VQ を用いた話者識別の検討*

侯　心 王月, 西　隆司（北九大）

1. はじめに

話者識別は個人情報を含み、コミュニケーションでもっとも使う「音声」に基づいて、だれが話しているかを自動的に判定するプロセスである。本研究では、五つの母音音素を連続して発音する「アオイウエ」を用い、MFCC 係数と Δ ケプストラム係数を特徴パラメータとして、ベクトル量子化 VQ(VQ：Vector Quantization) 法による話者識別システムを提案する。この手法により、高精度な本人識別が得られることを示す。

2. 話者認識の基本構造

全ての話者認識システムは話者識別と話者照合に分類することができる。話者識別は、入力された音声が登録話者中の誰であるかを判定する。話者照合は、音声を入力するとともに、自分が誰であるかを申告し、本人照合を行う。本研究では、音声による話者識別について検討した。

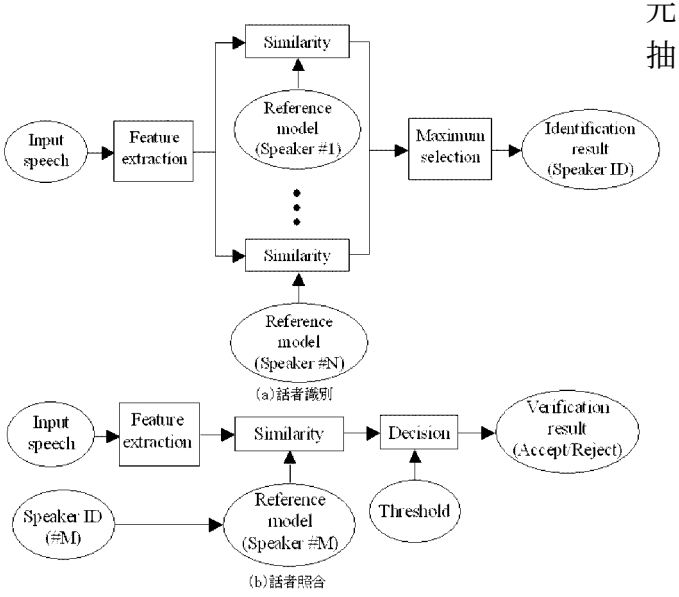

図 1: 話者認識システムの基本構造 [3]

3. 話者識別の特徴パラメータ

音声認識では、特徴量として人間の聴覚特性を考慮したメルケプストラム (MFCC：Mel Frequency Cepstrum Coefficient) と Δ ケプストラムが広く用いられている。本研究でも、この 2 種類を特徴量として採用した。

3.1 MFCC 係数

メルケプストラムは、音声波のスペクトルを人の聴覚に近い周波数間隔に切り分けてケプストラム化したものである。人の聴覚は低い周波数では細かく、高い周波数では粗い周波数分解能を持つことが知られている。これはメル (mel) 尺度と呼ばれ、対数に近い非線形の特性を示す。音声を認識するためには、音声スペクトルから周波数成分ごとの時系列データを抽出する必要があるが、人の聴覚に合わせるため、各帯域フィルタを対数周波数軸上、あるいはメルスケール上に等間隔に配置して抽出する。FFT によるスペクトルを元に、メルスケールの帯域フィルタ群出力を抽出する手順を図 2 に示す。

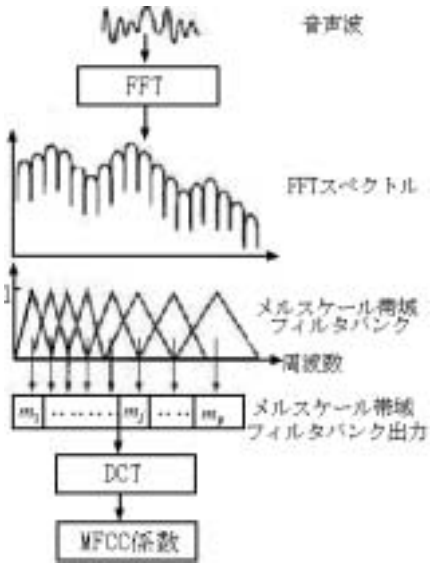

図 2: メルケプストラム抽出手順

*A Study on Speaker Recognition System with Vector Quantization
By Hou Xinyue , Nishi Takashi(The University of Kitakyushu)

ここで、メルスケールは

$$Mel(f) = 2595log_{10}(1 + \frac{f}{700}) \qquad (1)$$

で定義される。今、各帯域フィルタの出力を m_i とする。このとき、MFCC 係数は、DCT(discrete cosine transform) を用いて、

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j cos(\frac{\pi * i}{N}(j - 0.5)) \qquad i = 1, 2 \cdots, 40 \qquad (2)$$

