音声言語情報処理 37−9
（2001．7．14）

キーワードスポッティングによる 商品紹介映像の商品区間への分割方法の検討

藤本 雅清 † 鷹尾 誠一 † 有木 康雄 † 松本 宏 ‡

† 龍谷大学 理工学部
〒 520-2194 大津市瀬田大江町横谷 1-5 Tel: 077-543-7427
E-mail: {masa, tail}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

‡ 住友電気工業株式会社 CAE研究センター
〒 554-0024 大阪市此花区島屋 1-1-3 Tel: 06-6466-5606
E-mail: mh@sei.co.jp

あらまし 本研究では，社内で製作された商品の紹介映像を個々の商品区間へ分割 (トピックセグメンテーション) し，商品名をインデックスとして付与するシステムの検討を行った．本研究におけるシステムでは，商品紹介映像の音声から音楽などの雑音を除去した後にキーワードスポッティングを行い，抽出された商品名を用いてトピックセグメンテーションを行っている．また，キーワードスポッティングにより商品名を抽出するためには，商品名辞書が必要となるが，本研究では，商品名辞書が事前に存在していない場合に，映像中のテロップ文字を利用して，オンラインで自動生成する手法についても検討を行った．実験の結果，商品名辞書が事前に存在している場合で約82%，商品名辞書を自動生成した場合で約60%の精度で区間分割を行うことができた．

キーワード：トピックセグメンテーション，キーワードスポッティング，雑音除去，オンライン辞書

A Study on A Method to Segment Goods Catalog Video into Individual Sections Based on Keyword Spotting

Masakiyo Fujimoto † Seiichi Takao † Yasuo Ariki † Hiroshi Matsumoto ‡

† Faculty of Science and Technology, Ryukoku University
1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan Tel: +81-77-543-7427
E-mail: {masa, tail}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

‡ CAE Research Center, Sumitomo Electric Industries, Ltd.
1-1-3, Shimaya, Konohana-ku, Osaka, 554-0024 Japan Tel: +81-6-6466-5606
E-mail: mh@sei.co.jp

Abstract In this paper, we propose a method to segment goods catalog video into individual sections and index them. Our proposing method uses the keyword spotting which extract the keywords from noise reduced speech signal within the goods catalog video. In order to extract the keywords by using keyword spotting, the goods name dictionary is required. In this paper, we study a method to generate the goods name dictionary automatically, by using the video captions within the goods catalog video. As the experimental result, the proposed method could segment the individual goods sections with approximately 82% accuracy when the goods name dictionary is available, and with approximately 60% accuracy when goods name dictionary is generated automatically.

Key words: topic segmentation, keyword spotting, noise reduction, on-line dictionary

1　はじめに

近年，放送の多チャンネル化により，多くのニュース番組が放映されるようになった．このような状況においては，ユーザにとって興味のあるニュースだけを見たいという要求が生じてくる．これを背景に近年，ニュースに対するトピックセグメンテーション [1]-[3] やトピック検索 [4][5]，パッセージ検索 [6][7]，クロスメディア検索 [8]，クロスメディア·パッセージ検索 [9]，ブラウジング検索 [10] などの研究が行われてきた．また，ニュース以外のメディアにおいても，ユーザにとって興味のある部分だけを知りたいという要求が生じている．この点から本研究では，社内で製作された商品の紹介映像を，個々の商品区間へ分割 (トピックセグメンテーション) し，分割した区間に商品名を索引として付与する技術の検討を行った．

また，商品の紹介映像を商品区間へ分割するにあたり，商品の紹介映像に含まれる商品名が既知であり，商品名辞書が存在する場合と，存在しない場合が考えられる．商品名辞書が存在する場合には，映像中にどのような商品が存在するかが既知であるため，個々の商品区間への分割は比較的容易であると考えられる．しかし，商品名辞書が存在しない場合は，商品に関する情報は全くの未知であり，個々の商品区間への分割が困難な問題になると考えられる．この問題を解決するために，本研究では，商品名辞書が存在しない場合において，商品名辞書を商品の紹介映像の中から自動で生成する手法について検討を行った．

