音声・状況の同時認識に基づく野球実況中継へのメタ情報付与

佐古　淳†　滝口　哲也†　有木　康雄†

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

あらまし　近年，多くのマルチメディア・コンテンツの所有が可能となってきた．大量のコンテンツの中から欲しい情報を得るためには，検索のためのメタ情報を付与しておく必要がある．本研究では，マルチメディア・コンテンツの一例としてスポーツ実況中継，特に野球実況中継に注目し，実況中継音声から音声認識を用いてメタ情報を抽出することを目的としている．野球のメタ情報としては，今何が起こっているかを表すイベントと，その積み重ねである状況が存在すると考えられる．まず，現実にイベントや状況が存在し，これを基にアナウンサーは実況を行う．本研究では，実況音声から単語列だけを推定する音声認識を拡張し，実況音声から単語列・イベント系列・状況系列全てを同時に推定する音声認識手法を提案する．定式化により，イベント依存音響モデル，状況遷移モデル，イベント推定モデル，状況依存言語モデルを得る．これらを確率の枠組みで統合的に用いることで，単語列とメタ情報の同時推定を行う．実験により，イベント検出 F 値 0.87，イベント正解率 0.86，状況正解率 0.77 を得た．

Extracting Meta-Information for Sports Live Games based on Speech and Situation Recognition

Atsushi SAKO†, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Guraduate School of Science and Technology, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
E-mail: †sakoats@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　Recently a large quantity of multimedia contents are broadcast and accessed. In order to retrieve exactly what we want to know from multimedia database, automatic extraction of meta-information is required. We focused on live speeches, especially baseball commentary speeches as a kind of multimedia contents. The purpose of this study is to provide meta-information based on speech recognition techniques. Events and situations are defined as meta-information. First of all, an event is occured or a situation is changed, then an announcer speaks based on an event and a situation. In this paper, we propose a extended speech recognition technique that estimates not only a word sequence but also a event sequence and a situation sequence concurrently. As a result of formulation, event dependent acoustic model, situation transition model, event estimation model and situation dependent language model are derived. A word sequence and meta-information are estimated based on these models. The experimental results showed that the proposed method provided meta-information with a high degree of accuracy.

1. はじめに

近年，デジタルテレビや WWW などの発展により，映像や音声など，多くのマルチメディア・コンテンツを所有することが可能となってきた．このような大量のコンテンツに対しては，ユーザーが欲しい情報を検索できる必要がある．また，すべてのコンテンツを視聴するには時間がかかりすぎるため，要点だけ，または好みのシーンだけを抜き出して視聴したいという要求も存在する．一方で，放送局の立場としては，出来るだけコストをかけずに新たなコンテンツを生み出したいという要求が存在する．ユーザーが望むシーンのみを提供可能になれば，新たなコンテンツ・ビジネスとなる可能性がある．

このような要求を満たすためには，映像や音声などのコンテンツに対してメタ情報を作成しておく必要があると考えた．メタ情報を利用することにより，自動的にハイライトシーンを提供したり，ユーザーの望むシーンを検索することが可能になると期待できる．また，計算機により自動的にメタ情報

の付与が出来れば，人手で付与する場合に比べ，コストと労力の削減が可能となる．

本研究では，マルチメディア・コンテンツのひとつとして，スポーツ実況中継，特に野球実況中継を対象としてメタ情報の作成を目的としている．メタ情報として，イベントの種別（投球・投球結果・ヒット・アウト・ホームラン・タイムリーなど）と状況（イニング・表裏・アウトカウント・出塁状況・ボールカウント）を定義する．メタ情報の詳細については 2. 節で述べる．

野球中継に対してメタ情報を作成する研究としては，いくつかの手法が提案されている．カメラワークを抽出して映像を構造化する手法 [1] や，映像中のテロップを解析する手法 [2]，クローズドキャプションを用いる手法 [3] などが提案されている．しかし，野球の実況中継に対して正確に映像認識を行うことは難しい．そこで，本研究ではアナウンサーの実況中継音声を音声認識し，この言語情報を元にメタ情報を作成するアプローチを採用した．これは，放送局の立場として，コンテンツの成立に最低限必要な実況中継音声のみを用いてメタ情報を作成可能であるという利点がある．また，学術的観点からは，音声のみからどのように状況を認識しているかという理解の助けになるものと期待できる．本研究では，テレビではなく，ラジオの実況中継音声を用いたこれは，映像がないために，音声のみで状況が理解出来るように実況中継がなされることによる．そのため，ラジオの実況中継音声には，テレビのものより情報が多く含まれる．ただし，実験では，実際の放送音声ではなく，アナウンサーの発話のみを収録した音声を用いた．これは放送用に解説者の音声や環境音と合成される前の音声である．

