音響的特徴を利用した自動話者分類*

小林恵太† 西崎博光‡ 関口芳廣‡
† 山梨大学大学院医学工学総合教育部・‡ 医学工学総合研究部

1 はじめに

複数の話者による発話を話者毎に自動分類するためには，GMM や VQ 歪み [1] を用いた手法が一般的に使用される．しかし，いずれも発話者に対する事前の学習が必要になり，不特定の話者には対処できない．そこで発話者に対する事前の学習が必要ない音響的特徴量を利用した対話音声の自動分類を目指す．

本稿ではまず人間による話者分類を分析し，自動話者分類方法を考案した．分析によると人間が話者分類を行なう際，音響的特徴が有効であることも分かった．そこで複数の話者が発話しているニュース音声を対象に，話者分類における音響的特徴量の有効性等について検討した．

2 人間による話者分類実験

2.1 実験概要

話者数を知らせず複数話者による対話を被験者に聞かせて発話を話者ごとに分類させ，分類に用いた要因，及び話者数を回答させた．被験者は成人男女 13 名である．また実験音声には会議音声データベース [2] 中の，男女各 2 名の 3 発話ずつを使用した．発話時間はいずれも 1 発話当たり 5 秒程度である．

2.2 実験結果及び考察

話者分類の要因として得られた回答例をまとめたもの及び回答数を表 1 に示す．また分類の正解率は 92.3%であり，話者数を多く回答してしまう誤りが見られた．

表 1 中の声質・声の高さは音響的特徴，また話し方・話の内容は言語的特徴だと言え，分類に用いる要因として音響的特徴の比率がやや高いと言える．そこで本研究ではまず声質・声の高さに関連する音響的特徴を用いた話者の自動分類を目指す．

表 1　人間が話者分類に用いる要因とその回答数

回答例 (回答数)	要因	回答数
声質 (10)	声質	14
声のこもり具合 (2)		
鼻にかかった声か (1)		
声の落ち着き具合 (1)		
声の高さ (8)	声の高さ	9
男声，女声の違い (1)		
話し方 (4)	話し方	9
発話速度 (3)		
訛り，イントネーション (1)		
男性，女性的な話し方 (1)		
話の内容 (2)	話の内容	3
口癖 (1)		

3 話者分類の方法

3.1 話者分類のための音響的特徴量

人間が話者の分類に用いる要因に対応する音響的特徴量として，以下のものを使用した．ただし 1 から 3 は母音別に，4 は発話平均から特徴量を求める．

1. 低次 (0 次から 9 次) のケプストラム (声質)
2. スペクトル (0-8kHz) の傾き (声質)
3. スペクトルのパワー比 (0-2kHz/2-4kHz)(声質)
4. 基本周波数 (声の高さ)

3.2 話者比較に用いる線形判別関数

2 つの発話が同一話者のものであるか否かを判別するための線形判別関数を，式 (1)，(2) のように表す．

$$y_i = W_0^i + W_1^i|x_1 - x_1'| + \cdots + W_n^i|x_n - x_n'| \quad (1)$$

$$d = W_0 + W_1 y_1 + \cdots + W_5 y_5 \quad (2)$$

y_i は母音毎に用意し，i=1,2, ••• 5 とする．

ここで $W_0^i,..,W_n^i$，$W_0,...,W_5$ は定数，$x_1,...,x_n$ 及び $x_1',...,x_n'$ はそれぞれ 3.1 節の特徴量を表す．また式 (2) の d が負となれば同一話者，正となれば異なる話者の発話として判断している．

3.3 話者分類アルゴリズム

前述の音響的特徴量，線形判別関数を用いた分類アルゴリズムを以下に示す．また全ての発話がいずれかのクラスタに属したとき，分類を終了する．

1. 任意の発話を含む新しいクラスタを作成する．
2. 全発話から任意の発話を一つ選び判別関数を適用し，1. のクラスタに属する全データとの判別結果の多数決をとる．
3. 多数ならば 1. のクラスタに対象発話を分類し，多数決が同数なら対象発話への判断を保留する．
4. 2 から 3 の作業を繰り返し，全ての発話を 1. のクラスタ内の発話と比較する．
5. 保留された発話を再度 1. のクラスタ内の発話と比較し，同一ならば発話をそのクラスタへ分類する．
6. 上記の作業により 1. のクラスタに属さないと判断された全発話を新たな発話集合として，1 の作業から繰り返す．

*Automatic speaker classification using acoustic features. by Keita KOBAYASHI, Hiromitsu NISHIZAKI and Yoshihiro SEKIGUCHI (University of Yamanashi)

表 2　判別関数の学習用データ

使用音声	日本語話し言葉コーパス (CSJ) [3]
話者数	50 名 (男性 25 名，女性 25 名)
発話数	話者一人当たり 10 発話，計 500 発話 各発話はいずれも 5 秒程度

