2L-01

テンプレート切り出しによる不特定話者対応のワードスポッティング

下萩原 勉†，浜田 玲子‡，井手 一郎‡，坂井 修一‡，田中 英彦‡
{tsutomu，reiko，ide，sakai，tanaka} @mtl.t.u-tokyo.ac.jp
†東京大学工学部，‡東京大学大学院工学系研究科

1. 研究の背景と目的

筆者らは映像に付随するテキスト教材の存在する教養番組（料理番組）に着目し，映像とテキスト教材の内容の対応づけを目指している[1]．この中で，音声の内容は，対応づけのための大きな手がかりになると考えられる．

本稿では，このような手がかりを得るために，入力音声中から付随テキストより抽出したキーワードを検出し，同時に話者を識別することを目的とする．この際，語彙を限定することによる精度の向上が期待できるワードスポッティングの利用が適していると考える．

具体的には，(1)入力音声からテンプレートを切り出し，そのテンプレートを用いて再度スポッティングを行い，(2)一方で話者適応したテンプレートにより話者を識別する．これらにより，語彙限定・不特定話者に対するワードスポッティングの精度の向上を試みる．

2．テンプレート切り出しによる不特定話者対応のワードスポッティング

2-1 テンプレート切り出しによる不特定話者対応

音声信号は，話者や話し方，発話環境によって大きくばらつくが，音声認識時は，これらのばらつきを何らかの形で吸収することが求められる．音響的特性が大きく異なるものについて複数個のモデルを用いる手法や，入力音声の先頭のデータなどを用いて，入力話者個人のスペクトル特性を反映した音声言語単位モデルに変形する手法がある．

本手法では，ワードスポッティングを 2 回に分けて行うことで不特定話者対応を行う，最初はしきい値を低めに設定することで候補を挙げ，その候補に対して，同一話者・同一単語の発話区間を検出した情報と比較して話者識別をする．さらに，入力音声から検出された単語を切り出してテンプレートとすることにより，第 2 回目のスポッティングの精度向上を計る．

2-2 提案手法の概要

処理の流れを図 1 に示し，以下に順に説明する．

2-2-1 標準テンプレート作成

まず，テキスト教材からキーワードを抽出し，これに対する標準テンプレートを用意する．

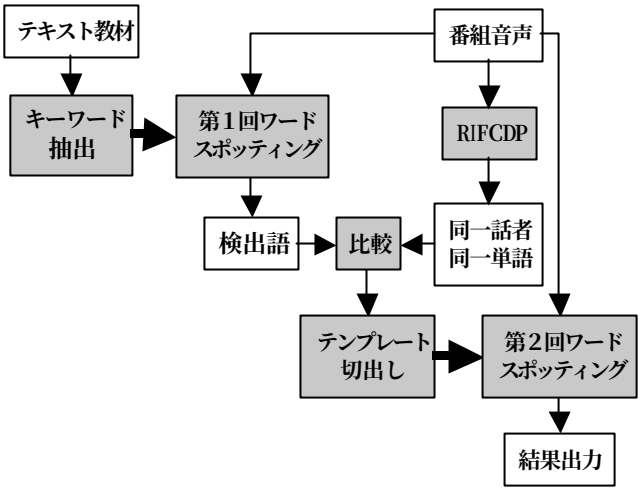

図 1: 提案手法の全体像

2-2-2 第 1 回目のワードスポッティング

用意した標準テンプレートを標準パターンに，番組音声を入力パターンにして，テンプレート候補抽出のためのワードスポッティングを行う(図 2)．ワードスポッティングには DP マッチングを用いる．ここでは，複数話者検出のため，しきい値をやや低めに設定し，検出の条件を緩めておく．

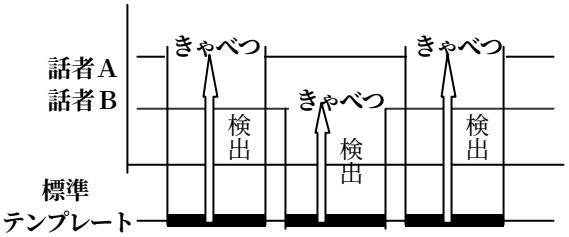

図 2: 第 1 回目のワードスポッティング

2-2-3 同一話者・同一単語発話区間の検出

話者識別を行い話者特定のテンプレートを作成するため，同一話者・同一単語の発話区間を検出する(図 3)．このために，2 つの時系列データ間で，任意の長さをもち，かつ互いに類似した区間の対を検出する RIFCDP という，連続DP を拡張した手法[2]を用いる．

2-2-4 特定話者のテンプレート切り出し

第 1 回目のワードスポッティングで検出された発話区間と RIFCDP で検出された発話区間を比較する．一致するものがあった場合は，同一のキーワードを同一人が 2 回以上発言したことを意味し，第 1 回目の

"Word-spotting designed for unspecific speakers referring to templates extracted from input voice",
Tsutomu Shimohagihara†, Reiko Hamada‡, Ichiro Ide‡, Shuichi Sakai‡, Hidehiko Tanaka‡, †Faculty of Engineering, ‡Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

