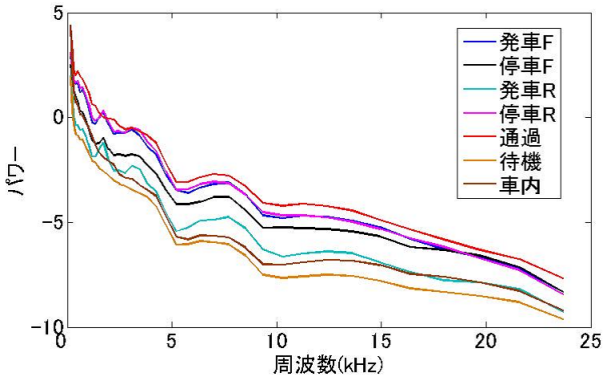

図 2. ショット別平均スペクトル包絡 (39 次)

本論文では，フィルタバンク分析はメル周波数軸上で 200，300，400 などの複数の帯域幅で行った．異なる帯域幅を使うことでフィルタ次数 (スペクトル包絡の点数) は表 1 のようになる．したがって，複数のモデルができるので，これらを実験により比較した．また，次数を低くすることで図 3 のようにスペクトル包絡の特徴が少ない点数に集約される．

表 1. 帯域幅とフィルタ次数

帯域幅	200	300	400	500	600	700
フィルタ次数	39	25	19	15	12	10

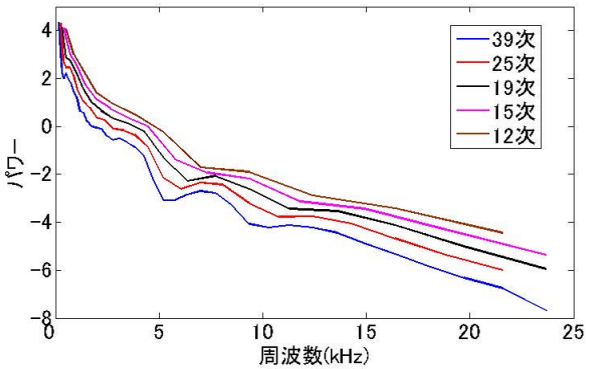

図 3. フィルタ次数別平均スペクトル包絡の例 (「通過」)

ショットの識別は学習データから求めた平均スペクトル包絡と共分散を用いて，式 (1) によって尤度を求めることで行った．SF_i は学習データから求めた各ショットの平均スペクトル包絡と共分散で，x は入力ショットの平均スペクトル包絡である．

$$\text{shot} = \underset{i}{\text{argmax}}\,[p(x|SF_i)] \tag{1}$$

3.3 スペクトルを用いたショット識別実験

3.3.1 データ収録

学習，実験に用いたデータは 2 つの駅とその間の電車内で収録をした．収録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．収録機材はバイノーラルマイク (adphox BME-200) と PCM 録音機 (EDIROL R-09) である．収録方法は，バイノーラルマイクを両耳に装着し，駅のホーム前方で 15 分程度収録をした後，電車に乗り込み車内の音を収録する．そして，次の駅で降り，ホーム後方で 15 分程度収録をする．その後再び電車に乗り車内で収録をして最初の駅へ戻る．同様にして最初の駅のホーム後方，次の駅のホーム前方で収録を行う．収録の時間帯は 10 時〜16 時であるが，多くは 11 時〜13 時に収録されたものである．以上のようにして収録したデータから人手によりショットを切り出し，学習データとテストデータとした．学習，テストデータ数とその平均時間を表 2，表 3 に示す．

表 2. 学習データ (中段：データ数，下段：平均時間 (秒))

発車 F	停車 F	発車 R	停車 R	通過	待機	車内
21	19	25	30	24	74	20
25	13	13	24	17	53	130

表 3. テストデータ (中段：データ数，下段：平均時間 (秒))

