社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（56）
2004／12／22

知識を用いた音声認識による野球実況中継の構造化

佐古　淳†　有木　康雄††

† 神戸大学大学院自然科学研究科　〒657-8501 神戸市灘区六甲台町 1-1
†† 神戸大学工学部　〒657-8501 神戸市灘区六甲台町 1-1
E-mail: †sakoats@me.cs.scitec.kobe-u.ac.jp, ††ariki@kobe-u.ac.jp

あらまし　本研究では，野球の実況中継に対して構造化を行うことを目的としている．構造化を行うために，大語彙連続音声認識を利用する．音声認識では，発音は似ているが意味に照らして考えればあり得ないような単語を認識することがある．このような単語に対しては，野球の知識を用いることで認識誤りを低減できる．また，状況依存の情報として，アナウンサーの感情も利用する．音声認識の結果を用いて構造化を行いながら，推定された情報を音声認識に利用する手法として「状態推定音声認識」の枠組みを提案する．実験により，構造化を行うために重要と考えられるキーワード正解精度が 2.3%向上し，また，約 73.3%の構造化正解率が得られた．
キーワード　構造化，言語モデル，情報統合，HMM 音響モデル，スポーツ中継

Structuring Baseball Live Game Based on Knowledge Dependent Speech Recognition

Atsushi SAKO† and Yasuo ARIKI††

† Graduated School of Science and Technology, Kobe University
Rokkodai 1-1, Nada, Kobe, 657-8501 Japan
†† Faculty of Engineering, Kobe University　Rokkodai 1-1, Nada, Kobe, 657-8501 Japan
E-mail: †sakoats@me.cs.scitec.kobe-u.ac.jp, ††ariki@kobe-u.ac.jp

Abstract　The purpose of this study is to automatically structure sports live speech, especially baseball live speech using Large Vocabulary Continuous Speech Recognition system. Since it is a difficult problem to recognize baseball live speech. We propose in this paper a speech recognition method of incorporating the baseball game knowledge such as counting of inning, out, strike, ball and announcer's emotion. This method is formalized in the framework of probability theory and implemented in the conventional speech decoding (Viterbi) algorithm. This method enables to seek a word sequence as well as a structure sequence. The experimental results showed that the proposed approach improved the structuring and segmentation accuracy to 73.3% as well as keywords accuracy by 2.3%.
Key words　structuring, language model, information integration, HMM acoustic model, live sports

1. はじめに

近年，デジタルテレビや WWW などの発展により，多くの人が映像や音声などのマルチメディアコンテンツを大量に所有できるようになってきた．しかし，このような大量のコンテンツの中から本当に欲しい情報のみを探し出すことは困難である．検索を容易に行うためには，映像や音声などのコンテンツに対してインデックス情報を付与し，データベースを構築しておく必要がある．コンテンツが大量に存在することを考えると，人手でインデックス情報を付与することは大変なコストと労力を伴うため現実的でない．そこで，コンピュータが自動的にインデックス情報を付与できることが望まれる．本研究では，マルチメディアコンテンツとしてスポーツ実況中継，特に野球実況中継を対象としている．

野球中継を構成する要素のイメージを図 1 に示す．野球中継は映像と音声のシーケンスによって構成され，各シーケンスは表 1 のような情報を保持している．本研究では，図 1 の情報のうち，「イニング，アウトカウント，ボールカウント，ストライクカウント」といった情報を，

表 1　野球中継のシーケンスが持つ情報
Table 1　Baseball information of each sequence.

構造情報	イニング，アウトカウント， ストライクカウント，ボールカウント

シーケンスに対して連続的に付与していくことを構造化と定義している．

実況中継に対してインデックス情報を付与する研究としては，カメラワークを抽出して映像を構造化する手法 [2] や，映像中のテロップを解析する手法 [3]，クローズドキャプションを用いる手法 [4] などが提案されている．しかし，野球の実況中継に対して映像認識を行うことは極めて難しい．そこで，アナウンサーの実況中継音声を認識して構造化を行うアプローチを採用する．

実況中継音声としては，ラジオの実況中継音声を用いる．これは，映像がないために，ラジオの実況中継音声の方がテレビの実況中継音声よりも情報量が多いためである．しかし，その分，ラジオの実況中継音声は発話速度が速い，言い間違いが多い，発音がなまるなどの問題が生じる．さらに，球場の雰囲気などを伝えるために，感情的な発話を行う．これらの要因により，実況中継音声の認識は困難なタスクとなっている．