で計算される [2]。

3.2 Δ ケプストラム

MFCC 係数はある分析フレームにおけるスペクトル包絡を表している。音声認識では、このほかにスペクトル包絡の時間的変換に対応し、動的特徴と呼ばれるパラメータが用いられる。これは Δ ケプストラム係数と呼ぶ。MFCC 係数の第 i フレームにおける n 番目の値を、$c_i(n)$ と記す。このとき、時刻 n を中心とした区間 $[n - \delta, n + \delta]$ における $c_i(n)$ の値に、直線を当てはめた場合の直線の傾きを $\Delta c_i(n)$ で表すと、

$$\Delta c_i(n) = \frac{\sum_{k=-\delta}^{\delta} k \cdot c_i(n+k)}{\sum_{k=-\delta}^{\delta} k^2} \qquad (3)$$

が成り立つ。Δ ケプストラム係数 $\Delta c_i(n)$ は、$c_i(n)$ の時間的な変化量（動的特徴）を表すものである。

今回の話者識別実験では、各フレームごとにケプストラム係数 (12 次元) と Δ ケプストラム係数 (12 次元) をまとめて 24 次元ベクトルとし、このベクトルに基づいて識別を行った。

4. VQ による話者識別システム

VQ はデータ圧縮技術の 1 つである。話者認識にも用いられ、高い認識率が得られることが報告されている。

4.1 ベクトル量子化の原理

入力するベクトルの中で、よく似たもの同士を 1 つのグループにまとめておき、全体をいくつかのグループに分類し、各グループの代表パタンを決める。このようにすると、各々のベクトルは、それ自身が属しているグループの代表パタンで近似できる。従って、代表パタンの集まりでベクトル全体が効率よく表現できる。

4.2 LBG の流れ

LBG アルゴリズムを用いてコードブックを作成する。LBG アルゴリズムでは、全データに基づいて、クラスタ分割を更新し、新たなセントロイドを計算する。本研究では、コードブックサイズが所望の個数になると終了するアルゴリズムでなく、任意のサイズのコードブックを作成する LBG アルゴリズムを用いた。

(a) クラスタ分割

図 3 に示すように、ベクトルを予め設定したコードワード数 m 個のクラスタに分け、クラスタ毎にセントロイドを計算する。

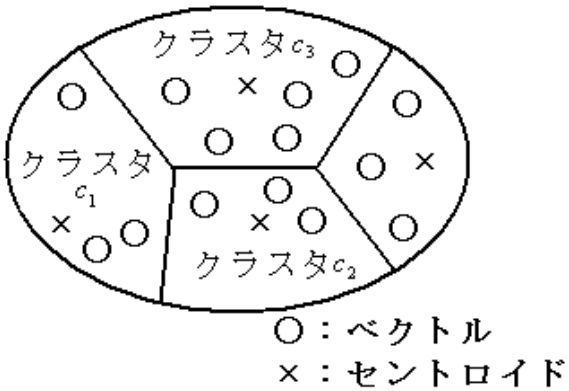

図 3: クラスタ分割の例

(b) 最小歪みのコードブックを作成する。

i. ベクトルと最も近いセントロイドを探し、歪みとセントロイドの番号を記録する。

ii. 同じセントロイドの番号を付けるベクトル群の平均を求め、新たなセントロイドをコードブックに書き込む。

iii. 平均歪みを計算する。

iv. 歪みの差分の絶対値が閾値より大きければ ii. と iii. を繰り返し、小さければ終了する。

話者ごとにコードブックを作成すれば、このコードブックよって個人性を表現することができる。未知の音声 x が入ると、その行き先を求めるのは VQ による話者識別の考え方である。初期値はランダムに選び、学習データを用いてコードブックを作成する。識別では同様にコードブックにより量子化し、量子化歪の最も小さい登録話者を取り、最終的な判定を行う。図 4 は、入力ベクトル x を代表ベクトルワード Y で近似することを示す。