2　トピックセグメンテーション

2.1　問題点

ニュース映像に対するトピックセグメンテーションは，ニュース映像から抽出したニュース音声の書き起こし結果を，統計的に言語処理することで行われている [1]-[3]．しかし，この方法では，ニュース音声の書き起こし結果の中に多くの誤り単語が含まれていると，後段の統計的言語処理が上手く機能しないといった問題が存在する．本研究で対象とする商品紹介映像には，常時，音楽などの雑音が音声に重畳しており，雑音除去 [11] を行ってから，大語彙連続音声認識により音声を自動的に書き起こしても，多くの誤り単語がその中に含まれてしまう．従って，従来法 [1]-[3] を用いて，トピックセグメンテーションを行うことは極めて困難である．

この問題点を回避するために本研究では，あらかじめ登録されたキーワードのみを抽出する，キーワードスポッティングを用いることにより商品紹介映像の音声から商品名を抽出し，トピックセグメンテーションを行った．

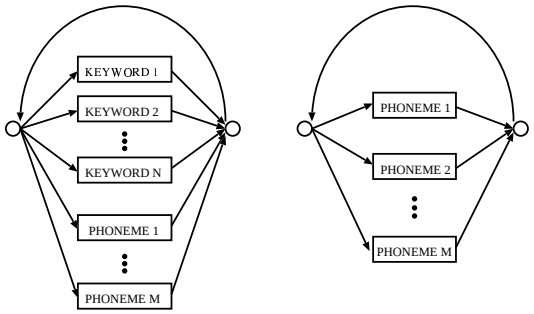
2.2　解決方法

商品の紹介映像において，個々の商品紹介映像の始まりでは，解説者が必ず商品名を発話している．従って本研究では，これらの発話された商品名に基づいて，トピックセグメンテーションを行う方法を採用する．つまり，商品名 A が最初に発話された区間から商品名 B を最初に発話する直前までを商品名 A の紹介映像区間であるとする．ここで，この方法では，システムが事前に商品名の辞書を持っていることが前提となる．

3　キーワードスポッティングによるトピックセグメンテーション

3.1　キーワードスポッティング

キーワードスポッティングは，サブワードデコーダーを併用して，リジェクションを行うことのできる方法を採用する [12]-[14]．具体的には，図 1に示す (a), (b)2 種類の言語モデルを用いる．(a) はキーワードとフィラーモデルにより構成されており，キーワードモデルと呼ぶ．フィラーモデルは，キーワード以外の区間を近似するもので，ここでは任意の音素を用いている．(b) は音素モデルの出現のみを許したもので，ここではバックグラウンドモデルと呼ぶ．(b) を用いたデコーディングは，サブワードデコーダーに相当する．入力音声に対して，2 種類の言語モデルを利用してそれぞれビタビアルゴリズムに基づき，式 (1), (2) で示される対数尤度を算出する．



(a)Keyword model　　(b)Background model

図 1: キーワードスポッティング用言語モデル

$$S_{KW}(W) = \frac{\text{単語}W\text{が認識されたときの対数尤度}}{\text{入力フレーム数}} \quad (1)$$

$$S_{BG}(W) = \frac{\text{単語}W\text{に対応する音素列の対数尤度}}{\text{入力フレーム数}} \quad (2)$$

ここで $S_{KW}(W)$ は，キーワードモデル中の単語 W が認識されたときの単語の対数尤度を，単語のフレーム数で割ったものであり，$S_{BG}(W)$ は $S_{KW}(W)$ で検出した単語 W の区間に対して，バックグラウンドモデルで音

素列の対数尤度を計算し，単語Wに対応するフレーム数で割ったものである．図2に単語Wが認識されたときの各尤度とフレーム数の対応を示す．

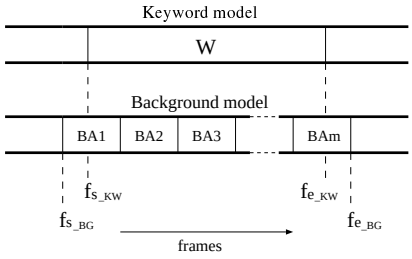


図 2: 2つのモデルでの尤度区間の対応

湧き出し単語のリジェクションは，式(3)により求めた対数尤度差$S(W)$が，ある一定の閾値以下のものに対して行なう．

$$\text{対数尤度差 } S(W) \quad = \quad S_{KW}(W) - S_{BG}(W) \quad (3)$$

以上に述べたキーワードスポッティングを，商品紹介映像の音声に対して適用することにより，商品名（キーワード）を抽出する．

ここで，商品紹介映像の音声は，常に音楽が重畳しており，この音楽により，キーワードスポッティングの精度が低下するという問題が生じる．この問題を解決するために，本研究では，商品紹介映像の音声に対して雑音除去[11]を適用した後に，キーワードスポッティングを行っている．