音声認識を用いてメタ情報を作成するにあたって，誤認識が問題となる．野球中継における実況放送音声の認識性能向上については，音響モデル適応・言語モデル適応を用いた手法が提案されている [4]．本研究も，同じく，音響モデル適応・言語モデル適応を行う．しかし，それでも単語正解精度は 7 割程度であり，依然として音声認識誤りが含まれている．従って，認識誤りに対して頑健なメタ情報推定方法が必要となる．本研究では，実況音声がイベント・状況といったメタ情報に依存して生成されると仮定し，実況音声から単語列だけでなく，イベント・状況まで同時に推定する音声認識を提案する．これにより，音声認識器が最終的に出力する 1-Best の認識結果だけでなく，はっきりと単語が確定していない認識仮説（本研究ではワードグラフを用いた）まで利用した統合的なメタ情報の付与が可能となる．

2. 野球中継に関するメタ情報

本節では，本研究で作成すべきメタ情報について述べる．本研究では，野球の実況中継音声を対象としている．

野球の実況中継音声は，アナウンサーが野球の試合進行に基づいて，これを伝えるために発話を行ったものである．ただし，試合進行とは直接関係しない，選手の情報や球場の様子，解説者との会話等の発話も行う．本研究では，アナウンサーの発話の内容を端的に表したものを「イベント」と呼ぶこととした．具体的なイベントとして，表 1 のうち「重要なイベント」「その他のイベント」の内容を設定した．重要なイベントは，投球やストライク，ボール，ヒット，アウトなど，試合が進行し，試合の状況が変化する際になされる発話に対して付与される内容を設定した．また，その他のイベントは,「解説者との会話」や，選手の説明，現在の試合状況の説明などの「実況一般」，打者がボールを打ち，試合の状況がどう展開するかはっきり決まらない段階を表す「守備の実況」を設定した．これらは，試合の状態が変化しない場合になされる発話に対して付与される内容である．検索の際に用いられるイベントは，試合進行に関係する前者のものが多いだろうという考えから，前者のイベントを「重要なイベント」とした．

表 1　メタ情報の定義
※についてはコーパス中には現れなかった
Table 1　The specification of meta-information

重要なイベント
投球　ストライク　ボール　ファール フォアボール　三振　牽制球　盗塁 ヒット　ツーベース　スリーベース ランニングホームラン※　ホームラン　得点 アウト　ダブルプレー　トリプルプレー※ ボーク※　デッドボール
その他のイベント
解説者との会話　実況一般　守備の実況
状況
イニング　表裏　得点　ストライクカウント ボールカウント　アウトカウント　出塁状況

本研究ではさらに，イベントの積み重ねとして,「状況」を定義した．状況は試合進行の状況と一致するように，イニング，及び表裏，得点，アウトカウント，ストライクカウント，ボールカウント，出塁状況を定義した．本研究では,「イベント」「状況」をあわせて，メタ情報と定義した．表 2 に実際の実況の書き起こし例と，人手で付与したメタ情報を記す．このような情報を自動的に付与することが本研究の目的である．ただし，ひとつのイベントが必ずひとつの発話で実況されるわけではない．この際，両方に同じラベルを付与すると状況の遷移がおかしくなるため，どちらかの発話を選んでラベルを付与するようにした．

3. 提 案 手 法

本節では，音声認識とメタ情報の付与を統合し，これらを同時に行う手法について述べる．以下，発話数を T，1 発話内のフレーム数を F_t，単語数を N_t とし，観測音声特徴の系列を $\mathbf{U} = \{\mathbf{O}_1, \cdots, \mathbf{O}_T\}$，$\mathbf{O}_t = \{o_1, \cdots, o_{F_t}\}$ とする．1 発話の長さは，おおよそ書き起こしテキストの句点から句点までの長さである．ただし，書き起こしテキストを作成する際，句点の挿入に明確な基準があったわけではない．このため，句点は恣意的に挿入されたものである．また，発話の系列を $\mathbf{D} = \{\mathbf{W}_1, \cdots, \mathbf{W}_T\}$，1 発話内の単語の系列を $\mathbf{W}_t = \{w_1, \cdots, w_{N_t}\}$ とする．イベントと状況のメタ情報は 1 発話毎に付与し，イベントの系列 $\mathbf{E} = \{e_1, \cdots, e_T\}$，状況の系列 $\mathbf{S} = \{s_1, \cdots, s_T\}$ とする．本研究においては，観測音声特徴の系列 $\mathbf{U}$ から，発話系列 $\mathbf{D}$，イベント系列 $\mathbf{E}$，状況系列 $\mathbf{S}$ を同時に推定することが目的となる．まず，次節において，本研究で仮定した実況中継音声の生成モデルについて述べる．

