

単語共起に注目した野球実況中継の構造化*

◎佐古 淳 滝口 哲也 有木 康雄（神戸大・工）

1 はじめに

近年、デジタルテレビやWWWの発展により、大量のマルチメディアコンテンツを所有できるようになってきた。しかし、大量のコンテンツを思い通りに扱うためには、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。我々は、音声認識技術を用いて、マルチメディアコンテンツ、特に野球実況中継に対しインデックス情報を付与する手法について研究を行ってきた[1]。また、より詳細なインデックス情報を付与するため、野球中継のシーケンスに対して、イニング・アウトカウント・ストライクカウント・ボールカウントを連続的に付与することを構造化と定義している。

音声認識技術を用いる際の問題点として、単語認識誤りがあげられる。現在の音声認識技術では、単語認識誤りを完全に無くすことは困難であると言える。この問題に対し、一般的な音声認識で用いられているよりも高次の情報を用いて、単語誤りを防ぐことが考えられる。この点に関し、我々は、“状況”というより高次の情報を音声認識に利用する“状況推定音声認識”について提案を行っている。これにより、インデックス情報を推定すると共に、単語認識誤りを低減させることが可能となる[2]。本稿では、単語の共起を基に“状況”を推定する手法について検討・報告する。

2 状況推定音声認識

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。

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

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

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。

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

これに対し、提案する状況推定音声認識では、観測される音声の特徴系列から、単語系列と共に“状況”の

系列を同時に推定する。状況の系列を $\mathbf{S} = \{s_1 \cdots s_N\}$ とすると、状況推定音声認識は以下のように定式化できる。

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

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

$P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ は状況に依存した音響モデルと考えられる。また、 $P(\mathbf{S}, \mathbf{W})$ は、さらに以下のように展開できる。

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(s_1, \cdots, s_N, w_1, \cdots, w_N) \\ &= \prod_i^N P(s_i|s_1^{i-1}, w_1^i)P(w_i|w_1^{i-1}, s_1^{i-1}). \\ &\approx \prod_i^N P(s_i|s_{i-1}, w_{i-M}^i)P(w_i|w_{i-1}, s_{i-1}). \end{aligned} \quad (6)$$

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。

- 音響モデルが状況に依存する。
- 発話された単語に依存して状況遷移する状況遷移モデルが存在する。
- 言語モデルが状況に依存する。

このように状況推定音声認識を用いることにより、状況の推定を行うと同時に、推定した状況を知識として音声認識に利用することが可能となる。

3 単語共起に注目した状況推定音声認識

アナウンサーは試合を見ながら発話を行うため、その発話は試合の進行、すなわち“状況”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。このような特徴は、状況推定音声認識における「状況に依存した言語モデル」によって表現可能である。

一方、「ボール」の様な一般的な単語の場合、「ボール」という単語が認識されると、投球がボールと判定されてボールカウントが増えるのか、単に物体の“ボール”を指しているだけなのか判別できない。投球がボール判定されたのであれば、状況を遷移させる必要があるが、そうでないならば遷移させてはいけない。この状況遷移は、本来、アナウンサーが発話した文章の“意味”を理解して行われることが望ましいが、これは困難であると考えられる。

ここで、野球中継において、状態を遷移させるような発話は、単語単独ではなく、他の単語と共に現れやすいことに注目する。すなわち、状態が遷移する

*Structuring the Baseball Game Based on Word Cooccurrences after Speech Recognition.
by Atsushi Sako, Tetsuya Takiguchi and Yasuo Ariki(Kobe Univ.)