3.2 実験

社内で製作された商品映像に対して，キーワードスポッティングに基づくトピックセグメンテーションを行った．

3.2.1 実験条件

音響モデルには，話者独立なmonophone HMMを用いた．HMMの学習には，日本音響学会新聞記事読み上げ音声コーパスのうち，男性話者137人分の21782発話を用いており，それぞれのデータに対してCMNを行っている．音響分析の条件，HMMの構造を表1, 2に示す．

表 1: 音響分析条件

標本化周波数	16kHz
高域強調	$1 - 0.97z^{-1}$
特徴パラメータ	12次MFCC + log Power $+\Delta + \Delta\Delta$
分析区間長	20ms
分析周期	10ms
時間窓	Hamming Window

表 2: 音素HMMの構造

状態数	5状態3ループ
混合数	12
音素数	41
タイプ	Left-to-Right HMM

3.2.2 実験結果

キーワードスポッティングの実験結果を表3に示す．単語正解率($Corr$)は式(4)，単語正解精度(Acc)は式(5)にそれぞれ示される．

表 3: キーワードスポッティングの結果(%)

データタイプ(id)	$Corr$	Acc
商品映像(003)	91.78	75.34

$$Corr(\%) \quad = \quad \frac{N - S - D}{N} \times 100 \quad (4)$$

$$Acc(\%) \quad = \quad \frac{N - S - D - I}{N} \times 100 \quad (5)$$

S： 置換誤り単語数　D： 脱落誤り単語数
I： 挿入誤り単語数　N： 全単語数

上記のキーワードスポッティング結果を用いた商品分割の実験結果を表4に示す．再現率は式(6)，適合率は式(7)，Fmeasureは式(8)にそれぞれ示される．再現率は商品区間の正確さを，適合率は過分割の少なさを表す．再現率と適合率はトレード·オフの関係にあるので，Fmeasureで実験結果の評価を行った．評価の結果，約82%の精度で連続した商品映像を各商品区間へ分割することができた．図3に分割した各商品区間の先頭フレームを示す．ユーザは，図3の商品画像をクリックすることで，その商品の紹介映像を視聴することができる．

表 4: 分割実験の結果(%)

データタイプ(id)	再現率	適合率	Fmeasure
商品映像(003)	82.4(14/17)	82.4(14/17)	82.4

$$\text{再現率} \quad = \quad \frac{\text{正しく検出された商品区間境界数}}{\text{人手で検出された商品区間境界数}} \quad (6)$$

$$\text{適合率} \quad = \quad \frac{\text{正しく検出された商品区間境界数}}{\text{システムが検出した商品区間境界数}} \quad (7)$$

$$Fmeasure \quad = \quad \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \quad (8)$$

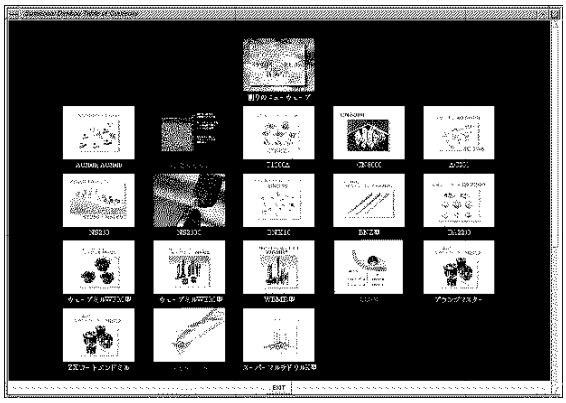


図 3: 商品区間への分割例

4　商品名辞書の自動作成

3にて，キーワードスポッティングを用いることにより，高い精度でキーワードを抽出し，商品区間に分割できることを示した．しかし，3の手法は，システムが事前に商品名の辞書を持っていることが前提となっており，事前に商品名の辞書が用意できない場合は適用することができない．そこで本研究では，商品名の辞書が存在しない場合に，辞書を自動で作成し，商品区間に分割する手法についても検討を行った．