発車 F	停車 F	発車 R	停車 R	通過	待機	車内
16	11	10	13	10	36	8
22	11	12	24	16	42	123

3.3.2 ショット識別実験

スペクトル包絡を用いた確率モデルで実験を行った．実験はテストデータを入力しそのショットを識別させた．フィルタ次数は 7 ショットの場合が 39 次，25 次，19 次，15 次，12 次で 3 ショット場合はそれらに加えて 10 次も試した．結果は入力データ中の正解の割合 (識別率) で評価した (表 4，5)．

表 4. フィルタ次数別の識別率 (7 ショット)

	識別率 (%)				
フィルタ次数	39	25	19	15	12
発車 F	32.6	18.8	12.5	12.5	6.3
停車 F	0	0	0	0	0
発車 R	80.0	80.0	80.0	90.0	80.0
停車 R	100	100	92.3	92.3	92.3
通過	90.0	90.0	90.0	80.0	80.0
待機	72.2	80.6	80.6	80.6	88.9
車内	100	100	100	100	100
平均	67.8	67.1	65.1	65.1	63.9

表 5. フィルタ次数別の識別率 (3 ショット)

	識別率 (%)					
フィルタ次数	39	25	19	15	12	10
電車	78.3	85.0	86.7	85.0	90.0	83.3
待機	75.0	86.1	86.1	86.1	94.4	100
車内	100	100	100	100	100	100
平均	84.4	90.4	90.9	90.4	94.8	94.4

3.3.3 考察

7 ショットで識別した場合は，発車 F，停車 F が低い識別率であった．識別の誤りの傾向としては，発車 F は停車 R として，停車 F は発車 R や待機として識別されることが多かった．発車 F は「電車が徐々に加速し走行して去っていく」ショットであり，停車 R は「電車が走行して徐々に減速して停止する」ショットである．それゆえに音響情報の時間変化を考慮していないこの手法では，区別できず誤って識別されたと思われる．また，停車 F と発車 R の対にも同様のことが言える．待機として誤って識別されたテストデータは，電車からの音が小さく聴取しても停車したことの判別が難しいデータであった．3 ショットの場合は 7 ショットの場合と比較して高い識別率が出ているが，電車のショットが全体的にやや低い．原因は停車 F や発車 R が待機と識別されたためである．

また，7 ショットの場合は 39 次，3 ショットの場合は 12 次のときに最も良い識別率が得られた．これは電車の発着や通過のようなある程度似ている音を識別するには，細かい特徴までモデル化した方が有効であることと，発着，通過を 1 つのショットとした場合は，大まかな特徴をとらえてモデル化した方が有効であることを示していると考えられる．

3.4 平均スペクトル包絡と Δ パワーを用いたモデル化

スペクトル包絡を用いてショットのモデル化を行った場合，時間変化は異なるが平均スペクトル包絡が似ているショット (発車 F と停車 R や発車 R と停車 F) の識別率が低かった．そこで時間変化に関する特徴量として，音量の時間変化 (Δ パワー) とスペクトル包絡を合わせてモデル化を行った．音量は時間波形で 2048 点のフレームを 1024 点ずつシフトさせながら切り出し，三角窓をかけて振幅の 2 乗和を求めることで算出した．Δ パワーは得られた音量データから式 (2) で求めた．$vol(n)$ は n フレーム目の音量である．電車が加速する場合は徐々に音量が上がるため Δ パワーは正の値になる．反対に減速する場合は負の値になる．また，待機や車内のように音量に大きな変化がないショットでは，0 に近い値をとると考えられる．したがって，停車と発車の識別，これらと待機の識別に有効であると考えた．

$$\Delta\text{Power} = vol(n+1) - vol(n) \tag{2}$$

この Δ パワーの平均値を求め，1 次の特徴量としてスペクトルの特徴量と合わせて，平均，共分散を求めてモデル化をした．ショット識別はスペクトル包絡の時と同様に尤度を用いて行った．

3.4.1 ショット識別実験

Δ パワーとスペクトルを結合したモデルを用いてショット識別実験を行った．フィルタ次数は 39 次，25 次，19 次，15 次，12 次とした．結果を表 6 に示す．

表 6. フィルタ次数別の識別率 (7 ショット)

