映像コンテンツのインデクシングのための音声・言語処理

林 良彦，松尾義博，大附克年，池田成宏，松永昭一，林 実，水野 理，別所克人，長谷川隆明
日本電信電話株式会社 NTT サイバースペース研究所

１．はじめに

ブロードバンドネットワークの進展に伴い，音声，動画像を含むマルチメディア映像コンテンツをネットワーク上で流通させ，様々な形で利用しようという動きが活発化している．このような映像コンテンツ群を有効に利用しようとすれば，コンテンツからその内容に関する情報を抽出し，メタデータとして構造化したうえでインデックスとして保持しておくことが必要となる．本稿では，このような映像コンテンツのインデクシングシステムにおける音声・言語処理の概要について述べる．

２．内容記述メタデータの生成と利用

マルチメディア映像コンテンツに付与すべきメタデータの基準について MPEG-7[1]などの標準化が進んでいる．しかしながら，メタデータの生成・付与においては，特に内容記述に関連する部分について人手に負うところが大きく，その多大なコストが問題となっている．このため，映像処理や音声・言語処理などのメディア処理[2]によって，メタデータ生成の過程を効率化しようという試みが行われている．

本稿で提案するインデクシングシステムの構成の概要を図１に示す．本システムにおいては，ニュース番組のように，特定の話題に関する区間(トピック区間と呼ぶ)が複数集まって一つの番組となっているような映像コンテンツを対象としている．

コンテンツ中の音声トラックの部分からは，音声インデクシングと呼ぶ処理により，検索・アクセスに有用な情報が抽出される．ここで抽出される情報は基本的に言語情報であるため，言語表現をキーとする検索の実現において重要である．

一方，コンテンツ中の映像トラックの部分については，映像インデクシングと呼ぶ処理により，シーン分割やテロップ文字認識結果などの情報を抽出する．

Overview of the Speech and Language Processing for Indexing Multimedia Content. by Yoshihiko Hayashi, Yoshihiro Matsuo, Katsutoshi Ohtsuki, Naruhiro Ikeda, Shoichi Matsunaga, Minoru Hayashi, Osamu Mizuno, Katsuji Bessho, and Takaaki Hasegawa. NTT Cyberspace Laboratories, NTT Corporation.

両者から抽出された情報は，情報統合と呼ぶ処理により整合性のチェックなどを経て XML 形式で表現されたメタデータとして出力される．検索対象としたいコンテンツ群に対するメタデータ集合を XML 検索エンジン[3]によりインデックス化することにより，キーワード入力によって特定の話題に関するトピック区間をコンテンツ群の中から高速に検索し，その区間のみを生成させるような検索・アクセスサービスを実現することが可能となる．

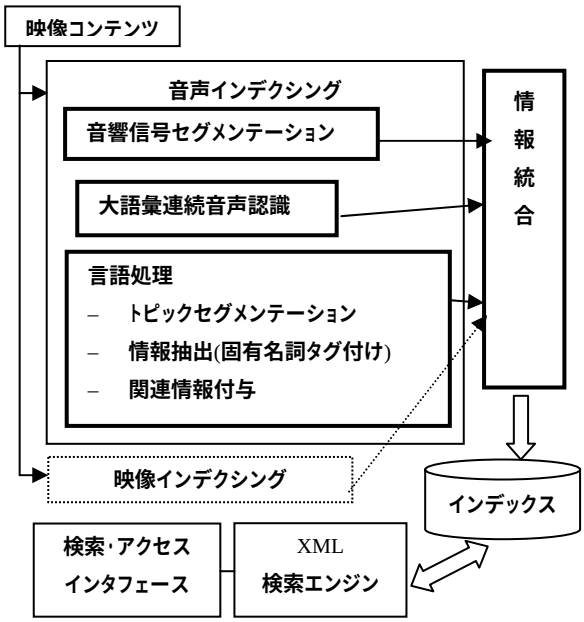

図 1: インデクシングシステムの構成概要

以下では，音声インデクシングに関わる部分について説明する．

３．大語彙連続音声認識による発声の文字化[6]

本システムで処理対象とするようなニュース番組のようなコンテンツにおいては，内容の主要な部分がアナウンサーなどにより発声される．このため，発声部分を音声認識により文字化することにより，内容記述の基本となる情報を得ることができる．音声認識の適用のためには，認識対象とすべき発声区間を求める必要がある．本システムでは，音響信号処理 [4]により，発声区間を BGM やノイズの区間から区別した後に，大語彙連続音声認識処理を施す．

用いる音声認識エンジン[5]は，適切に準備された音響・言語モデルのもとで，高速(実時間での認識)・高精度(ニュース番組で 95%程度の単語正解率)に発声内容を認識することができる．一定の長さのポーズで区切られた発声セグメントごとに認識結果の漢字かな混じり文字列，読み・品詞などの文法情報のほかに，認識信頼度を出力する．さらに，発話セグメントごとに，関連する話題ラベルを出力する．

