音響シーンクラスタリングによる番組効率視聴支援

広畑 誠†　井本 和範†　青木 恒†　上原 龍也†

† 株式会社 東芝 研究開発センター

1 はじめに

近年のハードディスクレコーダーや PC には，放送番組を録画する機能だけでなく，視聴，検索，編集の操作性を向上させる新たな機能が搭載されつつある．映像構造化技術は，新たな機能実現には不可欠な技術の一つである．[1]では，画像特徴をもとに，得られた類似ショットを用いたコーナー検出技術を提案しており，時間分割による構造化を可能にした．しかし，番組の種類によっては，必ずしも精度良く処理できるわけではない．そこで，本稿では，音響処理という異なるアプローチからコーナー検出を行う方法について報告し，その有効性について検討する．

2 画像特徴を用いたコーナー検出

2.1. 類似ショット検出

処理対象データとしては MPEG-2 形式のファイルを用いる．この方法ではまず，縮小画像を取り出し，色相ヒストグラムと輝度成分のモザイク画像を用いて類似検定を行い，カット検出，類似ショット検出を行う[1]．この処理で，映像はショットと呼ばれる細かい区間に分割され，類似したショット同士に同じ ID を付与することが可能となる．

2.2. コーナー検出

コーナー検出処理には，まず m 番目のショットから n 番目のショットまでの任意の区間[m, n]にて，ショット同士の絡み具合を表す対話度数 I_{mn} を求める（式(1)）．

$$I_{mn} = \frac{\sum_{i=m}^{n} \rho_{mn,i} \cdot \sum_{i=m}^{n} \rho_{mn,i} \lambda_i}{\left(\sum_{i=m}^{n} \lambda_i \right)^2} \qquad (1)$$

ここで，λ_i はショットの時間長であり，$\rho_{mn,i}$ はショット i と同じ ID を持つショットが区間[m, n]で出現すれば 1，しなければ 0 をとる値である．そして，映像の先頭から対話度数が極大になるようにショット区間の始点終端の組を探索し，極大値が閾値を越える場合に対話区間として定義していく．最後に，区間の分割点となった境界時刻をコーナーの境界時刻として検出する（図 1）．

2.3. 類似ショットの問題点

類似ショットは同一のカメラアングルで撮影

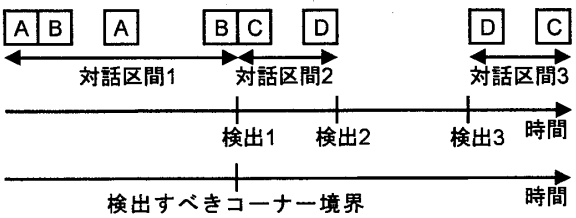

図 1. 対話区間検出とコーナー境界検出の例

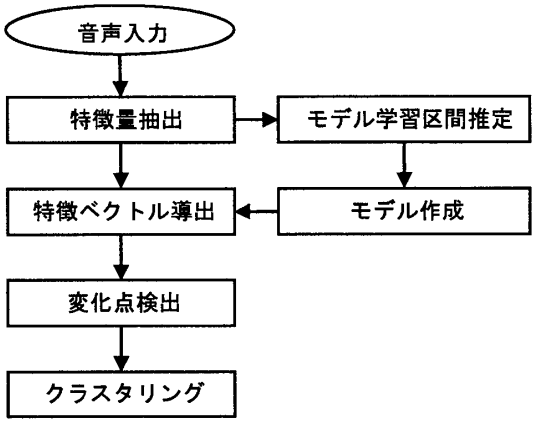

図 2. 音響処理でのフローチャート

されたショットをグルーピングしたものである．そのため，様々なアングルから撮影される番組において，類似ショットの番組全体に対する割合は少ない．それ故，各対話区間は必要以上に短くなり，コーナー境界を過剰に検出してしまう（図 1）．このとき，コーナー毎に発話者の構成が異なる番組ならば，同じ発話者の音声は類似した特徴を持つものとして観測することが可能である．したがって，音響的に類似したシーン毎に分類（以下，音響シーンクラスタリングと表記）すれば，発話者毎に分類することになり，コーナーの境界検出および継続するか否かの判定が容易にできる．次章では音響特徴を用いた手法について述べる．

3 音響特徴を用いたコーナー検出

3.1. 音響シーンクラスタリング

画像特徴の場合とは異なり，MFCC など音声データから抽出した特徴を用いて類似検定する方法では，グルーピングが難しい．そこで，[2]の方式による処理とクラスタリング処理を組み合わせる（図 2）．まず，入力音声から特徴量として MFCC を抽出した後，M 個のモデルを作成する．同時に，一定時間毎に分割し，M 次元の

Audio Scene Clustering for Indexing of TV Shows
Makoto Hirohata†, Kazunori Imoto†, Hisashi Aoki†, Tatsuya Uehara†
† Corporate R&D Center, Toshiba Corporation

