3J－4

ニュース音声中のアナウンサー発話区間の自動切り出し

西田 昌史　有木 康雄

龍谷大学 理工学部

1 はじめに

ビデオ・オン・デマンドを目指したニュースデータベースの構築では、ニュースを記事毎に切り出しておく必要がある[1]。しかし、大量のデータから人手でニュース記事を切り出すには、かなりの労力を費やさなければならない。ニュース記事の切り出しの1つの方法として、アナウンサーの発話区間の抽出が有効であると考えられる。

本研究では、入力音声中の話者特徴を自動的に学習し、話者の発話区間を自動的に切り出すことを目的として、話者照合に基づいた話者区間の切り出し法を提案する。この手法を用いて、NHKニュース30日分に対して、アナウンサー発話区間の切り出し実験を行なった。

2 アナウンサーの発話区間の切り出し

2.1 話者照合

話者照合[2]とは、入力音声と同時に自分が誰であるかのIDを入力して、その音声が本当にそのIDに対応する人の発話であるかを判定するものである。図1に示すように、本人の標準パターンとの類似度（尤度）が、閾値よりも大きければ本人の発話であると判定し、それ以外の場合は他人の発話であると判定するものである。

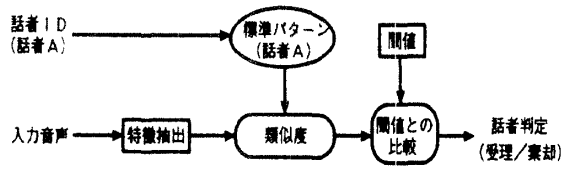

図 1: 話者照合

Automatic Segmentation of Announcer Utterance in News Speech
Masafumi Nishida and Yasuo Ariki
Ryukoku University
1-5, Yokotani, Oe-cho, Sete, Otsu-shi, 520-21 Japan

2.2 主成分分析法

本研究では、従来法に比べて計算量の少ない主成分分析法に基づく部分空間法[3]を話者照合のベースにしている。図2に主成分分析法の概念図を示す。

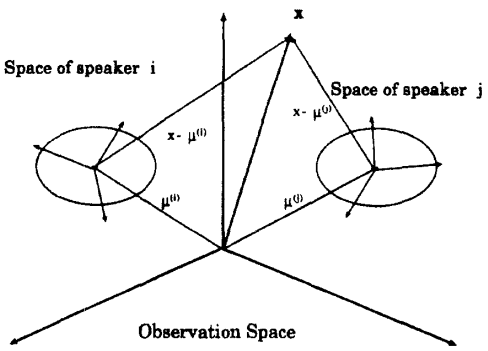

図 2: 主成分分析法

主成分分析法による話者照合では、話者ごとに入力音声データ $x_k(1 \leq k \leq N)$ から平均ベクトル $\mu^{(i)}$ を求め、式(1)により共分散行列 $R^{(i)}$ を求める。ここで、(i) は話者の識別番号を表す。

$$R^{(i)} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu^{(i)})(x_k - \mu^{(i)})^T \qquad (1)$$

話者ごとの共分散行列 $R^{(i)}$ を固有値分解して正規直交ベクトル $v_j^{(i)}, j = 1,2,\cdots,p^{(i)}$ を求め、式(2)により射影行列 $P^{(i)}$ を求める。

$$P^{(i)} = \sum_{j=1}^{p^{(i)}} v_j^{(i)} v_j^{(i)T} \qquad (2)$$

本人であると申告された話者の部分空間に対して、入力音声データ x_k との距離を式(3)により求める。この距離が閾値より小さければ、本人の音声であると判定する。

$$\begin{aligned} y^{(i)} &= \frac{1}{N}\sum_k \|x_k - \{P^{(i)}(x_k - \mu^{(i)}) + \mu^{(i)}\}\|^2 \\ &= \frac{1}{N}\sum_k \|(I - P^{(i)})(x_k - \mu^{(i)})\|^2 \end{aligned} \qquad (3)$$

