論文 / 著書情報
Article / Book Information

論題(和文)	野球中継番組を対象とした音響情報を用いたシーン認識
Title(English)	
著者(和文)	宮崎 太郎, 中川 弘充, 中川 竜太, 岩野 公司, 篠田 浩一, 古井 貞熙
Authors(English)	Taro Miyazaki, Hiromitsu Nakagawa, Ryuta Nakagawa, Koji Iwano, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2006年春季講演論文集, Vol. , No. , pp. 19-20
Citation(English)	, Vol. , No. , pp. 19-20
発行日 / Pub. date	2006, 3

野球中継番組を対象とした音響情報を用いたシーン認識*

◎宮崎太郎　中川弘充　中川竜太　岩野公司　篠田浩一　古井貞熙 (東工大)

1　はじめに

現在，スポーツ中継番組，特に野球中継番組を対象とした，検索用メタデータの自動生成についての研究が盛んに行われている．映像だけでは認識の難しいシーンが存在していることから，精度の高いシーン認識を実現するためには，映像情報と音響情報を併せて利用する必要がある [1]．そこで本稿では，映像情報を用いたシーンの認識と組み合わせる前段階として，音響情報のみを用いたシーン認識について検討を行う．

音響情報を用いた手法としては発話内容を用いたシーン認識手法 [2] が提案されているが，スポーツ中継番組の音声は背景雑音や発話スタイル，表現の揺れの問題から認識が難しい．発話内容の認識と会場音の解析の 2 つの手法を組み合わせた手法 [3] では歓声があがっている区間に重要シーンがあるという知見を用いて精度の向上を実現しているが，やはり音声認識の精度が課題として残る．また，アナウンサーの発話の興奮度とバット音を用いる手法 [4] も提案されているが，バット音に関しては認識精度が低い，他のスポーツへの応用が難しいといった問題がある．

上記のような問題から，発話内容を用いず，またスポーツの種類に依存しない手法として，ケプストラム情報，シーン全体の時間長，シーン内部での発話時間長を組み合わせて利用するシーン認識手法を提案する．

2　評価データ

評価実験には NHK 放送技術研究所から提供された野球中継データベースを用いた．これは実際に放映された野球中継番組 7 試合分 (約 25 時間) のデータで構成されている．各試合は，実況アナウンサー 1 名とゲスト解説者 1 名による発話が含まれている．アナウンサーは全試合において同一話者であるが，ゲスト解説者は試合により異なる．

各試合のデータには，手作業により作成されたインデックスが付属している．このインデックスには，シーンの開始時刻，シーンで起こったプレイの種類 (シーンラベル) とそれ以外の注釈が記されている．ここでのシーンとは，投球が始まる直前から次の投球が始まる直前までに相当する．このシーンラベルのうち，今回の実験では「ホームチーム得点」,「ビジターチーム得点」,「長打」,「短打」,「アウト」,「ストライク＋ボール＋ファウル (SBF)」の 6 種類の認識を考える．ラベル中の時刻情報に基づきシーンごとにデータを切り出した上で，モデル学習・シーン認識を行う．なお，データベース中にはリプレイやイニングの切り替え時などの，上記の 6 種類のシーン以外のシーンも含まれているが，今回の実験ではそれらのシーンは事前に取り除いている．

表 1. 各シーンの時間長の平均と標準偏差

シーンラベル	平均時間長 (秒)	標準偏差 (秒)
ホームチーム得点	37.5	14.3
ビジターチーム得点	34.3	11.1
長打	28.4	9.47
単打	24.0	10.2
アウト	20.0	12.3
SBF	20.1	8.74

3　シーン認識手法

3.1　概略

シーン認識においては，まず，6 種類のシーンごとにケプストラムを特徴量とした HMM を学習する．次にその各々に対し，評価データを用いた教師なし適応 (MLLR+MAP) を行い，その適応後の HMM を用いてシーン認識を行う．また，それに加えて，シーン全体の時間長，シーン内部での音響的なイベントの継続時間長をそれぞれモデル化して利用する．

3.2　ケプストラム情報を用いたモデル (モデル A)

歓声や音楽の有無，アナウンサーの発声の興奮度の違いといった音響的な特徴がケプストラム情報に反映されていると考え，この情報を用いてシーン認識を行う．HMM の構造としてはスキップなしの left-to-right 型，ergodic 型の 2 種類の比較検討を行う．

3.3　シーン時間長を用いたモデル (モデル S)

表 1 に，シーンの種類ごとのシーン全体の時間長の平均と標準偏差を示す．得点シーンや長打などの重要なシーンは比較的シーン全体が長時間になり，逆にストライクやボールはシーン全体の時間が短いことがわかる．このような時間長の分布の違いを用いてシーン認識を行う．

シーン時間長情報のモデル化にはガンマ分布を用いた．予備実験において，正規分布によるモデル化との比較の結果，より高い性能が得られることがわかっている．

3.4　音響的イベントの継続時間長を用いたモデル (モデル E)

ケプストラム情報によって学習された HMM の各状態は，それぞれが異なる音響的な特徴を持つイベントに対応している．left-to-right 型であれば投手の投げる前の静寂区間や歓声の上がっている区間などの音響的イベントの順序が考慮されたモデルとなっている．ergodic 型は状態遷移の順序の制限が緩いこ

---

* Scene Recognition for TV Baseball Program Using Acoustic Information.