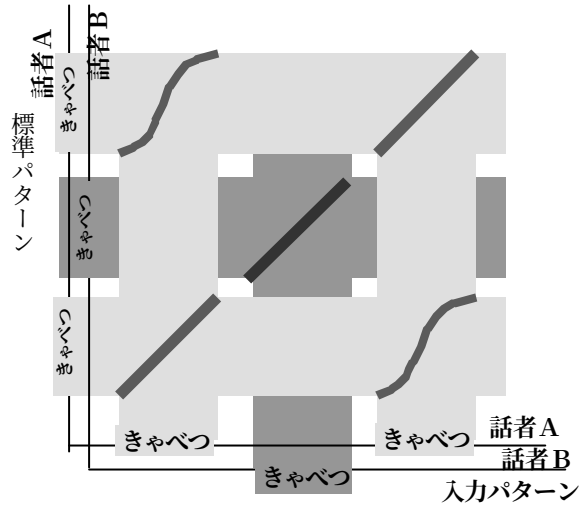

図 3: RIFCDP による同一話者・同一単語の検出

ワードスポッティングで検出された発話区間の話者を識別することができる．

次に，第1回目のスポッティングで検出された全ての発話区間を切り出し，そのうち，同一話者の発話であると識別されたものについて，DP マッチングによる平均化を行い，特定話者テンプレートを作成する．

2-2-5 第 2 回目のワードスポッティング

話者特定テンプレートを標準パターンに，番組音声を入力パターンにして，第 2 回目のワードスポッティングを行う(図 4)．テンプレートと番組音声中の該当する発話区間は類似しているはずなので，しきい値は高めに設定し，話者以外の発言を検出しないように注意する．

ここでの出力が最終結果であり，番組音声中より，テキスト教材から抽出したキーワードを，話者毎に区別して検出したことになる．

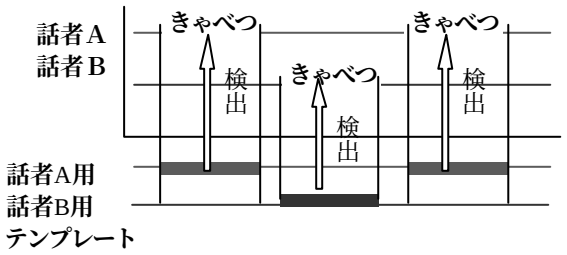

図 4: 第 2 回目のワードスポッティング

3．実験：話者特定テンプレートによるスポッティング

話者固有の差異，及び，入力音声から切り出した発話区間を平均化して話者特定テンプレートを作成する効果を検討するため，実験を行った．

3-1 実験条件

テレビで放送された料理番組から，番組音声約 7 分を録音し，実験に用いた．

- 話者：女性 2 人（A, B とおく）
- スポットする語：「ホワイトソース」
  (6 ヶ所存在：W_1〜W_6とおく)

次にこれを表 1 のように平均化したものをテンプレートとし，ワードスポッティングを行った．

表 1 テンプレートの平均化

平均化したデータ	元のデータ	話者
W_{123}	W_1,W_2,W_3	A
W_{456}	W_4,W_5,W_6	B
W_{56}	W_5,W_6	B

3-2 実験結果と考察

各発話区間に反応したテンプレートは表 2 のようになった．W_{123}, W_{456} はそれぞれ自身と同一話者の発話区間を検出するのに成功した．

一方，W_{56} は W_4 を含めず，W_5,W_6 のみを平均化したものであるが，W_4 を検出するのに成功している．これは，平均化により同一話者・同一単語の特徴が洗練されたためと思われる．すなわち，第 1 回目のスポッティングで W_4 を検出できなかった場合でも，W_5,W_6 を検出できていれば，W_{56} をテンプレートとし，第2回目のスポッティングで W_4 を検出できることを意味する．

表 2 スポッティングの検出結果

発話区間	反応したテンプレート
W_1,W_2,W_3	W_{123}
W_4,W_5,W_6	W_{456}, W_{56}

4．おわりに

本稿では，テキスト教材中のキーワードを入力音声中から検出する手法を提案した．実験で明らかになったように，音声データとは別に予め用意した標準テンプレートを用いて完全に検出ができなくとも，同一話者の他の発話区間を用いることにより検出できることが分かった．

謝辞

DP によるワードスポッティング及び RIFCDP のプログラムは，技術研究組合新情報処理開発機構（RWC）のご好意により提供して頂いた．深く感謝する．

参考文献

[1] 浜田玲子，井手一郎，坂井修一，田中英彦：“料理番組とテキスト教材の対応付け”，第5回知能情報メディアシンポジウム（IIM'99）論文集 pp.69-74，1999

[2]西村 拓一，古川 清，向井 理朗，岡 隆一：“時系列パターン検索のための重み減衰型 Reference Interval-Free 連続 DP について”，電子情報通信学会論文誌 Vol.J81-D-II，No.3，Mar. 1998.