3.1 実況中継音声の生成モデル

本研究で仮定する実況中継音声の生成モデルを図 1 に示す．実況中継音声は，試合の状況や起こったイベントをアナウンサーが “実況” したものである．このため，発話される単語は，試合の状況とイベントに依存して生成されると考えた．また，起こったイベントの種類によっては，アナウンサーは

表 2　実況中継に対するイベント・状況の具体例
Table 2　An example of events and situations for a commentary speech

発話	イベント	状況
ほんとうにこの人は鉄人ですね，ええ.	会話	1 回裏 0 対 0 1S 2B 2O 1 塁
全イニング出場，全試合全イニング出場しています，バッターボックスの○○.	実況一般	1 回裏 0 対 0 1S 2B 2O 1 塁
ボールカウントワンストライクトゥーボール，投球四球目.	実況一般	1 回裏 0 対 0 1S 2B 2O 1 塁
ピッチャーの××，足が上がって第四球を投げました.	投球	1 回裏 0 対 0 1S 2B 2O 1 塁
高目，ストライク決まりました.	ストライク	1 回裏 0 対 0 1S 2B 2O 1 塁
シュートぎみ，アウトコースいっぱいに決まっています.	実況一般	1 回裏 0 対 0 2S 2B 2O 1 塁
ボールカウントトゥーストライクトゥーボール.	実況一般	1 回裏 0 対 0 2S 2B 2O 1 塁

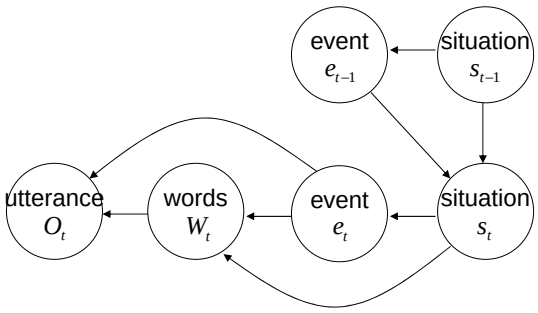

図 1　実況中継音声の生成モデル
Fig. 1　Generation model of commentary speeches.

興奮を交えて実況する場合がある．そこで，音声は，単語とイベントに依存して生成されるものと考えた．さらに，イベントは状況に依存して生成されると考えた．これは，状況に応じて起こるイベントと起こらないイベントが存在するためである．最後に，状況は，以前の状況と以前のイベントに依存して生成されると考えた．以前の状況においてイベントが発生し，それにより変化した状況が現在の状況であることを意味している．これらをふまえ，次節において提案手法の定式化について述べる．

3.2 定 式 化

通常の音声認識では，発話全体の観測音声特徴の系列 $\mathbf{U} = \{\mathbf{O}_1, \cdots, \mathbf{O}_T\}$ から，発話の系列 $\mathbf{D} = \{\mathbf{W}_1, \cdots, \mathbf{W}_T\}$ を推定する．すなわち，観測特徴系列 $\mathbf{U}$ が既知の条件で，発話の系列 $\mathbf{D}$ が生成される確率 $P(\mathbf{D}|\mathbf{U})$ が最も高くなるような発話系列 $\hat{\mathbf{D}}$ を推定する．

$$\begin{aligned}\hat{\mathbf{D}} &= \operatorname*{argmax}_{\mathbf{D}} P(\mathbf{D}|\mathbf{U}) \\ &= \operatorname*{argmax}_{\mathbf{D}} P(\mathbf{W}_1, \cdots, \mathbf{W}_T|\mathbf{O}_1, \cdots, \mathbf{O}_T) \\ &= \operatorname*{argmax}_{\mathbf{D}} \prod_{t=1}^{T} P(\mathbf{W}_t|\mathbf{W}_1^{t-1}, \mathbf{O}_1^{t-1}) \\ &\qquad \times P(\mathbf{O}_t|\mathbf{W}_1^{t}, \mathbf{O}_1^{t-1}). \end{aligned} \tag{1}$$