4　話者分類実験

4.1　実験概要

音響的特徴量による自動話者分類実験を行なった．音響的特徴量には 10 次の低次ケプストラム及びケプストラムに 3.1 節に示したその他の特徴量を適宜加えた 5 種類を用意し，分類精度の比較を行なった．

実験対象音声にはニュース音声中の男性話者 9 人の 70 発話を使用した．ただし実験音声の発話者数は未知として分類を行なった．

また 3.2 節に示した線形判別関数は，表 2 に示す学習データから作成した．

4.2　評価方法

自動分類結果の評価尺度として，RandIndex 及び BBN 尺度 [4] を用いた．

RandIndex は式 (3) で定義され，I_{Rand} の値が小さいほどその分類は優れていると言える．また BBN 尺度は式 (4) で定義され，I_{BBN} の値が大きいほどその分類は優れていると言える．式 (4) における Q は分類数が増えることに対するペナルティーである．この実験では Q を 0.5 とした．

$$I_{Rand} = \frac{1}{2}\left\{\sum_i n_i^2 + \sum_j n_j^2\right\} - \sum_i \sum_j n_{ij}^2 \qquad (3)$$

$$I_{BBN} = \sum_i \sum_j \frac{n_{ij}^2}{n_i} - Q \cdot N_c \qquad (4)$$

n_{ij} : クラスタ i に属する話者 j の発話文数
n_i : クラスタ i に属する文数
n_j : 話者 j の発話文数
N_c : 総クラスタ数

4.3　実験結果

5 種類の音響的特徴量の組み合わせに対し分類実験を行なった．それらの評価結果の比較を図 1 に示す．4 種類全ての音響的特徴量を使用した場合が最も精度が高く，その結果を表 3 に示す．

同一話者の発話を異なるクラスタへ分類してしまったことや，総クラスタ数が話者数より多いという問題があるが，音響的特徴のみで全体としては約 75%の発話を正しく分類できた．

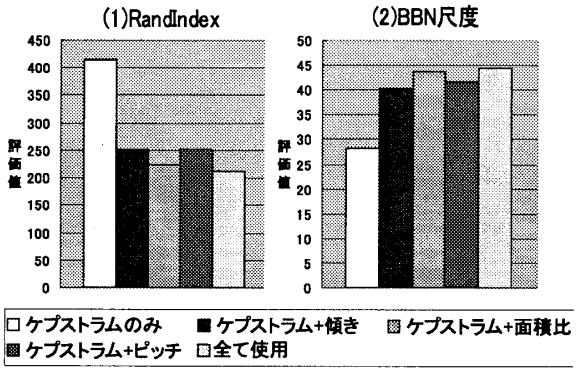

図 1　音響的特徴量と分類評価の関係

表 3　音響的特徴量による話者分類結果の例 (全特徴量を使用した場合)

C/S	S0	S1	S2	S3	S4	S5	S6	S7	S8	計
C0	9	0	1	0	0	0	0	2	0	12
C1	0	9	1	0	0	0	0	2	0	12
C2	0	0	3	0	0	0	0	0	0	3
C3	0	0	0	5	2	0	0	0	0	7
C4	0	0	0	0	3	0	0	0	0	3
C5	0	0	0	0	0	8	0	1	2	11
C6	0	0	0	0	0	0	5	0	0	5
C7	0	0	0	0	0	1	0	5	0	6
C8	0	0	0	0	0	0	0	0	6	6
C9	0	0	0	0	0	0	0	0	1	1
C10	0	0	0	0	0	1	0	0	0	1
C11	0	0	0	0	0	0	0	1	0	1
C12	0	0	0	0	0	2	0	0	0	2
計	9	9	5	5	5	12	5	11	9	70

C:Cluster, S:Speaker

5　おわりに

本稿では，人間が音声から話者を分類する際に用いる要因を調べ，それに対応する音響的特徴量を用いた話者の自動分類実験を行なった．この実験により自動分類における音響的特徴量の有効性が分かった．

今後は発話時間が短い音声に対して音響的特徴量が有効であるか否かを調べるとともに，言語的特徴の利用なども考えてゆく．また対話音声の分類に不可欠なリアルタイム性も検討する予定である．

参考文献

[1] 森一将，山本公一，中川聖一：発話間の VQ 歪みを用いたオンライン話者交代識別と話者クラスタリング，電子情報通信学会，信学技報，SP2000-18 (2000.6)

[2] RWCP：会議音声データベース
< *http* : *//unit.aist.go.jp/itri/itri* − *spg/rwc* − *db.htm* >

[3] K. Maekawa,"Corpus of Spontaneous Japanese: Its design and evaluation," Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003),pp.7-12(2003)

[4] A.Solomonoff,A.Mielke,M.Schmidt,H.Gish,"Clustering speakers by their voices", Proc. ICASSP '98,pp.757-760(1998)