	識別率 (%)				
フィルタ次数	39	25	19	15	12
発車 F	37.5	31.3	12.5	25.0	25.0
停車 F	0	0	0	0	0
発車 R	90.0	90.0	90.0	90.0	90.0
停車 R	100	92.3	92.3	92.3	92.3
通過	90.0	90.0	60.0	40.0	20.0
待機	97.2	94.4	94.4	94.4	91.7
車内	100	75.0	37.5	25.0	25.0
平均	73.5	67.6	60.6	57.0	49.1

表 7. フィルタ次数別の識別率 (3 ショット)

	識別率 (%)				
フィルタ次数	39	25	19	15	12
電車	75.0	76.7	71.7	66.7	65.0
待機	100	100	100	100	100
車内	100	75.0	37.5	25.0	25.0
平均	91.7	83.9	69.7	63.9	63.3

3.4.2 考察

フィルタ次数が 39 次の場合では識別率の向上が見られた．しかし，帯域幅が広くなるほど識別率が低下した．帯域幅が広くなると確率モデルの次数が低くなる．この低い次数のモデルに Δ パワーを加えたことで影響が大きく出たために，識別率が低下したと思われる．特に車内は音量の変化が小さく，同じく音量の変化の小さい待機として誤って識別されたため低い識別率となっている．また，停車 F は Δ パワーを利用しても，識別できなかった．識別の誤りとしてはほとんどが発車 R であった．結局，停車 F も停車 R も音量が，最初はホームに入ってくる，もしくは加速するので徐々に大きくなり，その後，減速する，またはホームから離れていくとなるので徐々に小さくなる．つまり，平均することでこれも特徴が似てしまったと思われる．これらの問題を解決するためには，Δ パワーを全体の平均とするのではなく，電車が「ホームに近づいてくるとき」,「ホームから去っていく時」,「加速時」,

「減速時」などの Δ パワーの符号が一致する区間で細かくモデル化を行うのが有効である可能性がある．しかし，それぞれの区間をどのように割り出すのかが問題である．

3.5 音源の移動を用いたモデル化

電車が発着，通過するときには電車が動いているため音源の移動が観測される．これをモデル化することによりショット識別を行う．音源の移動の導出は音源の方向を短時間のフレームで連続して求める事で行う．音源の方向は両耳に装着したバイノーラルマイクへの音の遅延により求める．例えば，両耳のバイノーラルマイクの距離が 0.3m でホームで正面を向いて立っているとし，電車の時速を 50km/h として，電車が矢印の方向に 30m 右から 30m 左に通過したとすると音の遅延は図 4 のようになる．ただし方向は正の値が右，負の値が左で値が 0 に近くなるほど音源が正面，もしくは背面にあることを示す．

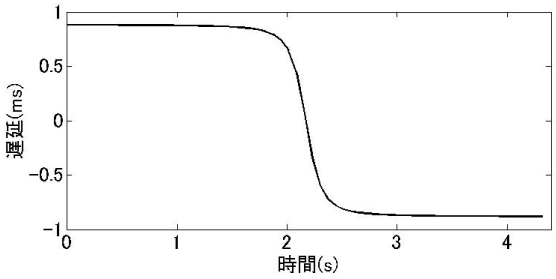

図 4. 通過の理想的な遅延

図 4 から特徴を抽出するために，1 次微係数を求める (図 5)．1 次微係数の傾きに着目してみると，正面を通過したのを境に傾きが負から正になっている．この傾きを利用してモデル化を行うが，移動が逆になると 1 次微係数は時間軸に対称となり傾きの正負が逆になる．向きの違いを無視するため，ショット全体を前半と後半に分け，それぞれの平均値を求め，時間順序は無視し，大小にだけ着目した 2 次元の特徴量とした．モデル化は 1 次微係数の大小それぞれの平均と共分散を求めて行った．