ベイズの定理により，式 1 が導かれる．ここで，

- 発話 $\mathbf{W}_t$ は他の情報に依存しない
- 観測音声 $\mathbf{O}_t$ は発話 $\mathbf{W}_t$ のにみ依存する

との仮定をおくと，式 1 は，

$$\hat{\mathbf{D}} = \operatorname*{argmax}_{\mathbf{D}} \prod_{t=1}^{T} P(\mathbf{W}_t) P(\mathbf{O}_t|\mathbf{W}_t) \tag{2}$$

となる．ただし，$P(\mathbf{U})^{-1}$ は $\mathbf{D}$ によらないため無視した．$P(\mathbf{O}_t|\mathbf{W}_t)$ は音響モデル，$P(\mathbf{W}_t)$ は言語モデルである．

これに対し，本研究では，発話全体の観測音声特徴の系列 $\mathbf{U} = \{\mathbf{O}_1, \cdots, \mathbf{O}_T\}$ から，尤もらしい発話系列 $\mathbf{D} = \{\mathbf{W}_1, \cdots, \mathbf{W}_T\}$，イベント系列 $\mathbf{E} = \{e_1, \cdots, e_T\}$，状況系列 $\mathbf{S} = \{s_1, \cdots, s_T\}$ を同時に推定する．

$$\begin{aligned}(\hat{\mathbf{D}}, \hat{\mathbf{E}}, \hat{\mathbf{S}}) &= \operatorname*{argmax}_{(\mathbf{D},\mathbf{E},\mathbf{S})} P(\mathbf{D}, \mathbf{E}, \mathbf{S}|\mathbf{U}). \\ &= \operatorname*{argmax}_{(\mathbf{D},\mathbf{E},\mathbf{S})} P(\mathbf{W}_1\ T, e_1^T, s_1^T|\mathbf{O}_1^T) \\ &= \prod_{t=1}^{T} P(s_t|\mathbf{O}_1^{t-1}, \mathbf{W}_1^{t-1}, e_1^{t-1}, s_1^{t-1}) \\ &\qquad \times P(e_t|\mathbf{O}_1^{t-1}, \mathbf{W}_1^{t-1}, e_1^{t-1}, s_1^{t}) \\ &\qquad \times P(\mathbf{W}_t|\mathbf{O}_1^{t-1}, \mathbf{W}_1^{t-1}, e_1^{t}, s_1^{t}) \\ &\qquad \times P(\mathbf{O}_t|\mathbf{O}_1^{t-1}, \mathbf{W}_1^{t}, e_1^{t}, s_1^{t}). \end{aligned} \tag{3}$$

ここで，3.1 節に留意し，以下の仮定を置く．

- 状況 s_t は直前の状況 s_{t-1} と直前のイベント e_{t-1} にのみ依存する
- イベント e_t は状況 s_t にのみ依存する
- 単語列 $\mathbf{W}_t$ はイベント e_t と状況 s_t にのみ依存する
- 観測特徴 $\mathbf{O}_t$ は，単語列 $\mathbf{W}_t$ とイベント e_t にのみ依存する

以上の仮定により次式が導かれる．

$$\begin{aligned}(\hat{\mathbf{D}}, \hat{\mathbf{E}}, \hat{\mathbf{S}}) &= \prod_{t=1}^{T} P(s_t|e_{t-1}, s_{t-1}) P(e_t|s_t) \\ &\qquad \times P(\mathbf{W}_t|e_t, s_t) P(\mathbf{O}_t|\mathbf{W}_t, e_t). \end{aligned} \tag{4}$$

$P(s_t|e_{t-1}, s_{t-1})$ は状況遷移モデル，$P(e_t|s_t)$ はイベント生成確率，$P(\mathbf{W}_t|e_t, s_t)$ はイベント・状況依存言語モデル，$P(\mathbf{O}_t|\mathbf{W}_t, e_t)$ はイベントに依存した音響モデルである．ここで，イベント・状況依存言語モデルは，各イベント・各状況毎に作成した trigram を用いた．また，イベント依存音響モデルには，各イベント毎に作成した HMM を用いた．本手法では，イベント・状況に依存して単語列を生成するモデル $P(\mathbf{W}_t|e_t, s_t)$（実際にはイベント・状況依存 trigram）がイベント及び状況の推定に大きな役割を果たす．

一方，$P(\mathbf{W}_t|e_t, s_t)$ を

$$P(\mathbf{W}_t|e_t, s_t) = \frac{P(\mathbf{W}_t|s_t) P(e_t|\mathbf{W}_t, s_t)}{P(e_t|s_t)}. \tag{5}$$