4.1　テロップ文字認識による商品名の抽出

音声認識全般において，単語辞書が存在しない場合，音声認識の結果は音素列の羅列でしかなく，この音素列の羅列からキーワードを抽出することは非常に困難である．そこで本研究では，映像中に含まれるテロップ文字に注目した．商品紹介映像おいて，多くの場合図4に示した例のように，各商品映像の先頭フレーム画像には商品名を表すテロップ文字が存在する．このテロップ文字に対して文字認識を適用し，商品名を抽出することにより，商品名の辞書を自動生成することが可能であると考えられる [15] .

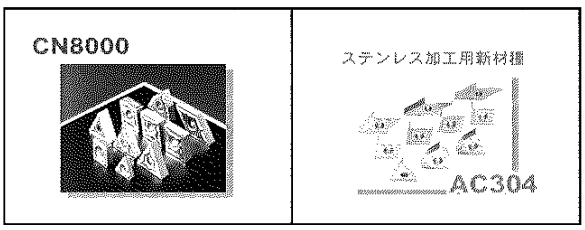


図 4: 各商品映像の先頭フレーム画像の例

4.2　音響パワーによる商品区間の始点検出

テロップ文字から商品名を抽出するためには，4.1で述べたように，各商品映像の先頭フレーム画像を検出して，文字認識を行う必要がある．しかし，各商品映像の先頭フレーム画像を検出することは難しく，従来のテロップフレーム検出手法では，多くのわきだしおよび，検出もれが発生してしまう．そこで本研究では，商品紹介映像の音響的な特徴を用いて，各商品映像の先頭フレーム画像を検出することを試みた．

商品紹介映像の音声には多くの場合，背景にBGMが存在している．このBGMは図5に示すように，商品区間の始点から演奏が始まり，商品区間の終点で終了し，BGM終了から次の商品の紹介までの間には無音区間が存在する．本研究では，音響パワー（対数パワー）を用いてこの無音区間を検出し，無音区間が終了した時刻を次の商品区間の始点と見なし，始点の時刻に同期するフレーム画像を，商品映像の先頭フレーム画像として抽出した．

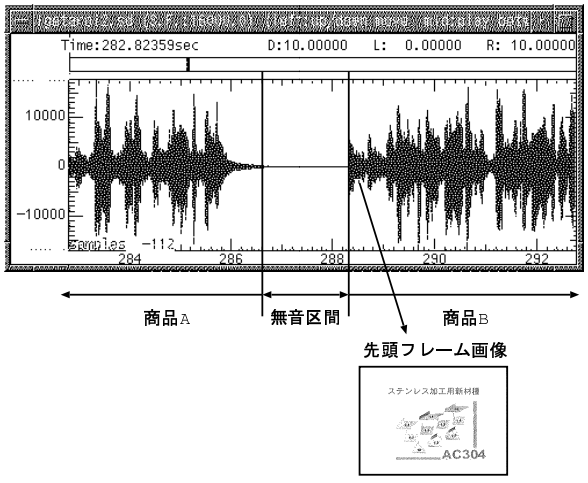


図 5: 音響パワーによる商品区間の始点検出の例

4.3　文字認識結果からのキーワード抽出

4.2で述べた手法により検出された，商品映像の先頭フレーム画像に対して，テロップ文字の認識を行った．テロップ文字の認識は，フレーム画像からテロップ文字領域を抽出した後に，市販のOCR（Win Reader Pro Ver.5.0）を用いて行っている [15]．ここで，図6に示すように，文字認識を行うと，'AC304'などのフレーム画像に存在する文字列の他に，'訓E櫨葎藍ｊ'といったような，自然語として意味不明な文字列がわきだすという問題がある．この問題を解決するために，文字認識結果に対して形態素解析（すもも Ver.1.3(NTT)[16]）を行い，自然言語として意味を持った文字列のみをキーワードと

