映像データベースの音声情報による話者索引付けの自動化について*

4U－8

張 宏斌　中村 哲　植村俊亮†
奈良先端科学技術大学院大学 情報科学研究科‡
{kouhin-c, nakamura, uemura } @is.aist-nara.ac.jp

1. はじめに

映像データベース構築における重要な作業の一つは，音声や画像による索引付けである．従来，このような索引付けは手作業に頼ることが多かったが，音声，画像認識の発達した技術を，映像データベース独自の要請に配慮しつつ，索引付けに応用することが可能になりつつある．ここでは，GMM (Gaussian Mixture Model) を用いた話者認識を利用し，従来の話者認識のように予め話者モデルを生成しておくのではなくて，話者モデルを生成しつつ話者認識を行ない，話者の索引付けを自動的に行う手法について論じ，実験結果を報告する．

2. 映像の音声による話者索引付けにおける話者認識

2.1 話者認識の在来手法の問題点と解決案

従来の話者認識システムを利用した，映像データベースの話者索引付けには，次の問題点がある．(1) 話者モデルの事前登録は繁雑で手間のかかる手作業になる．(2) 映像の音声は，話者ごとに分割されていない．複数の話者が登場する映像に対して，一人の話者の音声の範囲といった情報もない．ここでは，複数話者の連続音声を自動的に分割し，自動的に新しい話者を識別し，話者モデルをトレーニングしてシステムに登録しながら，話者認識を行うことにより，音声索引付けを実現する手法を提案する．

2.2 複数話者を含む連続音声の自動分割

一般に，複数話者の連続音声には話者が交替するとき，若干の無言時間が生じる．本研究では，このような無言時間を利用して，複数話者の連続音声を自動分割する．

2.3 話者認識および話者モデルの自動生成

(1) 話者照合

話者認識率をあげるために，本研究では，動的しきい値 θ，ϕ を利用する．

$$\begin{cases} \mathcal{L}_s > \theta & \text{認識結果を受理} \\ \mathcal{L}_s < \phi & \text{話者を追加} \\ \phi \le \mathcal{L}_s \le \theta & \text{話者不明リストに登録} \end{cases}$$

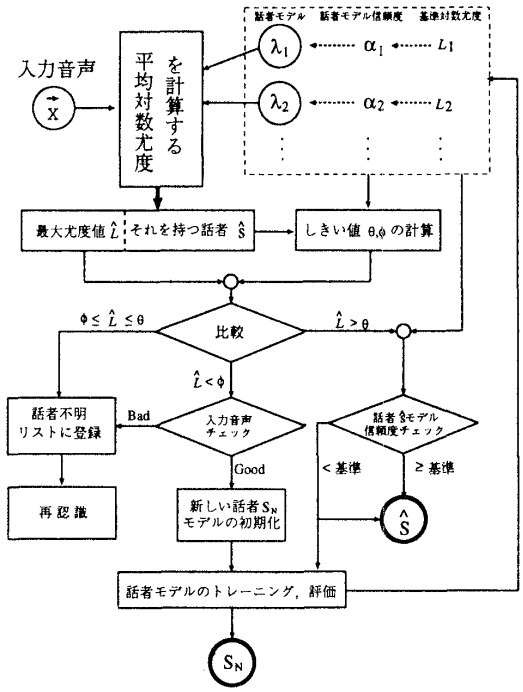

図 1 話者認識

θ，ϕ は$\theta_s = L_s \times \alpha_s \times \beta$, $\phi_s = L_s \times \alpha_s \times \gamma$である．

ここで L_sは，話者モデルλ_sの学習データを用いて，話者モデルλ_sに対して得られる平均対数尤度である．L_sを話者モデルλ_s の基準尤度とする．α_sは話者モデルλ_sの信頼度である（式 (1) ）．

$$\alpha_s = \frac{\text{話者モデル}\lambda_s\text{の学習データの長さ}}{\text{学習データの基準長さ}} \qquad (1)$$

β，γ (>1) は話者認識の誤り率をさげるための制御係数である．

(2) 話者モデルの自動生成

入力音声は，新しい話者の音声であると判定されると，それを学習データとして利用し，話者モデルを生成する．まず，システムは入力音声の発話内容についてチェックを行う．これは話者モデルの信頼度をあげるためである．本研究では，入力音声の中の発話時間を判定基準にした．発話時間が十分長ければ（例えば，$> 15s$），システムは自動的に新しい話者モデルを初期化し，入力音声を学習データに用いて，新しい話者モデルをトレーニングし，話者モデルを生成する．

(3) 話者モデルのトレーニング

---

*"Automatic Video Indexing by Speaker Recognition for a Digital Video Library"

†Zhang Hong Bin, S. Nakamura and S. Uemura
‡Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)