のように変形すると，言語モデルは状況のみに依存するようになり，認識仮説からイベントを推定するモデル $P(e_t|\mathbf{W}_t, s_t)$ が現れる．これを式 4 に代入すると次式が導かれる．

$$\begin{aligned}(\hat{\mathbf{D}}, \hat{\mathbf{E}}, \hat{\mathbf{S}}) &= \prod_{t=1}^{T} P(s_t|e_{t-1}, s_{t-1}) P(\mathbf{W}_t|s_t) \\ &\qquad \times P(e_t|\mathbf{W}_t, s_t) P(\mathbf{O}_t|\mathbf{W}_t, e_t). \end{aligned} \tag{6}$$

状況遷移モデル，及びイベント依存音響モデルについては式 4 と同様である．一方，イベント・状況依存言語モデル，及びイベント生成確率は，状況依存言語モデルとイベント推定モデルへ変化している．また，通常の音声認識である式 2 と比較すると，言語モデル・音響モデルがそれぞれ状況依存・イベント依存に変化し，新しく状況遷移モデル・イベント推定モデルが追加されている．本手法では，イベント推定モデル $P(e_t|\mathbf{W}_t, s_t)$ を用いて，イベント e_t の推定を認識仮説 $\mathbf{W}_t$ から識別的に行う．この点が，イベント及び状況を生成モデルによって推定する式 4 の手法と大きく異なっている．

以下，3.3 節から 3.6 節において，各モデルの詳細について述べる．

3.3 状況依存言語モデル

本研究で用いる言語モデルは trigram をベースとする．学習データ量の関係から，同じ状況の発話だけを集め，そこから trigram を構築することは困難である．そこで，言語モデル適応を用いて，状況依存の言語モデルを構築する．言語モデル適応には，N-gram 出現回数の重み付き混合による手法 [5] を用いた．

また，比較のためのイベントと状況の両方に依存した言語モデルについては，まず状況に応じた適応を行った後で，イベントに応じた適応を再度行うことにより構築した．状況については，状況依存言語モデルと同様，イニング・表裏・得点を無視した．イベントについては，全てのイベントを考慮し，適応を行った．

3.4 イベント推定モデル

イベント推定モデル $P(e_t|\mathbf{W}_t, s_t)$ は，認識仮説の単語列からイベントを推定するモデルである．認識仮説の単語列は 10 単語程度になることもあり，数式通りに確率を計算すると相当にスパースなモデルとなってしまう．そこで，本研究では，識別的な手法である AdaBoost [6] を用いて認識仮説の単語集合からイベントを推定する．AdaBoost は 2 値判別手法であるため，多種のイベントを識別するためには one-vs-rest 法などを用いた拡張が必要である [8]．one-vs-rest 法では，あるクラス k に着目し，クラス k とそれ以外のクラスを識別する強識別器 $f_k(\mathbf{W}_t)$ を全クラス数分作成する．

また，Adaboost の出力は確率ではないため，そのままでは提案手法と統合して用いることができない．しかしながら，AdaBoost の出力スコアは，識別境界面からの距離と捉えられることが報告されている [7]．すなわち，境界に近いほど確率 0.5 に近く，境界から遠いほど確率 0，もしくは確率 1 に近づくと考えられる．本研究では，完全な確率とは言えないものの，式 8 の sigmoid 関数を利用し，AdaBoost の出力スコアを擬似確率化する．

$$P(e_t = k|\mathbf{W}_t, s_t) = \frac{sigmoid(f_k(\mathbf{W}_t))}{\sum_l sigmoid(f_l(\mathbf{W}_t))}, \quad (7)$$

$$sigmoid(x) = \frac{1}{1 + exp(-w_1 x - w_0)}. \quad (8)$$

3.5 状況遷移モデル

状況遷移モデルは，試合進行の状況がイベントによって変化していくことを表すルールモデルとなる．モデルの一部を図 2 に示す．このモデルにより，ストライクカウントは 2 まで，一度に 1 ずつしか増えない，フォアボールイベントが生じるのはボールカウントが 3 のときのみ，といった野球のルールを表現することが出来る．また，「ストライクイベントが起きた際にはストライクカウントが 1 増える」といったルールが表現されていることも重要な点である．状況は，

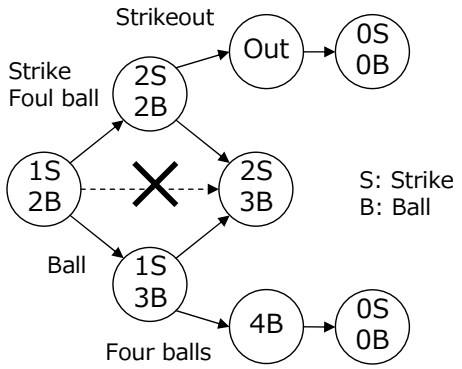

図 2　状況遷移モデル
Fig. 2　Situation transition model.