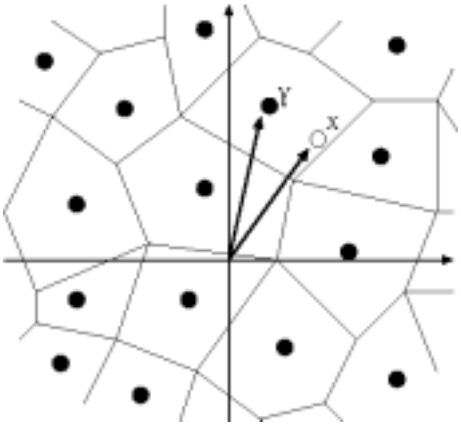

図 4: ベクトル量子化による話者識別

5. 実験と結果

実験全体の流れを図 5 に示す。

今回、無響室内に研究室のメンバー 13 人の音声を、三回繰り返し録音した。三回の中で、一回の音声を学習に使い、残る二回を識別に使った。

5.1 実験条件

実験条件を表 1 に示す。コードワードの数が少ないと歪（量子化雑音）が大きくなり、コードワードの数が増えれば、量子化雑音は減るが、処理が複雑になる。予備実験でコードワードが 16 個のとき、収束が最も良かったため、実験ではコードブックとして 16 個のコードワードを用いた。