表 1. 評価結果

種類	時間 (min)	境界数	再現率		適合率	
			画像	音響	画像	音響
コメディ	107	16	81.3	93.8	21.0	78.9
音楽	90	31	48.4	87.1	44.1	87.1
料理	59	15	100.0	100.0	40.5	71.4
ドキュメンタリー	120	20	95.0	80.0	38.8	61.5
カルチャー	90	23	82.6	87.0	26.8	74.1

特徴ベクトルを作成する．特徴ベクトルの m 次元目の要素は m 番目に作成したモデルに対する尤度とする．次に，時間軸上で隣接している特徴ベクトル同士の類似度を求めていき，極小になった時刻で分割する．この分割により区切られたそれぞれの区間毎に特徴ベクトルを求めた後，クラスタリング処理を行う．

クラスタリング処理には，MeanShift 法[3]を用いる．具体的には，まず，各特徴ベクトルを自身とユークリッド距離が閾値以下の全ての特徴ベクトルとの平均ベクトルに更新する．そして，更新後のベクトルが収束するまでこの処理を繰り返す．最後に，互いの特徴ベクトルのユークリッド距離が閾値以下となる区間を集めた音響クラスを作成する．

3.2. コーナー検出

音響特徴を用いたことで，同一コーナーの類似シーンがより長い時間検出できる．そこで，[4]のように，ある時刻を境として，一定時間からなる隣接する前後 2 つの時間区間の類似性に着目し，コーナー検出を行う．時刻 t 前後の時間区間は次元数がクラス数 K となるベクトル $\boldsymbol{a}_t$, $\boldsymbol{b}_t$ で表現する．ベクトルの k 番目の要素は k 番目のクラスに属する区間が出現した総時間の平方根とする．時刻 t 前後の時間区間の非類似度 D_t は式(2)で求める．

$$D_t = \frac{1}{K}\left\|\boldsymbol{a}_t - \boldsymbol{b}_t\right\|^2 \qquad (2)$$

この非類似度を一定時刻毎に求め，極大値をとる時刻をコーナー境界時刻として検出する．

4 評価実験

画像特徴を用いたコーナー検出（以下**画像**と表記）と，音響特徴を用いたコーナー検出（以下**音響**と表記）の検出性能を評価するため，実験を行った．実験は，予め用意したコーナー境界時刻に対し，自動で検出した時刻の検出精度を求めるもので，再現率，適合率を算出した．なお，コーナー境界時刻に対し±30sec 以内に検出できれば正解とした．実験には，5 種類の番組を対象として，それぞれ放送 2 回分，計 10 個のコンテンツを用いた．

コーナーの境界時刻は，各番組が持つ構成に従い，構成が切り替わる時刻とした．例えば，コメディ番組は，コーナー毎にタレントが変わる構成であり，音楽番組は，スタジオシーンの間に歌のシーンが挿入される構成である．また，料理，ドキュメンタリー，カルチャー番組は，スタジオと説明ビデオが繰り返される．

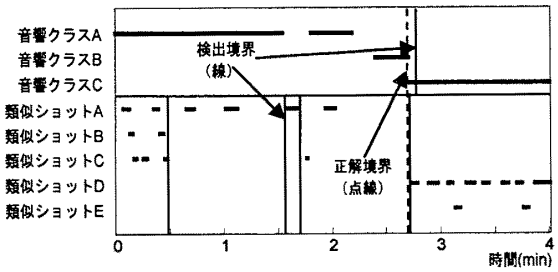

図 3. 実験結果の例

実験結果を表 1 に示す．結果より，再現率に関しては，**音響**は**画像**に比べ，同等以上の性能が得られ，特に音楽番組では高精度な結果が得られた．音楽番組にて**画像**の精度が低かった理由としては，カメラアングルのバリエーションが多く，類似ショットの検出数が極端に少なかったことが挙げられる．また，適合率に関しては，どの種類の番組においても，**画像**より**音響**の方が良い結果を得ている．図 3 は類似ショット検出結果および音響クラスの作成結果の例である．横軸が時刻を表し，各時刻においてどの類似ショットまたは音響クラスが出現したかをプロットしている．この例からも分かるように，類似ショットの出現時間は少なく，さらに，中心となるカメラアングルも変化するため，**画像**では，コーナー境界を過剰検出してしまう．よって，実験した番組においては，画像特徴よりも音響特徴の方が，コーナーの継続性を判断しやすく，高精度な検出が行えると言える．

5 まとめ

本稿では，音響特徴を用いてコーナー検出する手法について報告した．コーナー毎に発話者の構成が異なる番組では，画像特徴を用いるよりも，コーナーの継続性の判定がし易く，精度の高いコーナー境界時刻の検出が行えることが確認できた．

今後の課題としては，他の種類の番組に対する検討や，画像特徴と音響特徴を組み合わせた手法についての検討が挙げられる．

参考文献

[1] 青木, 信学論, vol.J88 D-II, no.1, pp.17-27, 2005.
[2] 広畑 他, 音講論(春), pp.109-110, 2007.
[3] D. Comaniciu, et al, IEEE TPAMI, vol.24, no.5, pp.603-619, 2002.
[4] M.A.Hearst, Computational Linguistics, vol.23, no.1, pp.33-64, 1997.