状況依存言語モデルからも推定可能である．「ボールカウント，ワンエンドツー．投げた．ストライク．ボールカウント，ツーエンドツーになりました」と実況している場合は，状況依存言語モデルからだけでも，どの点で状況が変化したか推定することが出来る．しかし，「ボールカウント，ワンエンドツー．投げた．ストライク」とだけ実況している場合は，状況依存言語モデルからでは変化を捉えることが出来ない．本モデルを用いて，“以前の状況”，及び “ストライクイベントの発生” から現在の状況を推定する必要がある．

3.6 イベント依存音響モデル

本研究で取り扱う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（講演音声 7.31mora/s に対し，実況中継音声 8.51mora/s），雑音レベルが強い，感情の起伏も激しいといった特徴がある [4]．本研究では認識性能向上のため，文献 [4] と同じ手法を用いて教師有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 [9] から作成した．適応手法として MLLR+MAP [10] を用いた．

これに加え，イベント毎の音響モデル適応も行う．まず，上記の手法で CSJ コーパスから作成したモデルに，全てのイベントを含む実況中継音声を用いて適応を行う．こうして得られたモデルをイベントの数だけ複製し，それぞれにイベント毎の実況中継音声を用いて再度適応を行う．これによりイベントに依存した音響モデルを得る．守備の実況やホームランイベント等において，臨場感を伝えるための興奮した音声に適応されたモデルが得られるものと考えられる．

4. 実　　験

実験の目的は，提案手法によってどの程度の精度でメタ情報を付与することが出来るか確かめることと，提案手法の中の 4 つのモデルそれぞれが，どの程度精度に寄与しているか確認することである．ここで，精度の指標として，以下の 3 点を用いた．

- 重要なイベント検出の F 値
- 重要なイベントの正解率（推定されたイベントと正解ラベルが一致しているか）
- 投球毎の状況正解率

1 つ目の重要なイベント検出の F 値は，実況一般や解説者との会話といった検索の条件になりにくいイベントの中から，検索の条件になるホームランや三振などの重要なイベントが正しく検出出来ているかを調べるための指標である．なお，

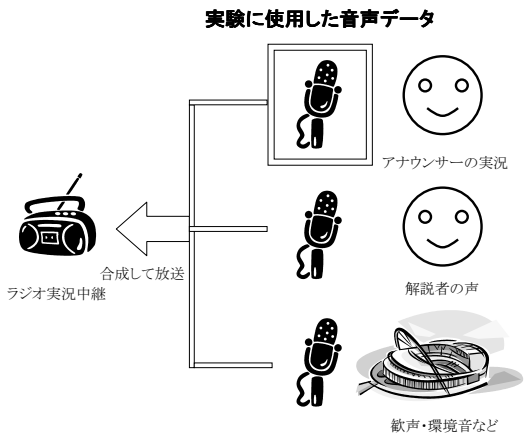

図 3　実験で使用した音声データ
Fig. 3　Recording environment of speech corpus.

重要なイベントは 2. 節で定義したものを用いた．2 つ目の重要なイベントの正解率は，重要なイベントが正しく検出された上でそのラベルが正解ラベルと等しい割合である．イベントの内容まで正しく推定出来ているかを調べるための指標である．3 つ目の投球毎の状況正解率は状況を正しく推定出来ているかを調べるための指標である．発話毎でなく投球毎である理由は，本来，投球毎にしか試合の状況が変化しないこと，実際はアナウンサーの言い直し等により投球以外の場合でも状況が変化する場合があることによる．以上の指標を用いて，提案手法の評価実験を行った．まず，人手による書き起こしテキストを用いたメタ情報付与実験を行い，システムの上限値を調べた．その後，実況中継音声に対して，提案手法を用いてメタ情報付与実験を行った．

ただし，実験に際しては，式 6 の各モデルに，重みを乗じて用いた．これは音声認識における言語重みと同様の調整パラメータである．

4.1　学習コーパスの仕様

本節では，実験で用いた学習コーパスについて述べる．実況中継音声は，ラジオの音声を用いた．これは，映像がないため，テレビの実況中継より音声の情報量が多いためである．使用した音声データは，図 3 の通りラジオ放送される前の，アナウンサーのみの音声を収録したものである．球場の歓声などはノイズとして重畳はしているものの，音声認識に大きく影響をあたえるほどではない．また，解説者との会話は存在するが，解説者の声は含まれていない．発話速度が速い，言い間違いが多い，発音がなまけているなどの特徴があり，音声認識にとっては困難なタスクとなっている．