5.2 実験の結果

(a) コードブックの作成

図 6 に作成したコードブックの例を示す。星印 ($\star$) の点はコードブックのコードワード（セントロイド）である。

(b) 識別の結果

LBG により生成されたコードブック

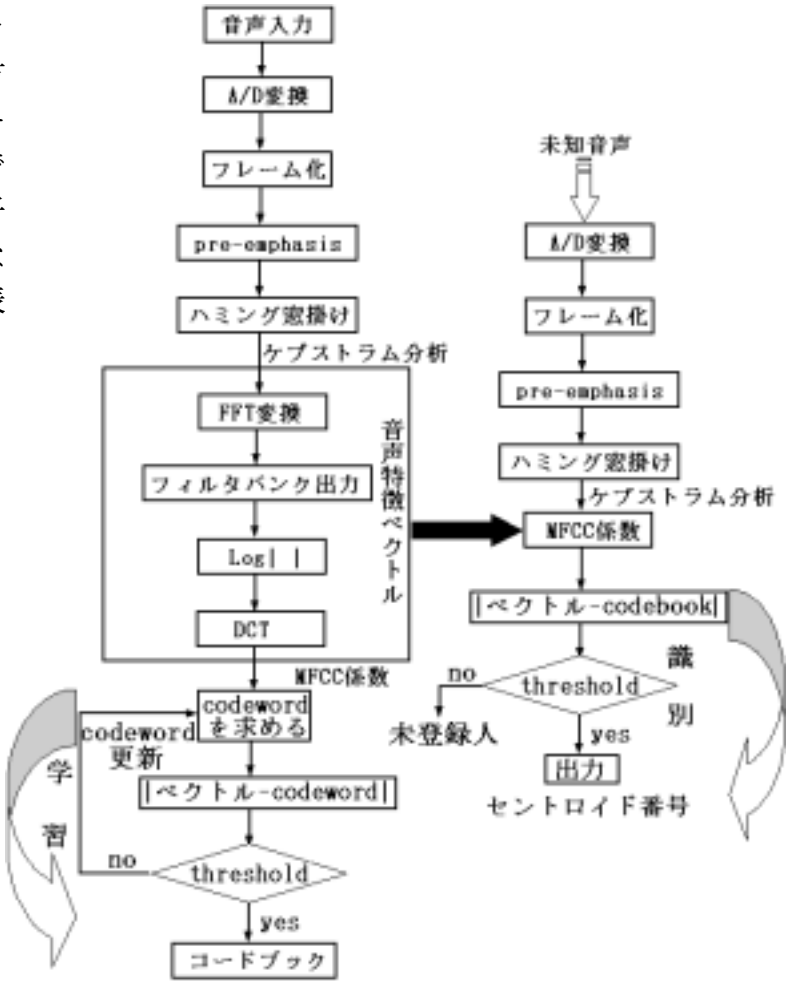

図 5: 話者識別の流れ

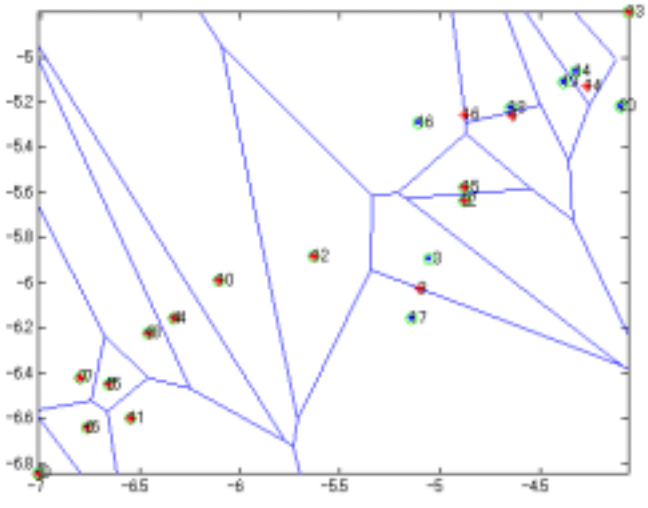

図 6: コードブックの作成例

表 1: 実験の条件

録音データ	連続「アオイウエ」
sampling 周波数	22.05kHz
量子化ビット幅	16bit
分析周期	30ms
シフト幅	10ms
分析窓	ハミング窓
pre-emphasis	0.99
mel フィルタ群	40 個
特徴パラメータ	12 次の MFCC 係数および Δcepstrum 係数
学習方法：VQ	コードワード 16 個、LBG method
学習人数	13 人

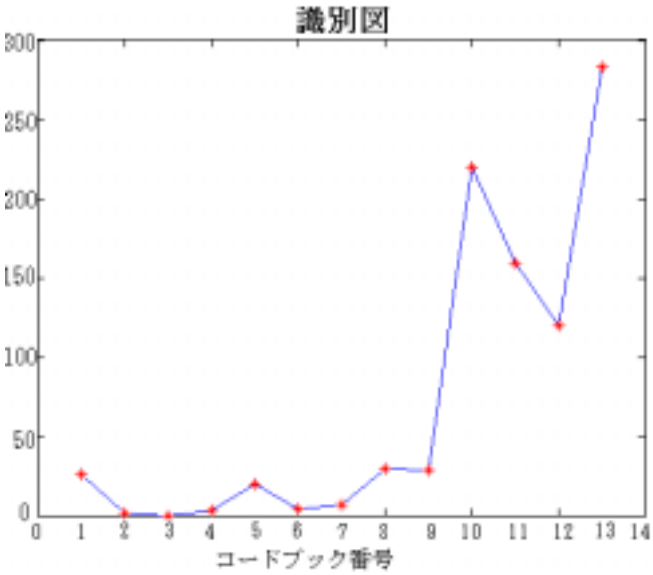

図 7: 実験結果

は初期値の選び方により違うので、局所解に陥る可能性がある。これを避けるために、初期値は 5 回選択した。1 回目はコードワード数 m で平均した値で、残る 4 回の初期値はランダム m 個を選んでアルゴリズム代入した。同じ点に収束したはコードブックにする。

識別の音声を代入すると、同様に局所解に陥ることを避けるために、初期値は 3 回を選択した。平均値の取る 1 回とランダムに選んだ 2 回である。結果は 2 回以上に同じ番号を表示すると、識別できた結果が得られた。

入力データが 3 番のコードブックの場合の識別結果の一例を 7 に示す。図から量子化の歪みが最も小さいコードブックは 3 番であることが分る。

今回の実験では、13 名全てで、100% の正解率を得た。

6. まとめと今後の課題

実験の結果から、本手法により話者が識別できることが分かった。少人数の規模ではあるが、MFCC 係数と Δ ケプストラム係数を特徴パラメータとし、VQ 法を用いる話者識別システムの有効性を確認できた。

今後の課題は、LBG アルゴリズムだけではなく、例えば様々な並列競合学習アルゴリズムを検討する必要がある。また、学習人数を増やすとともに、録音条件を下げた場合の誤識別に対する耐性について検討する。

MFCC 係数は、線形・時不変な受動フィルタバンクによる聴覚のモデル化を近似的に利用しているに過ぎない。一方、最近の研究より、聴覚モデルを構成する際には、非線形性や時不変だけでなく、能動性ももったフィルタ群が必要であることが指摘されている。また、異なったフィルタ間の相互作用も検討する必要がある。聴覚モデルの研究進展と、その新しい成果を利用した音声分析法の開発を進めていく [2]。

参考文献

[1] 古井　貞熙, 著建築・音声情報処理, 森北出版株式会社, 1998

[2] 安藤　彰男, リアルタイム音声識別, 社団法人　電子情報通信学会 ,2003

[3] Minh N. Do,An Automatic Speaker Recognition System,Audio Visua,Communications Laboratory,Swiss Federal Institute of Technology, Lausanne, Switzerland

[4] 今井　聖, 音声信号処理, 音声の性質と聴覚の特性を考慮した信号処理, 森北出版株式会社,1996

[5] 嵯峨山　茂樹, 東京大学　工学部　計数工学科　応用音響学, *http : //hil.t.u − tokyo.ac.jp/ sagayama/applied_acoustics*