話者情報を用いた映像検索システム
A Video Retrieval System based on Speeker Recognition

F-027

光宗 朋宏† 柘植 覚‡ 黒岩 眞吾‡ 獅々堀 正幹‡ 北 研二§
Tomohiro Mitsumune Satoru Tsuge Shingo Kuroiwa Masami Shishibori Kenji Kita

1. はじめに

インターネットのブロードバンド化や大容量メディアの登場により映像メディアの普及が進み，映像検索への期待が高まっている．映像メディアは画像，音声，言語などから成るデータストリームである．複数のデータストリームから成ることから，いずれかのデータストリームに着目することにより，見たい映像を検索することが可能である．本論文では，検索したい話者の音声情報に着目し，話者認識を用いて映像のインデキシングを行うシステムを提案する．

2. 映像インデキシング

映像は時間軸を持った連続メディアである [1]．従って，ある内容を見たい場合には，映像の始めから時間軸に沿って見ていくことになる．この不便さを無くすために，章や節のようなインデクスを付与する映像インデキシングといわれる手法がある．インデキシングすることにより，時間的な連続メディアである映像を，章や節ごとに取り出し利用することができる．本節では一般的な映像インデキシングの手法と本研究に用いる手法を述べる．

2.1 画像データによるインデキシング

映像では，画像が大きく変化する所がシーンの変わり目であることが多く，そこにインデクスをつける手法が提案されている．例として，連続した画像が変化する点を利用してインデキシングする手法 [2] や，映像の中で繰り返し現れる区間を検出しインデキシングする手法 [3] などがある．

2.2 音声データによるインデキシング

ニュース音声や討論番組の音声，TV 番組の音声などのあらゆるデータには必ず音声区間と無音区間が存在する．本論文では，この 2 つの区間を利用して話者認識によるインデキシングを行う．話者認識を用いる場合，話者のモデルをあらかじめ作成する教師ありインデキシングと，逐次的に作成する教師なしインデキシングの 2 通りに分類できる [4]．このうち，開発システムでは非リアルタイムのインデキシングを想定しており，教師ありインデキシングを用いている．

3. 音声情報を用いた映像検索システム

本システムでは，映像中の音声情報に基づきユーザの見たい人物が現れる映像シーンを検索する．実際の処理の流れを図 1 に示す．まず映像データから音声データを抽出し，有音・無音区間に分類する．次に，抽出された有音区間に対して話者識別を行い誰の発声であるかを判別する．システムの利用方法として web ブラウザ上で，映像の音声データをインデキシングし，特別なアプリケーションやソフトウェアを必要とすることなく検索できるように実装した．

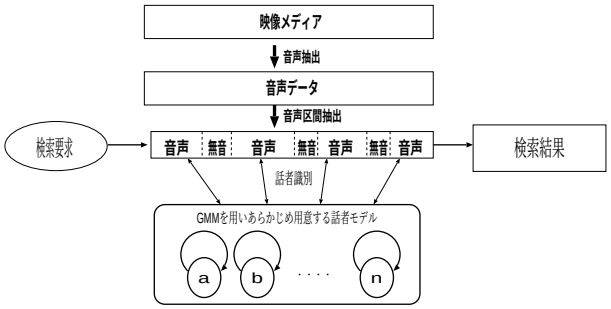

図 1: 映像検索システムの概要

3.1 音声区間検出を用いた自動セグメント

音声データには様々な種類があるが，必ず音声区間と無音区間が存在する． 音声区間と無音区間を区別するために，入力音声のフレームごと (フレーム長: $10ms$，フレーム周期 $5ms$) の平均パワーを求め，設定された閾値より高いパワーが一定フレーム以上続けば音声区間と判断し，それ以外は無音区間と判定する．そして連続した音声区間と判定された無音区間から無音区間の間の音声を一つの発話と考えインデキシングを行う．

3.2 話者認識

本システムではテキスト独立型話者識別を用いインデキシングを行う．テキスト独立型話者識別では，入力音声と登録されている話者の標準パターンとの類似度を調べて，最も類似度の高い話者を選びその話者が発話者であると判定する．話者モデルとして混合ガウス分布により音声データをモデル化する方法である GMM(Gaussian Mixture Model) を用いた [5]．音声データの n 次元特徴ベクトルを x_t，話者 s のモデルを λ_s とすると，話者モデル λ_s の GMM は，次のように表される．

$$b(x_t|\lambda_s) = \sum_{m=1}^{M} \omega_{sm} N(x_t|\mu_{sm}, \Sigma_{sm}) \qquad (1)$$

$N(x_t|\mu_{sm}, \Sigma_{sm})$ は多次元正規分布の密度関数であり，次式のように表される．

$$N(x_t|\mu_{sm}, \Sigma_{sm}) =$$

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp - \frac{1}{2}(x_t - \mu_{sm})^t \Sigma_{sm}^{-1}(x_t - \mu_{sm}) \qquad (2)$$

† 徳島大学 大学院 工学研究科
‡ 徳島大学 工学部
§ 徳島大学高度情報化基盤センター

表 1: 音響分析条件

サンプリング周波数	16kHz
フレーム周期	10msec
フレーム長	25msec(ハミング窓)
窓タイプ	ハミング窓
フィルタバンク数	24
エネルギー正規化	なし

表 2: 実行時間と識別率

先頭 n(秒)	実行時間 (秒)	識別率 (%)
5	35.90	90.9(70/77)
10	57.92	94.8(73/77)
15	79.76	97.4(75/77)
20	104.31	100.0(77/77)
全区間	557.20	100.0(77/77)

この GMM を映像メディアに含まれる全話者分および BGM について作成し識別を行う.