2.3 話者区間の切り出し法

話者区間の切り出し方法を図3に示す。

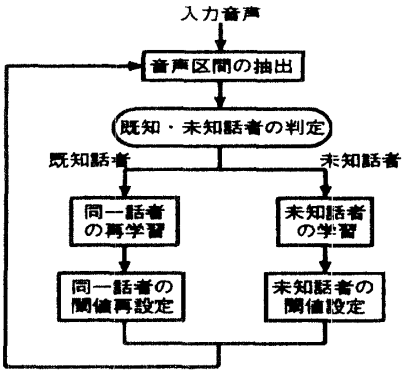

図 3: 話者区間の切り出し法

話者区間の切り出し方法は、まず入力音声の1秒の区間において平均パワーを求め、音声か無音かを判定し、これを1秒毎に繰り返す。これにより抽出された無音から無音までの音声区間を1話者の発話区間として、直前に発話した話者と照合する。既に学習した話者と判断されれば、今抽出された音声区間と既に抽出されている同一話者の音声区間を合わせて部分空間を再学習する。新しい話者と判断されれば、抽出された音声区間で部分空間を学習する。

既知話者と未知話者を判定する閾値 θ は、学習時に作成された部分空間と学習に用いた音声との距離の平均 μ と標準偏差 σ を求めて、次式のように設定している。

$$\theta = \mu + \frac{\sigma}{3} \tag{4}$$

3 アナウンサーの発話区間切り出し実験

3.1 実験条件

NHK の5分間のニュース30日分を用いて、アナウンサーの発話区間を切り出す実験を行なった。部分空間の次元数は、最も切り出し率が高かった7次元とした。

実験の評価は、アナウンサー推定率・切り出し率・適合率で評価した。これらは、次式で定義される。

$$\text{推定率} = \frac{\text{正しくアナウンサーを推定できた日数}}{\text{30 日}} \tag{5}$$

$$\text{切り出し率} = \frac{\text{正しく認識した発話区間数}}{\text{全ニュースの発話区間数}} \tag{6}$$

$$\text{適合率} = \frac{\text{正しく認識した発話区間数}}{\text{切り出した発話区間数}} \tag{7}$$

3.2 実験結果と考察

アナウンサーの発話区間切り出し実験を行なった結果を表1に示す。

表 1: アナウンサーの発話区間切り出し結果 (%)

アナウンサー		
推定率	切り出し率	適合率
76.7	93.4	98.7

発話区間の切り出しでは、話者の初期学習に用いる音声が短かすぎると、閾値が低く設定され、以後その話者の音声は棄却されやすくなる。また、発話区間に雑音が重畳していると、誤って話者を認識する場合がある。

本研究では、アナウンサーの発話の後にレポータなどの発話が続き、再びアナウンサーの発話に戻る直前までを一つの記事と判定している。このため、1つのニュースに対して話者の発話区間を切り出した結果、アナウンサーの発話区間が誤って切り出されると、記事を正確に切り出せなくなるといった問題がある。このため、切り出し率が高いにもかかわらず推定率が低くなっている。

4 おわりに

主成分分析法に基づく部分空間法をベースとする話者照合に基づいて、話者区間を自動的に切り出す方法を提案し、アナウンサーの発話区間の切り出し実験を行なった。

今後の課題として、発話区間の切り出し精度を向上させる予定である。

参考文献

[1] 齊藤陽子, 有木康雄 : “ニュース映像のデータベース化に向けて −ニューススタジオの映像検出と記事切り出し−”, 画像電子学会研究会, 149, pp.13-16 (1994).

[2] 松井知子, 古井貞熙: “VQ ひずみ、離散 / 連続 HMM によるテキスト独立型話者認識法の比較検討”, 信学論 (A), J77-A, 4, pp.601-606, (1994).

[3] エルッキ・オヤ, 小川英光, 佐藤誠訳: “パターン認識と部分空間法”, 産業図書 (1986).