これらの問題に対するアプローチとして，我々は，音響モデル，及び言語モデルを野球実況中継音声に対して適応することにより，大幅に認識精度を向上させる手法について報告を行ってきた [1]．しかし，認識結果を見てみると，「阪神」と「三振」を間違って認識したり，「ファールボール」と「フォアボール」を間違えたりしている部分が見られる．これらは，発音はよく似ているものの，野球の知識に照らして考えれば全く異なるものである．「三振」や「フォアボール」といった単語は，野球のルールと関係して発話される前提や条件がある．この前提や条件を満たしていない発話は認識誤りである可能性が高い．このような野球に関する知識を用いることで，音声認識精度を向上させることができると考えられる．本稿では，野球実況中継の認識精度向上を目指して，野球中継に関する知識を用いた実況中継の音声認識手法を提案する．

以下，第 2. 章では，本研究で考える野球実況中継の知識について述べる．第 3. 章では，野球の知識を音声認識に用いる手法である状態推定音声認識について，第 4. 章では，認識，及び構造化実験とその結果について述べる．

2.　野球実況中継の知識

野球中継は，まず，攻守のチームによって「イニング」「表裏」にセグメントされる．さらに，イニングの中でも「ストライクカウント」「ボールカウント」「アウトカウント」にセグメントされている．また，これらの情報は，「アウトカウントは 0 から 2 まで」「ストライクカウントとボールカウントが同時に増えることはない」など，野球のルールに基づく制約に従っている．このような野球のルールに基づく情報は，全てピッチャーの投球を起点としている．そのため，構造化を行う際，映像と音声のシーケンスに対して連続的に情報を付与するのではなく，ピッチャーの投球毎に情報を付与するものとする．

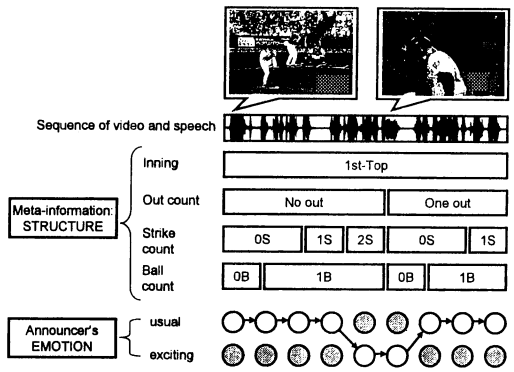

図 1　野球中継の構成要素
Fig. 1　Components of baseball game.

また，野球のルールとは別に，野球中継が持っている情報としてアナウンサーの感情があげられる．アナウンサーは試合の臨場感を伝えるため，しばしば感情的な発話を行う．このような発話は主に試合が大きく動くような場合にされることが多い．例えば，ヒットやホームランが出た場合，得点チャンスの場面での三振などである．このような情報は野球中継の状況に依存した知識と考えることができる．

次に，野球の試合進行とアナウンサーの発話との関係について考察する．アナウンサーは試合を見ながら発話を行うため，アナウンサーの発話は試合の進行，すなわち構造情報に依存している．例えば，「三振」という単語は，ストライクカウントが 2 である場合に発話されやすい，「フォアボール」という単語は，ボールカウントが 3 である場合に発話されやすい，といった特徴がある．また，これらの単語は，「第 3 球投げました，直球ストライク，三振」「ボールカウントツースリーから，投球ボール，フォアボール」といったように，投球に関する発話とセットでなされる場合がほとんどである．これらから，投球に関する発話とセットでない場合に「三振」と認識された場合は，「阪神」のような発音の似た単語の認識誤りではないかと推定したり，ボールカウントが 3 でない場合の「フォアボール」という発話は，同じく発音の似

た「ファールボール」の認識誤りではないかと推定したりできると考えられる．

本研究では，野球のルールやアナウンサーの感情の状態，構造情報と発話との関係を考慮した音声認識を行うことで野球実況中継の構造化を行うことを目的とする．次章で，このような野球実況中継の知識を，大語彙連続音声認識の枠組みで利用するための手法である，「状態推定音声認識」について述べる．

3. 状態推定音声認識

3.1 状態推定音声認識の定式化

一般的な音声認識は，観測される音声の特徴系列を $\mathbf{O} = \{o_1 \cdots o_t\}$，単語系列を $\mathbf{W} = \{w_1 \cdots w_N\}$ とすると，特徴系列 $\mathbf{O}$ に対して，尤もらしい単語系列 $\mathbf{W}$ を求めることに相当し，以下のように定式化される．