４．トピック区間への分割

ニュース番組のようなコンテンツ群をキーワードにより検索する場合，各番組コンテンツ中に存在する指定されたキーワードの発話それぞれを検索結果として返すよりも，各番組中に含まれる各トピック区間を一つのまとまりとして検索の単位とするほうが適切であると考えられる．このように，トピック区間を検索文書とみなせば，通常の情報検索のように，ランキングされたトピック区間を検索結果として返すことが可能となる．

本システムでは，内容語に対してあらかじめコーパスにおける共起情報から獲得した概念ベクトルに基づく手法[7]を用いてトピックセグメンテーションを行う．この手法では，発話セグメントの認識結果の文字列を概念ベクトルの系列へと変換し，ある窓幅において，概念ベクトルの系列が大きく変化する発話セグメント境界をトピック区間の境界として抽出する．実験の結果，本手法はニュース番組のようなコンテンツに対して比較的良い精度を与えることが確認できたが，さらに，トピックの推移を示す手がかり語(「さて，次のニュースは…」など)を併用することにより，精度の向上を図っている．

５．検索・アクセスに有用な情報の抽出

本システムにおいてメタデータ化され，インデックスとして蓄積される情報は音声認識結果に依存している．認識結果には誤りが含まれることが避けられないため，認識誤りの影響を低減させる手段が望まれる．例えば，本来発声されていたであろう語彙を内容記述のメタデータに補完できるとよい．

このために，認識結果をクエリとして用いて外部データベースから関連文書を検索することにより，そのような語彙を補完する手段を実現している．基本的な考え方は，[8]と同様であるが，クエリとする単語を抽出する際に音声認識の信頼度を考慮することにより，適切な関連文書検索を行うようにしている．さらに精度を向上させるために，サイドコーパスにおける共起情報をもとにクエリとすべき重要語を選別することを試みている[9]．

また，認識誤りの原因の一つとしてボキャブラリの問題があり，特に固有名詞を含む固有表現の認識において問題となる．固有表現として認識されるべき区間を隠れマルコフモデルにより統計的に認定する手法を実現しており[10]，認識誤りを含む場合もある程度の精度で固有表現の区間を認定し，そのタイプを判定することができる．固有表現のタイプに応じたタグ付けをすることができれば，これを利用した構造化検索が可能となる．

６．おわりに

現在までのところ，ニュース番組のようなコンテンツに対して本システムが有効に動作することを確認している．これは，アナウンサーなどによる発声が良好な音声認識結果をもたらすこと，コンテンツ自体が比較的明確なトピック構造を持つこと，コンテンツの意味内容がインターネットなどの外部データベースから関連する語彙を補完することに適していることなどが理由である．

今後は，対象とするコンテンツの範囲を，例えばドキュメンタリ番組などに広げていくために，大語彙連続音声認識の適応性・ロバスト性の向上などを進めていく．また，検索インタフェースの向上を目的とした話題表現の抽出[11]や要約生成，認識誤りに対して頑健なトピックトラッキング手法などの検討を進めていく．

参考文献

[1] http://www.itscj.ipsj.or.jp/mpeg7/.

[2] 有木: マルチメディア情報の解析と統合，人工知能学会情報統合研究会，SIG-CII-2000-Nov,2000.

[3] 富田: XML 文書検索システム:LISTA，NTT R&D，Vol.52，No.2，2003.

[4] 水野，大附，松永，林: ニュースコンテンツにおける音響信号自動判別の検討，電子情報通信学会 2003 年総合大会，2003.

[5] 野田，山口，大附，小川，中川，今村: 音声認識エンジン VoiceRex の開発，日本音響学会 1999 年秋季研究発表会，1999.

[6] 大附，松永，別所，松尾，林: 大語彙連続音声認識を用いた音声・映像コンテンツのインデクシング，日本音響学会 2003 年春季研究発表会，2003.

[7] 別所，大附，松永，林: 概念ベクトルの結束性によるトピックセグメンテーション精度の評価，言語処理学会第 9 回年次大会，2003.

[8] Singhal,A. and Pereira,F.C.N.: Document Expansion for Speech Retrieval，Proc.of SIGIR-99，1999.

[9] 松尾，林: 認識誤りに頑健な重要語抽出，言語処理学会第 9 回年次大会，2003.

[10] 長谷川，林: 隠れマルコフモデルに基づく音声認識結果からの固有表現抽出，言語処理学会第 9 回年次大会，2003.

[11] 池田，松尾，林: パターンと重要語に基づく関連記事からの話題抽出，語処理学会第 9 回年次大会，2003.