して抽出した．文字認識結果に対して形態素解析を行うと，’訓Ｅ櫨葎藍ｊ’といったような，自然言語として意味不明な文字列は，1文字もしくは2文字単位の単語に分解される．一方，’ステンレス’という形態素解析の辞書に存在する文字列や，’AC304’という記号的な意味合いをもつ文字列は，一つのまとまった単語として解析される．この特徴を利用して，形態素解析の結果において，2文字以下の単語を削除し，3文字以上の単語をキーワードとして利用した．

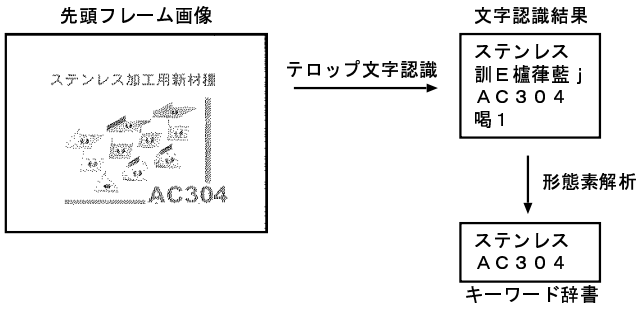


図 6: 文字認識結果からのキーワード抽出の例

4.4 自動生成された商品名辞書を用いたキーワードスポッティング

4.3で述べた商品名辞書の自動生成は，4.2で検出された区間ごとに行い，それぞれの区間で生成された商品名辞書を用いてキーワードスポッティングを行う．ここで，4.3の手法で生成された商品名辞書には，図6の例のように複数のキーワードが存在する場合があり，どのキーワードが商品名であるかを判定する必要がある．この問題において，商品名はその商品の紹介区間において，最も重要なキーワードであり，繰り返し発話される回数が最も多いキーワードであると考えられる．このことから，自動生成された辞書を用いてキーワードの抽出を行い，最も多く抽出されたキーワードを商品名であると見なした．

また，生成された辞書にキーワードが存在しない，もしくは，キーワードスポッティングによりキーワードが抽出されなかった区間については，4.2の手法により，過剰検出された区間であり，先頭フレーム画像には，意味を持ったテロップ文字が存在しないとみなして，1つ前の商品区間とマージしている．

5 実験結果

4の手法により，自動生成された商品名辞書を用いて，各商品区間への分割と，分割された区間に対して商品名による索引付与を行った．実験条件は3.2の条件に従っており，実験結果を表5に示す．

表 5: 分割実験の結果(%)

データタイプ(id)	再現率	適合率	Fmeasure
商品映像(003)	58.8(10/17)	62.5(10/16)	60.6

表5より，自動生成された商品名辞書を用いることにより，Fmeasureで60.6%の精度で個々の商品区間に分割し，索引づけを行うことができた．

分割精度が60.6%と低くなった理由として，テロップ文字認識精度の問題が考えられる．本研究で用いた商品紹介映像に用いられているテロップ文字は，青やピンクなどの色付きの文字が多く存在している．このような色付きのテロップ文字は，文字自身の輝度値が高く，背景画像との輝度差分が小さくなるため，文字領域の切り出し精度が低下し，テロップ文字の取りこぼしが生じやすくなる．これにより，テロップ文字からのキーワード抽出精度が低下し，商品名辞書の精度に影響したと考えられる．また，商品名辞書の精度の低さが，分割精度に影響したと考えられる．この問題を解決するために今後，色付きのテロップ文字を高精度に切り出し，認識を行う手法について検討を行う必要がある．

また，テロップ文字の情報だけでなく，紹介映像の音声からもなんらかの情報を抽出して，画像，音声それぞれの情報を統合し，十分に活用した上で，より精度の高い商品名辞書を生成することについても検討を行う必要がある．