3.3 先頭数秒を用いた識別

音声区間全体を GMM により話者識別を行った場合, 識別に時間がかかってしまいシステムに用いるのには実用的ではない. そこで, セグメントされた音声区間の先頭の数秒のみを用い識別することを考えた. これにより実行時間の大幅な削減が期待できる.

4. インデキシング実験

4.1 実験方法

システムを評価するために, 市販されている DVD の映像データを使用し実験を行った. 同データ中では, 女性 10 名が BGM の流れている環境で, 各々2 分前後の発話を数回行っている (延べ 120 分). 各話者の発話の間に効果音, 歌などの音声も混在している. 音響分析条件を表 1 に示す. 特徴パラメータは MFCC1〜12 次, その一次回帰係数 12 次, 対数パワーの一次回帰係数を用いた. MFCC に対しては, 1 データごとに CMS(Cepstrum Mean Subtraction) を行った. 話者モデルは 16 混合の GMM を用いた.

1. 不特定話者 GMM および背景音 GMM の作成
   話者モデルを作成する際の初期モデルとして使用する不特定話者モデルを学習した. 学習データには女性 10 名の各々の発声区間をあらかじめ取り出したもの 60 秒と話者の発話以外の音 (効果音, 歌など) をランダムに選んでに編集したもの 90 秒を用いた. 学習のくり返し回数は 10 回とした.

2. 話者 GMM の作成
   不特定話者 GMM を初期モデルとして, 話者 GMM を学習した. 学習データは, 1. と同様である. 学習のくり返し回数は 10 回とした.

4.2 識別実験結果

自動セグメントされた音声区間の先頭 n 秒 (n=5, 10, 15, 20) を用い識別実験を行った結果を表 2 に示す. この実験は, CPU:Celeron1.2GHz, メモリ:256MB の PC で行った. 実行時間は, 自動的にセグメントを行い全音声区間を識別するまでの総時間である. なお, 本実験では複数話者の発声が 1 つの音声セグメントになったしまうことはなかったため, 識別率のみでの評価を行っている.

4.3 考察

表 2 より評価データの音声が短くなればなるほど話者の識別率は下がっていることがわかる. GMM を用いた話者識別は比較的短い評価音声長においても高い識別率が得られるが, 15 秒以下の音声では識別率が下がる. しかし, 音声の先頭 20 秒を用いることにより, 音声区間全てを用い識別したときと同様な識別率が得られ, このとき実行時間も 1/5 以下に下がっている. これより, 先頭 20 秒程度の音声を用いることがシステムに適当であると結論できる. n の値を下げた場合, ある話者の識別率が極端に低いのが目立った. そこで, 話者モデルを作成した音声の映像と, 誤識別された音声の映像を見くらべてみた. その結果, その話者では映像間で明らかな体型の変化が見られた (撮影時期に差があることが予測される). 時期差による性能の低下は話者認識における問題点の 1 つであるが, 映像情報においては複数の時期に撮影したものが一本の映像となっていることもあり, 新たな検討が必要である.

5. おわりに

本論文では, 映像データ中の音声情報に着目し, 話者識別を用いてインデキシングを行い検索するシステムを構築し, その動作を検証した.

6. 謝辞

本研究の一部は, 財団法人放送文化基金および文部科学省科学研究費, 基盤研究 (B)(2) 14350204) の援助によった.

参考文献

[1] 馬場口 登: メディア理解による映像メディアの構造化, 電子情報通信学会パターン認識メディア理解研究会 特別講演 PRMU99-42, (1999)

[2] 金子 敏充, 堀 修:ゆう度比検定を用いた MPEG ビットストリームからの動画像カット検索手法電子情報通信学会論文誌 D-II, Vol.J82-D2 No.3, pp.361-370, (1999)

[3] 多田 秀玲, 杉山 喜明, 有木 康雄：ニュース映像中の共通区間検出による記事クラスタリングと要約・検索画像電子学会研究会, 98-05-06, pp.33-36, (1998)

[4] 西田 昌史, 有木 康雄： 自動学習における話者セグメンテーション, 電子情報通信学会論文誌, SP97-57, pp.1-6, (1997)

[5] D.A. Reynolds and R.C. Rose： “Robust textindependent speaker identification using Gaussian mixture speaker models” IEEE Trans. on Speech and Audio Processing, Vol.3, No.1, pp.72-83, (1995)