フレーズに基づく状態推定音声認識の検討*

◎佐古 淳，滝口 哲也，有木 康雄（神戸大・工）

1 はじめに

高精度な音声認識を行う上で，言語モデルは重要な役割を果たしている．現在までに様々な言語モデルが提案されてきた．中でも，N-gram モデルは学習の容易さ，単語予測性能の高さなどから広く用いられている [1]．しかし，従来の言語モデルは，発話が行われる状況を考慮していなかった．状況によって発話内容は影響をうけるため，これをモデル化することでより高精度な認識が可能になる．この点に対し，我々は，状況と単語系列を同時に推定することで言語モデルの状況に依存させることのできる状態推定音声認識について提案を行ってきた [2]．野球実況中継構造化タスクにおいて，野球の試合状況を状態とすることで，試合状況に関連するキーワードに関して認識性能を向上させられることを確認した．

一方で，試合状況を状態とした場合では，試合状況に関連しない単語については改善が得られなかった．そのため，全体としての単語正解精度はあまり改善されていない．一般的な単語について改善を得るには，試合状況よりも汎用的な状態を設定する必要がある．

ここで，野球実況中継は試合進行を解説することが主な目的であることに注目する．アナウンサーは，試合進行を視聴者に伝えるために発話を行う．このとき，試合進行の表現には，「投げた，直球，ストライク」「４回の表ベイスターズの攻撃」など試合進行に応じて特定のフレーズが現れる．そのため，どのような試合進行を表現したいかと，どのようなフレーズを発話したいかは同様の概念を示しているものと考えられる．そこで，試合進行を状態としてモデル化するよりも，フレーズの集合を状態としてモデル化する方が汎用的であり，また，自動的に状態を学習可能なことから，本研究ではフレーズを状態とした状態推定音声認識について検討を行う．

2 状態推定音声認識

ここで，状態推定音声認識について概説する．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．

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

$P(\mathbf{O}|\mathbf{W})$ 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．

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}) \prod_{i=1}^{N} P(w_i|w_{i-1}). \quad (2)$$

これに対し，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 “状態” の系列を同時に推定する．本研究では，状態を “発話しようとしているフレーズ” と定義している．状態の系列を $\mathbf{S} = \{s_1 \cdots s_N\}$ 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．

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{S}, \mathbf{W}|\mathbf{O}) \quad (3)$$

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

$P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ は状態に依存した音響モデルと考えられるが，簡単のため状態からは独立とし，通常の音響モデルと同一の $P(\mathbf{O}|\mathbf{W})$ とする．また，$P(\mathbf{S}, \mathbf{W})$ は，さらに以下のように展開できる．

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(s_1, \cdots, s_N, w_1, \cdots, w_N) \\ &= \prod_{i}^{N} P(s_i|s_1^{i-1}, w_1^{i-1})P(w_i|w_1^{i-1}, s_1^{i}). \\ &\approx \prod_{i}^{N} P(s_i|s_{i-1}, w_{i-1})P(w_i|w_{i-1}, s_i). \quad (5) \end{aligned}$$

$P(s_i|s_{i-1}, w_{i-1})$ はフレーズの遷移確率を表現するモデルとなる．ただし，フレーズとは，ひとかたまりの単語の集合であると考えられるため，フレーズの途中で別のフレーズへ遷移することは好ましくない．そこで w_{i-1} がフレーズの末尾に相当する単語の場合にのみ遷移確率を与え，それ以外の場合は遷移確率をゼロとする．次に，$P(w_i|w_{i-1}, s_i)$ はフレーズに依存した言語モデルとなる．これは，各フレーズにおいて bi-gram 確率を計算したものとする．提案手法では，一般的な音声認識の定式化と比較して，

- フレーズ遷移モデルが存在する．
- 言語モデルがフレーズに依存する．

という点が異なっている．以上のような認識機構を用いることで，各フレーズにおいてどのような発話がなされやすいかという点を考慮した音声認識を行うことが可能となる．

---

*Studies on state dependent speech recognition based on phrases. by Atsushi Sako, Tetsuya Takiguchi and Yasuo Ariki (Kobe Univ.)