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{O}) \tag{1}$$

$$= \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W})P(\mathbf{W}). \tag{2}$$

ベイズの定理により式 2 が導かれる．$P(\mathbf{O}|\mathbf{W})$ は音響モデル，$P(\mathbf{W})$ は言語モデルである．言語モデルは通常，N-gram によって表されるため，一般的な音声認識は最終的に以下の式で表される．

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}) \prod_{i=1}^{N} P(w_i|w_{i-1}). \tag{3}$$

次に，音声認識に知識を用いる枠組みとして本研究で提案する状態推定音声認識の定式化について述べる．状態推定音声認識では，音声認識に利用したい知識を「状態の系列」として表し，観測される音声の特徴系列から，単語系列と状態系列を同時に推定する．本研究では，「状態」として「ストライクカウント，ボールカウント，アウトカウント，アナウンサーの興奮状態」を用いた．状態系列を $\mathbf{S} = \{s_1 \cdots s_N\}$ とすると，状態推定音声認識は以下のように定式化できる．

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{S}, \mathbf{W}|\mathbf{O}). \tag{4}$$

式 4 は，ベイズの定理により次のように展開できる．

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}, \mathbf{S})P(\mathbf{S}, \mathbf{W}). \tag{5}$$

ここで，$P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ は状態に依存した音響モデルと考えられる．特に，状態に含まれる「アナウンサーの感情」に依存する．また，$P(\mathbf{S}, \mathbf{W})$ は，さらに以下のように展開できる．

$$P(\mathbf{S}, \mathbf{W}) = P(s_1, \cdots, s_N, w_1, \cdots, w_N)$$
$$= P(w_1)P(s_1|w_1)$$
$$\times \prod_{i=2}^{N} P(w_i|w_1^{i-1}, s_1^{i-1})P(s_i|s_1^{i-1}, w_1^i). \tag{6}$$

このままでは式が複雑すぎるため，近似によって簡単化を行う．まず，$P(w_i|w_1^{i-1}, s_1^{i-1})$ について，

- 単語は直前の単語と直前の状態にのみ依存

と近似することにより，以下の式を得る．

$$P(w_i|w_1^{i-1}, s_1^{i-1}) \approx P(w_i|w_{i-1}, s_{i-1}) \tag{7}$$

これは，状態依存バイグラムと考えることができる．これによって，第 2. 章で述べた，「フォアボール」という単語はボールカウントが 3 のときに発話されやすい，といったような野球の知識を表現できると考えられる．次に，$P(s_i|s_1^{i-1}, w_1^i)$ について，

- 状態は直前の状態と現在の単語，及び状態遷移確率を最も高くするような，M 単語以内の現在の単語とのペアの単語

という近似を行う．これにより以下の式を得る．

$$P(s_i|s_1^{i-1}, w_1^i) \approx \max_{y=(i-1\cdots i-M)} P(s_i|s_{i-1}, w_i, w_y) \tag{8}$$

これは，単語のペアに依存した状態遷移モデルと考えることができる．「三振」という発話だけによってアウトカウントの遷移をとらえるよりも，「投げた，三振」のように，単語のペアによって遷移をとらえた方がより精度を高められる．ただし，単語のペアは，隣接するとは限らず，「投げた，直球ストライク，三振」のように，離れて現れることが多いため，M 単語以内の単語とのペアのうち，最も遷移確率が高いものを選択するようにした．

最終的に，状態推定音声認識は以下のように定式化できる．

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}, \mathbf{S})P(w_1)P(s_1|w_1)$$
$$\times \prod_{i=2}^{N} P(w_i|w_{i-1}, s_{i-1})$$
$$\max_{y=(i-1\cdots i-M)} P(s_i|s_{i-1}, w_i, w_y). \tag{9}$$

一般的な音声認識の定式化と比較して異なる点は以下の通りである．

- 音響モデルが状態に依存する．
- 言語モデルが状態に依存する．
- 発話された単語に依存して状態遷移する状態遷移モデルが新たに存在する．

状態系列とは，音声認識に利用したい「知識」を表現する系列である．状態に依存した音響モデル・言語モデルは，すなわち，知識に依存した音響モデル・言語モデルであると言える．このように，単語と同時に状態を推定することで，知識を用いた音声認識が可能となる．以下，これらの確率モデルの学習方法について述べる．