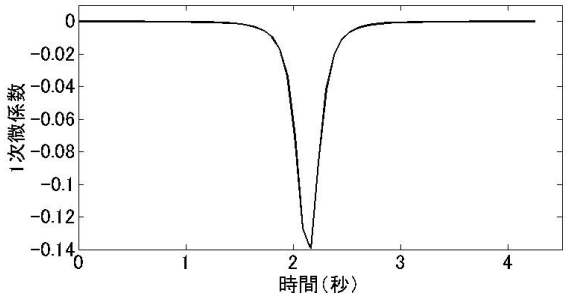

図 5. 「通過」1 次微係数の理想曲線

以上のように特徴量を求めるが，実際のデータでは電車の移動の様子は未知なので，式 (3) で両耳の信号の相互相関の最大値を求めて，そのときの n を音源の方向とする．L は左チャネル，R は右チャネルの信号であり，n は -100 から 100 まで計算する．この処理を 5000 点のフレームを 2500 点ずつずらしながら行い，音源の移動を推定する．図 6 に実際に通過のショットに対して移動を求めた例を示す．横軸はフレーム番号で，縦軸はそのフレームの時の最大相関値になった時のサンプルシフト数である．つまりこれが音源の方向で，負の時は左，正の時は右，0 は正面 (背面) である．このサンプルシフト数の 1 次微係数を求めるが，より移動の様子を明確にするため，微分の前に 20 点移動平均フィルタによって平滑化を行った．この平滑化を行ったデータに対し微分を行い，1 次微係数を求め，再度平滑化をして (図 7)，傾きを算出する．また，図 8,9 に待機に対し同じ処理をした例を示す．このように異なるショット間では類似していないためショット識別に有効であると考えられる．ショット識別は他のモデルと同様に，平均，共分散を用いた 2 次元確率モデルから尤度を求めることで行った．

$$\mathrm{direction} = \underset{n}{\mathrm{argmax}} \left[\frac{\sum_{i=1}^{4800} L_i^2 R_{i+n}^2}{\sqrt{\sum_{i=1}^{4800} L_i^2} \sqrt{\sum_{i=1}^{4800} R_{i+n}^2}} \right] \quad (3)$$