3　フレーズモデルの学習

試合進行を表現するようなフレーズは，頻繁に発話されるものが多い．そこで，学習テキスト中に出現する複数個の単語列のうち，しきい値以上の回数出現するものをフレーズとした．ただし，あまり短い単語列は意味を持たないため，4 単語以上の単語列に限定した．この中でフレーズには含まれなかったものについては,「フレーズ外状態」として bi-gram を作成した．

これだけでは，フレーズの数が多くなりすぎてしまう．また，フレーズ内の単語が少し入れ替わる程度のものについては同一のフレーズと考える方が自然である．そこで，類似したフレーズのクラスタリングを行う．クラスタリングは，以下の手順で行った．まず，ひとつのフレーズを文書とみなして LSA(Latent Semantic Analysis)[3] を行う．次に LSA によって求められた意味ベクトルが最も類似しているフレーズの組を求め，これをひとつにまとめる．ここで，意味ベクトルの類似度は以下の式によって求めた．

$$\cos(\mathbf{x},\mathbf{y}) \quad = \quad \frac{\mathbf{x}^T \cdot \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||} \qquad (6)$$

以上の操作をフレーズの数が一定のしきい値以下になるまで繰り返した．その後，各フレーズにおいて bi-gram 確率を計算した．

以上の操作によって求められたフレーズが，学習テキスト中にどのような順で現れるかを調べることでフレーズ遷移モデルを学習する．ただし，多くのフレーズについて，フレーズからフレーズ外状態へ，フレーズ外状態からフレーズへ，という遷移がみられた．このことから，フレーズ遷移モデルは，フレーズの連鎖モデルとしての意味はあまりもたず，フレーズの途中で別のフレーズに遷移してしまわないようにするという点において意味を持っていると考えられる．

4　実験・結果

以上のような認識機構，及びフレーズモデルを用いて認識実験を行った．認識タスクは野球実況中継音声である．野球実況中継音声はテレビではなくラジオのものを用いた．デコーダーには Julius 3.4.2[4] を用いた．まず，最初に状態推定音声認識は用いずに，通常のトライグラムモデルを用いて認識を行う．このとき，ベストパスだけではなく，十分に深い複数のパスを出力する．その後，状態推定音声認識の認識機構を用いてリスコアリングを行い，最終的に最もスコアの高いパスを認識結果とする．すなわち，bi-gram をファーストパス，tri-gram をセカンドパス，フレーズモデルをサードパスとして用いる．

tri-gram で用いる言語モデルは，実況中継音声の書き起こしテキスト（約 8 万形態素）から palmkit[5] を用いて作成した．学習データは，話し言葉である実況中継音声を書き起こしたものである．テストセット・パープレキシティは 60.3 であった．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．ベースラインの音響モデルを，日本語話し言葉コーパス（CSJ:Corpus of Spontaneous Japanese）モニター版 [6] のうち男性話者 200 名の講演音声を用いて作成した．これに対し，MLLR+MAP によって実況中継音声（約 1 時間）の教師あり適応を行ったものを用いた．音響モデル適応に用いた音声は，テストセットと同一の話者のものを用いた．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．これらの音響モデル及び，言語モデルを用いて実験を行った．

表 1 に結果を示す．実験結果から，本研究で提案した音声認識により，単語正解精度が 1.3%向上していることがわかる．これは，フレーズを考慮することにより，不自然な発話が減少したことによるものと考えられる．

Table 1　実験結果

	従来手法	提案手法
単語正解精度	70.9%	**72.2%**

5　おわりに

「どのようなフレーズを発話しようとしているか」を状態として，単語列と共に推定する状態推定音声認識について検討を行った．実況中継音声書き起こしテキストコーパスから頻出するフレーズを学習しフレーズモデルとして認識に用いた．実験より，単語正解精度において 1.3%の改善が得られることがわかった．

参考文献

[1] 北，“確率的言語モデル”，東京大学出版会，2001.
[2] 佐古，滝口，有木，音講論 (春)，149-150，2005.
[3] S. Deerwester *et al.*, “Indexing by latent semantic analysis,” 1990.
[4] 河原ら，情処学会，SLP-49-57, 2003.
[5] http://palmkit.sourceforge.net/
[6] 古井他, 第 2 回話し言葉の科学と工学ワークショップ, pp.1-6(2002).