発話の単位は，人手による書き起こしテキストにおいて句点で区切られた単位とした．音声認識，及びメタ情報の付与はともに，一発話毎に行った．メタ情報付与のための学習データは，発話毎にメタ情報ラベルを人手で付与し，作成した．学習データの分量を表 3 に示す．全 4 試合で，1 試合当たり約 2000 発話であった．総発話数は約 9000，単語数は 80K，異なり単語数（辞書サイズ）は約 3000 語であった．

4.2　実況中継音声を用いたメタ情報の付与

次に，実況中継音声から提案手法を用いてメタ情報を付与する実験を行った．本節では，実況中継音声からのメタ情報付与精度，音声認識とメタ情報付与の統合の効果，各モデルの精度に対する寄与，音声認識率とメタ情報付与精度の関係の 4 点について明らかにする．まず，音声認識の条件と結果について述べる．

表 3　学習コーパスの仕様
Table 3　The specification of our corpus

日時	話者	時間	発話数	単語数
2003/09/05	A	1.73	2232	21K
2003/09/06	B	1.81	2210	22K
2003/09/15	B	1.76	2320	21K
2003/09/16	A	1.61	2010	20K

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　音声認識結果の単語正解精度
Table 5　Word accuracy of the speech recognition results.

単語正解精度	キーワード F 値 (再現率/適合率)
63.8%	0.76 (0.74/0.77)

表 6　AdaBoost によって選択された素性語の例
Table 6　An example of features selected by AdaBoost.

メタ情報識別素性例
あたり あんまり きのう ほんと よく アウトー
インサイド ストライク スリー ツーアウト バッター ボール
ワン 一塁 回っ 外れ 甘い 監督 詰まっ 球 空振り 牽制
三振 始まり 送球 打ち 直球 変化球 方向 etc.

4.2.1　音声認識条件と結果

ベースラインの音響モデルは，日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese) モニター版 [9] 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．これらの条件で音響モデルを作成し，さらに，MLLR+MAP [10] により音響モデル適応を行った．音響モデル適応は，同一話者の別の日時の実況中継音声を用いて行った．適応データの分量は，約 1 時間半であった．

言語モデルは，野球実況中継音声の書き起こしテキストから trigram モデルを作成した．異なり単語数は約 3,000，コーパスサイズは約 8 万形態素であった．

4 つの試合の音声認識結果の単語正解精度の平均，及びキーワードの F 値の平均を表 5 に示す．ここでのキーワードは AdaBoost によって学習された素性を用いた（表 6）．以後，これらの認識結果を用いて実験を行った．

4.2.2　音声認識結果に対するメタ情報の付与

実況中継音声に対し，提案手法によるメタ情報付与実験を行った．学習とテストは 4 fold のクロス・バリデーション法により行った．音響モデル・言語モデルともにオープンの条件において，提案手法を用いた場合のメタ情報付与実験結果を表 7 に示す．ここで，“1-best” の結果は，提案手法ではなく，音声認識とメタ情報付与を統合しない場合の結果である．すなわち，まず通常の音声認識結果を出力し，その 1 通

表 7　提案手法によるメタ情報付与実験結果
Table 7　Results of extracting meta-information using recognized transcription.

	提案手法	1-best	SE trigram
イベント検出 F 値	**0.87**	0.85	0.75
(再現率/適合率)	(0.88/0.85)	(0.85/0.84)	(0.73/0.76)
イベント正解率	**0.86**	0.83	0.78
状況正解率	0.77	0.74	0.67
単語正解制度	63.9%	63.8%	63.9%

りの結果に対して提案手法の各モデルを用いてメタ情報付与を行ったものである．認識と統合せず，認識仮説を用いないところが提案手法と異なる点である．音響モデルについても，イベント依存のものではなく，通常の音響モデルを用いた．また，“SE trigram (Situation and Event dependent trigram)” は，式 6 の提案手法ではなく，イベント・状況依存言語モデルを用いる式 4 に基づく手法である．