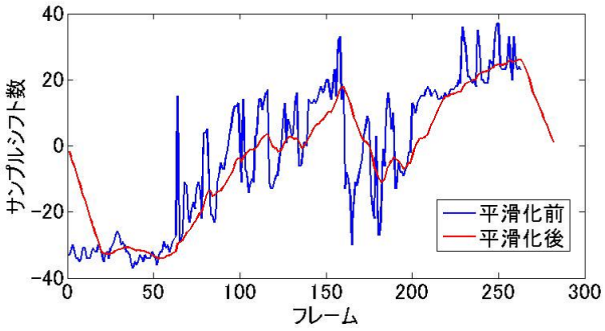

図 6. 「通過」の移動推定

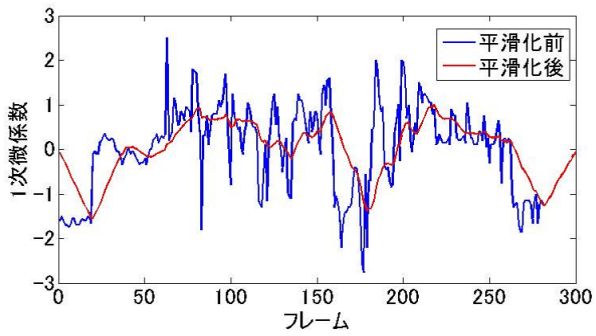

図 7. 「通過」の 1 次微係数

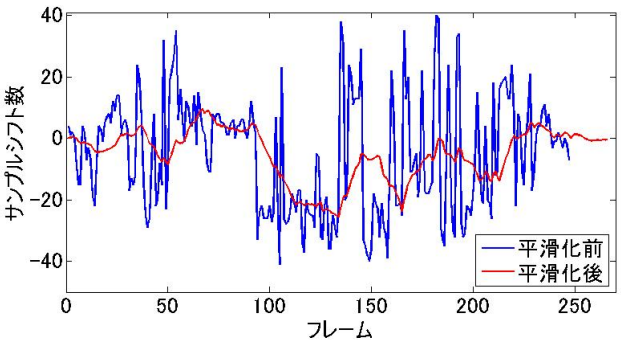

図 8. 「待機」の移動推定

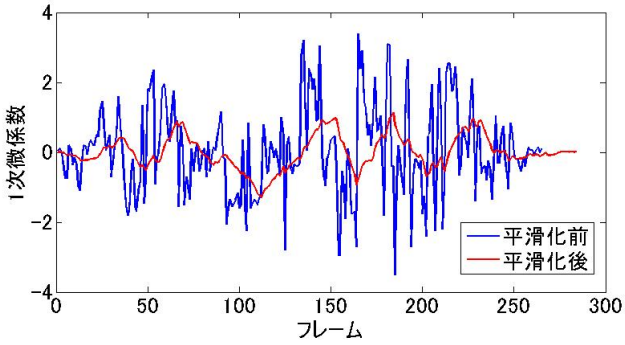

図 9. 「待機」の 1 次微係数

3.5.1 ショット識別実験

実験は他のモデルと同様にショットを入力し，識別をさせその識別率で評価を行った．テストデータ数はこれまでと同じである．実験は 7 ショットと 3 ショットの両方で行った．結果を表 8 に示す．

表 8. ショット別識別率

ショット	識別率 (%)
発車 F	12.5
停車 F	9.1
発車 R	30.0
停車 R	30.8
通過	10.0
待機	41.7
車内	87.5
平均	31.7

ショット	識別率 (%)
電車	61.2
待機	52.8
車内	87.5
平均	67.2

3.5.2 考察

7 ショット，3 ショットのどちらの場合も他の手法に比べ低い識別率となった．駅構内での音には電車以外にもアナウンスや周辺の雑音がありこれらが影響して正しい方向推定を行えなかったためと思われる．また，データ収録のときには，常に正面を向いているわけではなく，頭部の動きは特に意識をしていないため，図 4 のような理想的な両耳遅延が得られなかったことも原因と考えられる．これらの点を解決するには，電車の音を追跡したり，受音側の変化を考慮したものにしなければならない．

4 あとがき

ライフログ映像のシーン検出を行うための駅構内でのショット識別の手法を提案した．提案手法はスペクトル包絡，Δ パワー，移動音源のモデル化を用いた手法である．それぞれの手法で確率モデルを求めショット識別実験を行った結果，平均識別率はスペクトル包絡を用いた手法では 67.8%(7 ショット，フィルタ次数 39)，94.8%(3 ショット，フィルタ次数 12)，スペクトル包絡に Δ パワーを加えてモデル化した手法では，73.5%(7 ショット，フィルタ次数 39)，91.7%（3 ショット，フィルタ次数 39），移動音源のモデル化をした手法では，31.7%(7 ショット)，67.2%(3 ショット) であった．今後は識別率の向上とともに識別手法をどのようにシーン検出に利用していくかを考えなければならない．

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，法政大学大学院情報科学研究科の高田勝裕氏に深く感謝いたします．

参考文献

[1] M. Lamming, et. al., ““ Forget-me-not ”Intimate Computing in Support of Human Memory”, *Proceedings of FRIEND21*,1994

[2] J. Gemmell, et. al.,“MyLifeBits: Fulfilling the Memex Vision”, *Proceedings of the tenth ACM Multimedia*, pp.235-238, 2002

[3] K. Aizawa, et. al., “Capture and Efficient Retrieval of Life Log”, *Proceedings of the pervasive 2004 workshop on memory and sharing experience*, pp.15-20, 2004

[4] 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

[5] 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

[6] 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

[7] 島田他，“マイクロホンアレーによるクラクションの方向定位” **信学技報**. *EA*, Vol. 105, No. 54, pp.25-33, May., 2005

[8] 土屋他，“リニアアレーを用いた移動音源の方向推定” **信学技報**. *EA*, Vol. 100, No. 724, pp.21-28, Mar., 2001

[9] 井上他，“統計モデルを用いた音源方向推定” **信学技報**. *EA*, Vol. 105, No. 651, pp.23-28, Mar., 2006

[10] 堀畑他，“両耳聴モデルを用いた 3 次元音源方向定位システムの開発” **機学論** *C*, Vol. 72, No. 723, pp.3567-3575, Nov., 2006