6 おわりに

本研究では，連続した商品紹介映像をキーワードスポッティングにより，各商品区間へ分割する方法について検討を行った．実験の結果，商品名辞書が存在する場合において，約82%の精度が得られた．一方，商品名辞書が存在しない場合は，個々の商品区間の先頭フレーム画像のテロップ文字を認識することにより，商品名辞書を自動生成し，約60%の精度を得ることができた．これら2つの実験から，商品区間分割の精度は，商品名辞書の精度に依存すると言える．このことをふまえて，今後，商品名辞書を精度よく自動生成する手法について検討する予定である．

また，今回使用した商品紹介映像の音声は，ニュースキャスターのように日本語文法及び，発音のしっかりとした音声であったため，高い音声認識精度及び，分割精度が得られた．しかし，音声がいわゆる自由発話音声のように，くだけた文法と怠けの生じた発音であった場合，音声認識精度，分割精度が低下すると考えられる．このことより，雑音だけでなく，自由発話音声に対しても頑健な音声認識手法が必要となる．さらに，商品紹介映像

にはニュース形式のもの，コミカルな形式なものなど，様々な形式が考えられ，それぞれの形式において様々な話者が登場することが考えられる．このことから，認識に用いる音響モデルを，様々な話者に対して適応させ，認識環境に応じて高精度な音響モデルを構成する必要がある．

以上の問題を解決し，様々な映像データに対して，高い精度で分割を行うことについても，今後検討する予定である．

参考文献

[1] S.Takao, J.Ogata and Y.Ariki: "Topic Segmentation of News Speech Using Word Similarity", *Proc. ACM'00*, pp.442-444(2000).

[2] P.Mulbregt, I.Carp, L.Gillick, S.Lowe and J.Yamron: "Text Segmentation and Topic Tracking on Broadcast News Via A Hidden Markov Model Approach", *Proc. ICSLP'98*, Volume VI, pp.2519-2522(1998).

[3] J.Ponte and W.Croft: "Text Segmentation by Topic", First European Conference on Research and Advanced Technology for Digital Libraries(1997).

[4] S.Takao, J.Ogata, and Y.Ariki, "Study on New Term Weighting Method and New Vector Space Model Based on Word Space in Spoken Document Retrieval", *Proc. RIAO'00*, Volume I, pp.116-131(2000).

[5] M.Siegler : "Experiments in Spoken Document Retrieval at CMU", *TREC7*(1998).

[6] 鷹尾誠一, 緒方淳, 有木康雄: "ニュース音声に対するパッセージ検索法の比較", 日本音響学会, 平成12年度秋季研究発表会発表会, 2-Q-5, pp.139-140(2000).

[7] J.Xu and W.Croft: "Query Expansion Using Local and Global Document Analysis", In Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp.4-11(1996).

[8] S.Takao, J.Ogata and Y.Ariki: "Expanded Vector Space Model based on Word Space in Cross Media Retrieval of News Speech Data", *Proc. ICSLP'00*, Volume II, pp.1085-1088(2000).

[9] 鷹尾誠一, 有木康雄, 緒方淳: "テロップやフリップ文字を検索質問とした発話文書に対する検索方式", 第6回知能情報メディアシンポジウム, pp.87-88(2000).

[10] 鷹尾誠一, 緒方淳, 有木康雄: "ニュース音声記事データベースにおける観点の自動抽出と構造化", 信学技報, DE00-12, pp.89-96(2000).

[11] M.Fujimoto, Y.Ariki: "Continuous Speech Recognition under Non-stationary Musical Environments Based on Speech State Transition Model", *CD-ROM Proc. ICASSP'01*(2001).

[12] K.M.Knill and S.J.Young: "Speaker Dependent Keyword Spotting for Accessing Stored Speech", CUED/F-INFENG/TR 193(1994).

[13] 緒方 淳，有木 康雄: "ニュース記事分類におけるディクテーションとワードスポッティングの比較", 信学技報, SP98-32, pp.67-72(1998).

[14] 中川 聖一: "音声認識研究の動向", 信学論, D-II, Vol, J83-D-II, No.2, pp.433-457(2000).

[15] 鷹尾誠一, 有木康雄, 松本宏: "テロップ文字認識による商品紹介映像の商品区間への分割方法", 電子情報通信学会総合大会, SD-5-6，pp.361-362(2001).

[16] "形態素解析システム「すもも」", http://www.t.onlab.ntt.co.jp/sumomo/index.html