3.2 確率モデルの学習

本節では，状態推定音声認識で用いる各確率モデルの学習方法について述べる．

3.2.1 状態依存音響モデル

本研究では，ラジオの実況中継音声を音声認識することによって構造化を行っている．ラジオの実況中継音声は，講演音声と比較しても，発話速度が速く，雑音レベルや感情も強いといった特徴があり，難しいタスクとなっている．そのため，文献[1]と同じ手法を用いて教師ありの音響モデル適応を行う．適応のベースラインとなる音響モデルは，比較的「話し言葉」に近い特徴を持つ日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese) モニタ版[5]から作成した．また，適応手法として，MLLR[6]やMAP推定[7]を単独で用いるよりも高精度に適応できるとされるMLLR+MAP[8]を用いた．これは，MLLRによってモデルパラメータの変換を行い，それを事前知識としてMAP推定を行う手法である．

ただし，ここで，本研究では一般的な音声認識ではなく，状態推定音声認識を用いることから，音響モデルも状態に依存したものを学習する必要がある．野球の実況中継音声において，音声特徴が依存している状態として，アナウンサーの興奮状態が考えられる．そこで，アナウンサーの興奮状態「平常」「興奮」に応じて，2種類の音響モデルを作成する．以下に,「平常」「興奮」2種類の音響モデルを作成する手順について述べる．

- 適応データを人手で書き起こす．このとき適応データを「平常」「興奮」に分割する．
- 「平常」の適応データについて，ベースラインとなる音響モデルに対しMLLR+MAPを行う．
- 「興奮」の適応データは十分なデータ量があるとは言えない．そのため，まず,「平常」のデータを用いてベースラインとなる音響モデルに対し適応を行い，その後,「興奮」の適応データを用いて適応を行う．

以上のような状態依存の音響モデルを用いて認識を行う．

3.2.2 状態依存言語モデル

本節では，状態依存の言語モデルの学習法について述べる．状態依存の言語モデルは，$P(w_i|w_{i-1}, s_{i-1})$という式で表される．これは，第2.章で述べた,「フォアボール」という単語はボールカウントが3のときに発話されやすいといった特徴を表現する．状態依存言語モデルは，以下の手順で作成する．

- 実況中継の書き起こしテキストに対し，状態が変化すると考えられる単語にタグを付与する．
- MeCab[9]を用いて形態素解析を行う．
- タグを参照しながら，各単語に状態を割り当てる．

表2 状態に依存する単語
「感情に依存」とは，感情を伴って発話されやすい単語であり，「平常」「興奮」によって単語遷移確率が変化する．
Table 2 State dependent words.

構造に依存	三振，ボールカウントワンストライクツーボール，フォアボール，ツーアウトランナー二塁など
感情に依存	アウト，ヒット，ホームラン，三振など

- 状態毎に単語連鎖確率を計算する．

ここで，多くの単語はどの状態においても連鎖確率が変化しないことがわかる．その中でも，状態への依存が強く，状態に応じて連鎖確率が変化していた単語の例を表2に示す．特に,「ボールカウントワンストライクツーボール」など，状態を説明している単語は状態の推定誤りを修正する効果が期待できる．

状態に依存しない多くの単語については，上記の書き起こしテキストは少量であるため，これから求めた単語遷移確率よりも，より多くのコーパスから計算した単語遷移確率の方が精度が良いと考えられる．そこで，状態への依存が低い単語については，文献[1]の手法による言語モデルを用いた．すなわち，インターネットから集めたテキスト集合（Webテキストコーパス）から得られる言語モデルと，発話を書き起こしたテキスト集合（書き起こしテキストコーパス）から得られる言語モデルを融合して得られる言語モデル（結合モデル）を作成し，さらに「結合モデル」と「書き起こしテキストコーパス」から得られる言語モデルをパープレキシティが最小となるように重み付き融合したものを用いた．

3.2.3 単語依存状態遷移モデル

本節では，単語に依存した状態遷移モデルについて述べる．単語依存の状態遷移モデルは$\max_{y=(i-1\cdots i-M)} P(s_i|s_{i-1}, w_i, w_y)$という式で表される．状態遷移のイメージを図2に示す．まず，野球のルール上の制約により，ストライクカウントとボールカウントが同時に増えたり，一度に2つ増えたりすることはない．また，アウトになったり，フォアボールになったりすると，次の打者となり，カウントは0-0に戻る．状態遷移確率のうち，野球のルールが許さない遷移については，単語の依存を考慮せずに確率値ゼロを与えている．