結果の示す通り，高精度が識別が可能であった．ただし，イベントの正解率を詳細に見てみると，投球・投球の結果については比較的高い精度を保っているものの，ヒットやアウトといった試合が大きく動くイベントの正解率に低下が見られた．このようなイベントの際には，臨場感を伝えるためにアナウンサーが興奮して発話を行う場合があり音声認識率が低下する．認識率の低下がイベントの正解率に影響を与えているものと考えられる．提案手法では，イベント依存音響モデルの効果により興奮した音声の音声認識率改善を期待していたが，効果は限定的であった．

次に，提案手法による結果を “1-best” と比較した場合について述べる．認識の結果，イベント検出 F 値及びイベント正解率については提案手法が高い性能を示している．これらについては，1-best の結果との比較で，二項分布の平均の差の検定により有意水準 5%で有意であった．提案手法では，状況やイベントまで考慮に入れた上で認識仮説を選択することが可能である．これに対し，“1-best” では，音声認識誤りを修正する手段を持たない．提案手法では，例えば具体例として，ボールカウントが 3 でない場合の「フォアボール」を正しい「ファールボール」に修正出来ている例や,「三振」と「阪神」の誤認識が修正されている例などが見られた．ただし，上記のようなキーワードについてはいくつか改善が見られたが，大部分の認識結果は共通しており，単語正解精度としてはほとんど変化がなかった．状況に依存する単語が，ルールに関連する用語などにある程度限定されているものと考えられる．認識結果が変わらない部分については提案手法と “1-best” は同じ結果となった．また，“SE trigram” では，精度の低下が見られた．AdaBoost を利用したイベント推定と異なり，イベント・状況依存言語モデルでは，識別性能が低下してしまうものと考えられる．状況正解率については，1-best との比較で，有意水準 5%では有意な差とならなかった．原因として，ヒットなどの認識性能が大きく低下する発話に対しては，提案手法でもほとんど改善が得られていないことが考えられる．さらには，一度の誤りがしばらく連鎖して続くため，全体的に正解率が低下してしまうことが考えられる．

5. ま と め

本稿では，音声認識とイベント・状況推定を同時に行うことにより，メタ情報を付与する手法について述べた．実況音声が生成される過程をモデル化し，観測音声特徴から発話系列・イベント系列・状況系列を同時に推定するよう定式化を行った．これにより，イベント依存音響モデル・状況遷移モデル・イベント推定モデル・状況依存言語モデルを得た．特に，イベント推定モデルについては，識別的手法である AdaBoost を用いた．AdaBoost の出力スコアは確率ではないため，sigmoid 関数を利用し擬似確率化して用いた．

実験の結果，提案手法でイベント検出 F 値 0.87，イベント正解率 0.86，状況正解率 0.77 を得た．これは，AdaBoost の素性を学習する際，音声認識結果を学習データとすることにより，認識誤りに対して頑健な素性が選択されるためと考えられる．提案手法を用いることにより，実況音声が生成される背後にある知識が利用可能となり，これによって音声認識誤りが改善する例が見られ，メタ情報付与性能も向上した．

本研究では，状況やイベントが定義しやすい野球を対象として研究を行った．しかしながら，他のスポーツへ展開するためには，本研究で定義した状況は “堅すぎる” であろう．特に，球技では，ボールや選手の位置が状況として重要な意味を持つ可能性がある．音声からでは伝えにくいこのような状況を理解するためには，映像情報との統合も検討する必要がある．今後の課題として，その他のスポーツへのメタ情報の付与を行うため，映像の利用による位置情報の抽出，ボールや選手の位置といった，緩い状況を表現可能なモデルの提案などの研究が必要である．また，同時に，メタ情報には様々な粒度が存在する（単なるアウトか，フライでのアウトか，ファールフライかなど）．緩い状況の表現とともに，より詳細な状況やイベントを表現するためにも，階層的なイベント・状況モデルの研究が必要である．

文　献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 伊藤彰則，好田正紀，“N-gram 出現回数の混合によるタスク適応の性能解析,” 信学論 (D-II), vol.J83-D-II, no.11, pp.2418–2427, Nov.2000.

[6] R.Schapire, Y.Freund, P.Bartlett, and W.Lee, “Boosting the margin: A new explanation for the effectiveness of voting mathods,” Annals of Statistics, vol.26, no.5, pp.1651-1686, Oct. 1998.

[7] 小野田崇, “Boosting の過学習とその回避,” 電子情報通信学会論文誌, Vol.J85-D2, No.5, pp. 776-784, 2002 年 5 月.

[8] Ethem Alpaydin, “Introduction To Machine Learning (Adaptive Computation and Machine Learning),” The MIT Press, 2004.

[9] 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.

[10] 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.