By Taro Miyazaki, Hiromitsu Nakagawa, Ryuta Nakagawa, Koji Iwano, Koichi Shinoda, Sadaoki Furui (Tokyo Institute of Technology)

表 2. 各手法及びそれらを融合した場合のシーン認識率 (F 値)

HMM の構造	A	S	E	$A+S$	$A+E$	$S+E$	$A+S+E$
Left-to-right	0.44	0.22	0.18	0.51	0.48	0.30	**0.54**
Ergodic	0.43	0.22	0.21	0.49	0.47	0.27	0.49

とから，より細かいイベントの遷移に追従することが可能なモデルとなっている．

HMM の各状態の継続時間長は盛り上がっていた区間の長さや，アナウンサーの発話時間長を表現していると考えられ，その情報を用いることにより性能の向上が期待できる．

シーン時間長のモデル化と同様に，音響的イベントの継続時間長のモデル化にはガンマ分布を用いた．なお，HMM の状態遷移とデータの対応付けには Viterbi algorithm を用いた．

3.5 尤度の融合

上記の手法により求まった尤度を融合することで，全体の尤度を計算する．モデル A から得られる対数尤度を L_A，モデル S から得られた対数尤度を L_S，モデル E で得られた対数尤度を L_E とし，以下の式に従って全体の融合対数尤度 L を求める．

$$L = w_A L_A + w_S L_S + w_E L_E \quad (1)$$
$$(w_A + w_S + w_E = 1, \quad 0 \leq w_A, w_S, w_E \leq 1)$$

4 実験条件

実験では，7 試合のデータのうち，6 試合をモデル学習用データ，1 試合を評価データとした leave-one-out 法により評価を行った．性能評価基準として 6 種類のシーンの F 値の平均を用いた．ケプストラム情報としては，MFCC12 次元，ΔMFCC12 次元，ΔΔMFCC12 次元，Δ 対数パワー 1 次元，ΔΔ 対数パワー 1 次元の計 38 次元のベクトルを用いた．特徴量はフレーム長 25ms，フレーム周期 10ms で抽出し，入力シーンごとに CMS を行った．

ケプストラム情報を用いたシーン認識において使用する各 HMM の状態数と混合数は，実験的にそれぞれ最良の値を用いており，今回の実験ではともに 9 状態 32 混合とした．また，融合の際の融合重みは事後的に最適な値を求めて使用している．

5 認識実験

各手法を用いた認識結果，及び，それらを融合した場合の認識結果を表 2 に示す．表中の A，S，E はそれぞれモデル A，モデル S，モデル E を用いた手法を表している．

モデル S，モデル E をそれぞれ単独で用いる場合を比較すると，モデル S の方が，モデル E よりも性能が高いことがわかる．また，この二つのモデルからの尤度を融合 (S+E) することで，それぞれを単独で用いる場合よりも良好な結果を得られている．この傾向はそれぞれをモデル A と融合した場合 ($A+S$, $A+E$) にも見られる．このことより，これらの 3 種類の情報はそれぞれがシーン認識性能の向上に役立つことがわかる．

さらに，3 つのモデルからの尤度の融合を行った場合 ($A+S+E$)，left-to-right 型 HMM で F 値の平均が 0.54 であった．これは，モデル A を単独で用いた場合と比べ，この尤度融合で 22.1%の精度の向上である．また，ergodic 型 HMM を用いる場合は left-to-right 型 HMM に比べ，若干性能が低いことがわかる．このことは，野球のシーン内において時間の経過に合わせて音響的イベントが特徴的に推移していることに起因しているためである可能性が高い．

モデル A に left-to-right 型 HMM を用い，3 種類のモデルからの尤度を融合した場合のシーンラベルごとの F 値は，SBF，ホームチーム得点，ビジターチーム得点がそれぞれ 0.85，0.72，0.68 と高く，長打，単打，アウトが 0.18，0.25，0.53 と低かった．このことから，盛り上がりが大きい得点シーンや，シーン時間長に特徴のある SBF は認識がしやすいことがわかる．また単打とアウトと SBF は混同しやすいなど，シーンラベルにより識別性能に差が出た．今回の実験で識別が困難であったシーンの認識性能を向上させるためには，さらに他の手法を用いる必要があると考えられる．

6 まとめ

本研究では，野球中継番組の音響情報を用いたシーン認識を行った．ケプストラム情報と，シーンの時間長，音響的イベントの継続時間長をそれぞれ用いる 3 種類の手法を融合することで，それぞれを単独で使う場合と比較して精度を向上できることを確認した．今後の課題としては，映像情報を用いた手法との融合などによるさらなる精度の向上，融合の際の重みの自動設定がある．また本研究で定義したシーン境界は，極めて高い精度で検出可能であることが報告されている [1]．そこで，シーン境界に自動検出したものを用いた性能評価を行いたい．また，今回の実験でシーンラベルとして用いなかったリプレイなどのシーンを用いた性能の検討も必要である．

謝辞 この研究は NHK 放送技術研究所との共同研究として行われました．ここに感謝の意を表します．また，21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤」の援助を受けました．

参考文献

[1] Nguyen Huu Bach, 篠田浩一, 古井貞熙, “隠れマルコフモデルを用いた野球放送の自動的インデクシング”, MIRU2005, pp.1113-1120 (2005-7).

[2] 山田一郎, 佐野雅規, 住吉英樹, 柴田正啓, 八木伸行, “アナウンスコメントを利用したサッカー番組メタデータ自動生成”, PRMU2004-122, pp.37-42(2005-2).

[3] 佐野雅規, 住吉英樹, 八木伸行, “サッカー中継に置ける会場音とスピーチを利用したメタデータ生成”, PRMU2005-120, pp.33-38 (2005-11).

[4] Yong Rui, Anoop Gupta, and Alex Acero, “Automatically Extracting Highlights for TV Baseball Programs”, ACM Multimedia, pp.105-115.