次に，野球のルールが許す遷移について，その遷移が起こるときにアナウンサーがどのような発話を行っているかを考察する．例えば，ボールカウントが増加するような遷移においては，確かに「ボール」という発話が行われるが,「ボール」という単語は,「ボールが高く上がった」のような発話の中でも現れる．このことから，単純に単語ひとつを見て状態を遷移させることは出来ないと考えられる．本来は，$P(s_i|s_{i-1}, w_i \cdots w_{i-M})$のように過

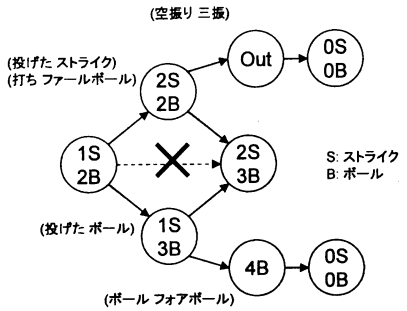

図 2　状態遷移モデル
Fig. 2　State transition model.

表 3　状態遷移を引き起こしやすい単語ペア
Table 3　Word pairs tend to raise state transition.

状態遷移	w_y	w_i
アウト	空振り	三振
	取り	アウト
ストライク	投げ	ストライク
	直球	ストライク
ボール	アウトコース	ボール
	第一球	ボール
フォアボール	投球	フォアボール
	ボール	フォアボール
ファールボール	投げ	ファール
	打ち	ファールボール

去 M 単語の系列によって状態を遷移させることが望ましい．しかし，これではデータスパースネスの問題が生じてしまう．そこで，過去 M 単語の中から，現在の単語との組み合わせで，最も状態遷移確率が大きくなる単語を選択する．これにより,「投げた，アウトコース，ボール」のような発話と「ボールが高く上がった」のような発話を区別し，状態遷移の精度を高めることができると考えられる．

状態遷移確率は以下の手順で計算する．学習データは，第 3.2.2 節で作成したものを用いる．この学習データは，実況中継の書き起こしテキストに，状態が遷移すると考えられる場所にタグを付与したものである．単語ペアに依存した状態遷移確率は以下の式で求められる．

$$P(s_i|s_{i-1}, w_i, w_y) = \frac{N(w_i, w_y, s_i|s_{i-1})}{N(w_i, w_y)} \qquad (10)$$

ここで，$N(x, y)$ は単語 x と y が M 単語の距離内に出現する回数である．表 3 に，状態遷移を引き起こす確率が高いと計算された単語のペアを示す．

4. 実　　験

4.1 実験条件

本研究で用いた音声データは，球場においてアナウンサーの接話マイクから録音したものを用いている．雑音については，観客の歓声があがったり，応援が盛り上がったりする場合には若干の雑音が混入するが，全体としては比較的クリーンな音声である．

テストセットとして，2003 年 9 月 7 日の阪神 vs 横浜戦の実況中継音声を用いた．音声は，約 120 分の音声データである．このうち，前半の 60 分を適応データとし，後半の 60 分を評価データとした．実況の形態は，アナウンサーが解説者と会話を行いながら実況するものである．ただし，音声に解説者の声は全く混入していない．

実況中継音声の発話スタイルは「読み上げ音声」よりも「話し言葉」に近い特徴を持つ．このため，ベースラインの音響モデルには日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese) モニター版 [5] のうち，男性話者 200 名の講演音声を用いて作成した．音響分析条件と HMM の仕様を表 4 に示す．

ベースラインの音響モデルには，長母音化を考慮した音節 HMM（音節数 244）を用いた．1 状態あたりの，混合分布数は 32 とした．また，母音 (V) は 5 状態 3 ループ，子音＋母音 (CV) は 7 状態 5 ループである [10]．サンプリング周波数は 16kHz，音響特徴量は 12 次元 MFCC と対数パワー，12 次元 MFCC の一次微分を加えた 25 次元である．また，ベースラインとなる言語モデルは，Web テキストコーパス（約 57 万形態素）から作成した．

認識の手順は次ぎの通り行った．まず，一般的な音声認識システムを用いて認識を行い，認識途中のワードグラフを出力する．このとき，平常の音響モデル，興奮の音響モデルをそれぞれ用いて，2 つのワードグラフを得る．この 2 つのワードグラフを状態依存音声認識のシステムに取り込んだ上で，スコアを計算し直した．一般的な音声認識システムには，最ゆう単語 back-off 接続によるものを用いた [11]．

イニング，及びイニングの表裏は実況中継音声中の長い無音区間から簡単に判別できるため，これらの情報は既知のものとして認識を行った．これらの条件で，構造化正解率を求めた．構造化正解率は，ピッチャーの各投球に対し，ストライクカウント，ボールカウント，アウトカウントが正解している割合である．また，野球の知識を音声認識に用いることでどの程度認識誤りを低減できるかを示すために表 5 のようなキーワードについて，キーワード正解精度を求めた．表 6 に結果を示す．

実験結果より，キーワード正解精度が改善しているこ

表 4　音響分析条件と HMM の仕様

Table 4　Condition of acoustic analysis and HMM specification.