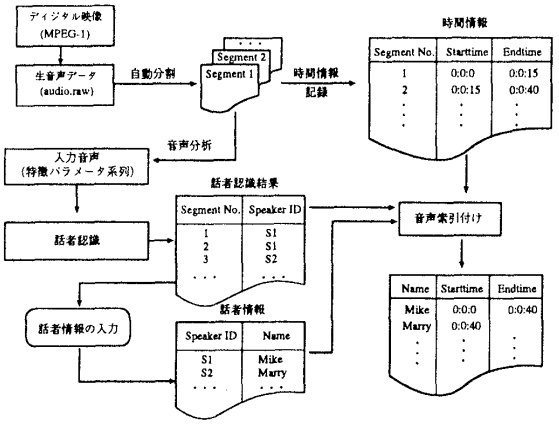

図 2 映像の音声による話者索引付け

(1) で入力音声に対して話者 s として認識されたとき，システムは話者 s のモデルを信頼度をチェックする．話者モデルの信頼度が 100%に達していない場合，システムは入力音声を用いて，話者モデルをトレーニングする．同時に，話者モデルの基準尤度と信頼度を更新する．

2.4 音声による話者索引付け

本システムは以下のステップで映像の音声による話者の索引付けを自動的に行う（図 2 ）．

(1) ディジタル映像（例えば，1 時間の会議ビデオ）から生音声データを抽出する．
(2) 生音声データを自動分割し，個々の音声セグメントの時間情報を記録する．
(3) 個々の音声セグメントの音声分析を行ない，特徴パラメタ系列を得る．
(4) (3) で得られた特徴パラメタ系列を入力として，話者認識を行う．
(5) 話者認識の結果は，仮の話者識別情報（ID）を用いて整理記録する．
(6) 話者ごとのサンプル音声あるいは映像をユーザに提示し，話者の名前の情報を得る．
(7) (2) の時間情報と,(5) の話者認識結果と,(6) の話者名の情報から音声の索引付けを行う．

(1) から (4) は，これまでに検討してきた話者認識の部分である．この段階では，話者を識別できても，それが誰であるかまではわからない．実際，個別の話者が誰であるかは，映像内容を熟知した人でないと識別できない．そこで，(5) における話者認識の結果は，セグメント番号と話者の識別番号の対応表になる．ある識別番号の話者が具体的にだれであるかを認識する作業は，映像を点検するなり，音声を聞くなりして，(6) で行い，話者番号と話者名の対応表を作る．こうした作業を結合して，最後に必要な話者索引を生成する．

3. システムの実装と評価

映像の題材には，衛星会議システムによる重点領域研究「高度データベース」の研究集会のビデオを MPEG によって圧縮して使用した．システムが自動生成した音声の索引を評価するために，索引の正確率を用いる．索引正確率は，式 (2) のように定義する．

$$\text{索引正確率} = \frac{\text{正しく索引付けされた映像時間}}{\text{映像の全時間}} \times 100 \quad (2)$$

実験結果を表 1 に示す．今回，15 分の映像を 8 個，30 分の映像を 4 個，60 分の映像を 1 個と 90 分の映像を 1 個を用いて，実験を行なった．表 1 に示す索引正確率と処理時間は平均正確率と平均処理時間である．実験には SGI Origin を使用したが，複数のユーザが使用しているので，ここで示す処理時間はおよその目安である．

映像時間	話者数	索引正確率	自動分割	話者認識
15 分	4	96.28%	数秒	約 30 分
30 分	5	95.13%	数十秒	約 50 分
60 分	8	90.33%	数十秒	約 110 分
90 分	10	88.12%	数十秒	約 140 分

表 1 正確率および処理の所要時間

4. まとめ

このシステムでは，複数の話者を含む連続映像を入力として，事前話者モデルの登録を必要とせずに，話者を識別しながら，話者モデルのトレーニングを行い，話者の識別情報を用いて，自動的に索引付けする．最後に，ユーザが話者名前情報を入力して，話者名前を用いた索引付けを行う．従って，本方式では，話者の名前情報の付与が人間の手作業に頼るだけで，あとは索引付けが自動化されている．

今回の実験では，MPEG による圧縮映像から抽出した音声を使用しているが，高い話者認識率を得られることが分かった．しかし，参加局のマイクの状況や送受信状況によって，映像会議の映像には，残響や反射音や環境騒音などがたくさん含まれている．これらの話者認識性能を劣化させる要因は，索引正確率に影響を与えることがある，システムの耐環境処理は今後の課題である．単一話者の話者モデルから話者混合モデルを生成手法や話者混合モデルを用いた話者認識の実現も今後の課題である．

参 考 文 献

[HAG95] A. G.Hauptmann: “Speech Recognition in the Informedia Digital Video Library:Uses and Limitations”, ICTAI-95,Nov 6-8,1995

[RGA95] D. A. Reynolds and R. C. Rose: “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, *IEEE Trans. on Speech and Audio Processing* Vol 3, No. 1. January 1995

[HAG97] A. G. Hauptmann and H. D. Wactlar: “Indexing and Search of Multimodal Information”, ICASSP-97 , April 1997, http://informedia. cs. cmu. edu/html/enter.html