表 1: 状況遷移を引き起こしやすい単語ペア

状況遷移	w_y	w_i
アウト	空振り 取り	三振 アウト
ストライク	投げ 直球	ストライク 空振り
ボール	投げ 第一球	ボール ボール
フォアボール	投球 ボール	フォアボール フォアボール
ファールボール	投げ 打ち	ファール ファールボール

場合の「ボール」は、「投げた、ボール」や「第一球、外側ボール」のように投球に関する単語「投げ」や「第一球」などと共に現れやすい。そこで、本研究では、次のように、近似的に M 単語以内に共起する単語のうち、状況遷移確率の最も高い 2 単語を選択し、次の状況を推定する。

$$P(s_i | s_{i-1}, w_{i-M}^i) \approx \max_{y=(i-1 \dots i-M)} P(s_i | s_{i-1}, w_i, w_y) \quad (7)$$

これにより、「投げ」と「ボール」が M 単語以内に共起している場合は、状況遷移の確率が高くなり、ボールカウントが増える確率が高くなるという推定が可能となる。この式に従って、野球実況中継の書き起こしテキストから、自動的に単語共起に依存した状況遷移確率を推定した。その結果推定された、状況遷移を引き起こしやすい単語共起関係の一部を抜粋し、表 1 に示す。M としては 10 単語を設定した。10 単語中に、投球に関する発話（「投げ」や「投球」、「第一球」など）と、その結果（「ストライク」や「ボール」、「三振」など）を同時に含むだけでなく、投球に関する単語共起だけでなく、「取り-アウト」や「打ち-ファールボール」のような単語共起関係も自動で学習できていることがわかる。

4 実験・結果

以上のような認識機構を用いて、野球実況中継の音声認識及び、構造化の実験を行った。野球実況中継の音声は、発話速度が速い、言い直しが多い、発音がなまけているなど、音声認識にとって困難なタスクとなっている。ただし、アナウンサーの数は限られているので、予め音響モデル適応を行うことが可能である。本研究では、約 1 時間のデータを用いて MLLR+MAP による教師有り適応を行った。ベースラインとなる音響モデルとして、日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese) モニター版を用いて、男性話者 200 名の講演音声から、長母音化を考慮した音節 HMM (音節数 244) を作成した。1 状態あたりの混合分布数は 32、母音 (CV) は 5 状態 3 ループ、母音+子音 (CV) は 7 状態 5 ループである。サンプリング周波数は 16kHz、音響特徴量は 12 次元 MFCC と Δ MFCC、 Δ パワーの計 25 次元である。これに対し、MLLR+MAP によって実況中継音声を「平常」と「興奮」に切り分けたものを用いてそれぞれ教師有り適応を行った。

表 2: キーワード一覧

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

また、言語的にも、話し言葉であり、独特の言い回しが多く存在する。このため、言語モデルには、Web から収集した野球に関するテキストコーパス (57 万形態素) と、実際の中継の書き起こしテキストをパープレキシティが最小となるように重み付け融合したものを用いた。

状況推定音声認識に用いるモデルは、以下のよう

- 状況依存音響モデル: 「平常」「興奮」に切り分けられた音声を教師有り適応したものを平常状況依存の音響モデル及び、興奮状況依存の音響モデルとした。
- 状況遷移モデル: 3 章の手法を用いて、単語共起に依存した状況遷移確率を計算した。
- 状況依存言語モデル: 学習データより自動的に状況依存のバイグラム確率を求めた。

これらの認識機構を用いて実験を行い、キーワード正解精度 (KW-Acc) 及び、構造化正解率を求めた。構造化正解率は、各投球毎にストライクカウント及び、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。キーワードは、特に構造化に関係が深いと考えられる表 2 の単語を設定した。“従来”は、音響・

表 3: 実験結果

	従来	状況推定	状況推定 単語共起
KW-Acc	66.8%	67.9%	69.1%
構造化正解率	-	71.4%	73.3%

言語モデル適応のみで状況推定音声認識を用いない従来手法、“状況推定”は、単語共起を用いない状況推定音声認識である。結果から、単語共起を用いることで、従来の状況推定音声認識よりもキーワード正解精度、構造化正解率共に向上していることがわかる。

5 おわりに

状況推定音声認識において、単語共起に依存して状況遷移するモデルについて検討を行った。実験より、単語共起を考慮しない場合に比べて、キーワード正解精度が向上することが確認できた。また、状況遷移の推定がより確かになることから、構造化正解率も上昇した。

参考文献

- [1] 有木他, 信学論, pp.1208-1215(2004).
- [2] 佐古他, 音響学会, pp.35-36(2004-03)