音響分析	サンプリング周波数	16kHz
	特徴パラメータ	MFCC(25 次元)
	フレーム長	20ms
	フレーム周期	10ms
	窓タイプ	ハミング窓
HMM	タイプ	244 音節
	混合数	32 混合
	母音 (V)	5 状態 3 ループ
	子音+母音 (CV)	7 状態 5 ループ

表 5　キーワード一覧

Table 5　Keyword list.

ストライク，ボール，ファールボール， フォアボール，アウト，空振り，三振

表 6　実験結果

Table 6　Experimental results.

	ベースライン	提案手法
キーワード正解精度	66.8%	69.1%
構造化正解率	-	73.3%

とがわかる．これは，状態依存の言語モデル，及び単語依存の状態遷移モデルによって，より野球の試合の流れをを適切に表現できるような単語が選択された結果であると考えられる．また，状態推定音声認識が推定した「構造情報」の正解率は，73.3%であった．構造情報の推定を間違ったシーンとしては，きわどいクロスプレーの後のシーンや，満塁ホームランの後のシーンなどがあった．このようなシーンでは，アナウンサーがかなり興奮して発話するため，興奮の音響モデルでもほとんど認識ができなくなってしまう．そのため，状態遷移の推定を間違い，その間違いの為に，その後しばらく誤推定を続けるというような点が見られた．逆に，淡々と投球を続けるようなシーンは比較的精度よく推定できていた．

5. ま と め

本研究では，野球実況中継の構造化を目的として，状態推定音声認識の定式化，及び実験を行った．状態推定音声認識によって，野球実況中継の構造化を行いながら，同時に野球の知識を音声認識システムの中に利用することができる．構造化の情報が野球の知識に見合ったものとなるように音声認識を行うことでキーワード正解精度を 2.3%改善させることができた．また，73.3%の正解率で構造情報を正しく推定することが出来た．さらなる精度向上のためには，アナウンサーが激しく興奮した状態での発話を精度よく音声認識できるような手法について検討を行う必要がある．

文　献

[1] 有木康雄，緒方淳，藤本雅清，塚田清志，“音響・言語モデルの適応処理によるスポーツ実況中継の音声認識，”信学論 (D-II), vol.J87-D-II, no.6, pp.1208-1215, Jun.2004.

[2] 山本拓，佐藤宏介，千原國宏，“野球中継映像における各種プレイシーンの自動検索/編集システム，”2000 信学総大, 情報・システム 2, D12-77, p.247, 2000.

[3] 館山公一，川嶋稔夫，青木由直，“野球中継におけるシーン検索，”第 3 回知能情報メディアシンポジウム論文集, pp.195-202, 1997.

[4] 新田直子，馬場口登，北橋忠広，“言語の画像の情報統合によるスポーツ映像からの人物・アクション・イベント抽出，”信学技報, PRMU99-256, 2000.

[5] 古井貞熙，前川喜久雄，伊佐原均，“『話し言葉工学』プロジェクトのこれまでの成果と展望，”第 2 回話し言葉の科学と工学ワークショップ, pp.1-6, 2002.

[6] C.L.Leggetter and P.C.Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, ”Comput. Speech Lang., vol.9, pp.171-185, 1995.

[7] J.L.Gauvain and C.H.Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, ”IEEE Trans. Speech Ausio Process., vol.2, no.2, pp.291-298, 1994.

[8] 緒方淳，有木康雄，“音素事後確率に基づく信頼度を用いた音響モデルの教師なし適応，”信学技報, SP2001-105, 2001.

[9] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer, ”http://chasen.org/ taku/software/mecab/

[10] 緒方淳，有木康雄，“日本語話し言葉音声認識のための音節に基づく高精度な音響モデルの検討，”信学技報, SP2002-129, 2002.

[11] 緒方淳，有木康雄，“大語彙連続音声認識における最ゆう単語 back-off 接続を用いた効率的な N-best 探索法，”信学論 (D-II), Vol.J84-D-II, No.12, pp.2489-2500, 2001.