対話音声を対象とした母音の音響的特徴による話者分類*

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

筆者らはこれまでに，発話中の音響的特徴のみを利用し，発話者に対する事前の学習を行わない自動話者分類手法を提案した [1][2]．その手法がニュース音声のような発話時間の長い読み上げ発話や，話し言葉に近い短時間発話中の話者の分類に効果があることを示した．

本稿では，対象音声を対話音声として分類アルゴリズムを見直し，より実用的な条件での自動話者分類について検討した．また提案手法の応用としてタスクによる条件を付与した場合の話者分類についても検討したので，併せて報告する．

2 話者分類手法

2.1 母音中の音響的特徴

本研究では母音中の音響的特徴に注目し，次の 4 種類の特徴量を使用した．ある 2 つの発話からこれらの特徴量を抽出し，その差の比較を行うことでこれら 2 つの発話が同一話者か否かを判別する．発話中の母音区間の特定には大語彙連続音声認識エンジン Julius [3] を使用する．

1. MFCC11 次元 (1 次から 11 次)
2. 周波数スペクトル (0-8kHz) の傾き
3. 周波数スペクトルのパワー比 (0-2kHz/2-4kHz)
4. 平均基本周波数 (声の高さ)

2.2 発話者判別手法

ある 2 つの発話が同一話者による発話であるか否かを判別するために SVM を用いる．SVM の入力は各母音の特徴量の差である．SVM のカーネルには式 (1) で表わされるガウシアン型カーネルを利用する．σ の値は 0.05 とした．

$$K(x,y) = exp(-\frac{||x-y||^2}{2\,\sigma^2}) \qquad (1)$$

判別の手順は，まず母音毎に判別結果を求めた後，その判別結果を統合する．この統合結果を最終的な判別結果とする．そのため SVM は母音毎の特徴量の差を学習データとしたモデル (5 種類) と，母音毎の SVM スコアを学習データとする結果統合用モデルの計 6 種類を用意した．また SVM の学習用データは表 1 の通りである．

表 1 話者分類用 SVM の学習用データ

使用音声	日本語話し言葉コーパス (CSJ)[4]
話者数	男性話者 50 名
発話数	同一話者： 話者一人当たり 70 発話，計 3500 発話 異なる話者: 話者一人当たり 10 発話，計 500 発話

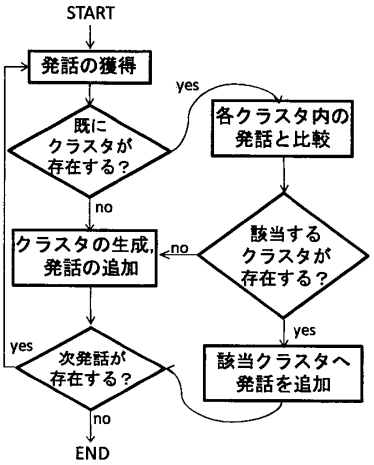

図 1 話者分類アルゴリズム

2.3 話者分類アルゴリズム

文献 [1][2] では，実験対象の発話を全て獲得した上で分類を始めることが前提であった．しかし対話音声を対象とした場合，発話の獲得順に処理を行ってゆくことが必須である．そのため分類アルゴリズムを文献 [1][2] の手法から図 1 のように変更した．

また発話をどのクラスタへ分類するかは，次の 3 種類を検討した．

1. クラスタ内の全ての発話との判別結果の多数決 (Maj)
2. クラスタ内の全ての発話の平均特徴量との SVM による判別結果 (Avg)
3. 2. の判別結果と，平均特徴量と分類対象発話の特徴量間のユークリッド距離の合計 (Avg_Dist)

3 対話音声を対象とした話者分類実験

3.1 実験概要

実際の対話を書き起こし，それを読み上げた音声を分類実験に用いる．実験対象音声についての説明を表 2 に示す．表 2 中の 8 名の話者を適宜組み合わせ，

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

表 2　実験対象音声

読み上げ対象	人間 4 名によるカードゲームプレイ時の音声対話
話者数	8 名 (成人男性)

4 通りの実験対話を作成した (EX1〜EX4)．これら 4 通りに対し，2.3 節の 3 種類の分類手法を適用する．

また音声対話型カードゲームタスクを想定し，上記の各実験にタスクに対応した以下の条件を与えた実験も行った．

- 話者数の限定 (4 名)
- 話者毎の初期クラスタの設定

3.2　評価手法

自動分類結果の評価尺度として，RandIndex [5] 及び発話単位の分類精度を使う．

RandIndex は式 (2) 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．また発話単位の分類精度とは，発話が最も集中したクラスタを正解とした，全発話に対する正しく分類できた発話の割合である．ただし同一クラスタに複数の話者の発話が挿入されてしまった場合，そのうち最も発話数の多い話者の発話数のみを使用している．

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

n_{ij} : クラスタ i に属する話者 j の発話数
n_i : クラスタ i に属する発話数
n_j : 話者 j の発話数

3.3　実験結果

分類手法と RandIndex の関係を図 2 に，分類手法と分類精度の関係を図 3 に示す．

まずタスクによる条件がない場合だが，RandIndex は EX1 と EX3 では Maj に比べ Avg，Avg_Dist の順に精度が良くなった．一方 EX2，EX4 では逆に Maj が最も良い精度であった．ただし EX2，EX4 ともに Avg に比べ Avg_Dist の方が精度が良いことから，判別に特徴量間のユークリッド距離を付加する効果はあると言える．また判別精度は EX4 以外の実験で特徴量間のユークリッド距離利用の効果が見られた．

タスクによる条件を付与した場合では RandIndex・判別精度とも，特徴量間のユークリッド距離導入の効果が表れているものが多いが，その効果がないものもある．その原因として，タスクによる条件を付与しない場合は冗長なクラスタに分類されていた発話が，クラスタ数に制限を設けることでいずれかのクラスタへ強制的に分類されることが挙げられる．そのため同一話者のクラスタへ分類された発話が分類精度を向上させ，逆に他の話者のクラスタへ分類された発話が RandIndex を悪化させたと思われる．

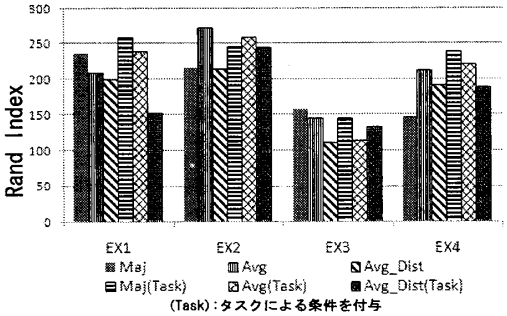

図 2　分類手法と RandIndex の関係

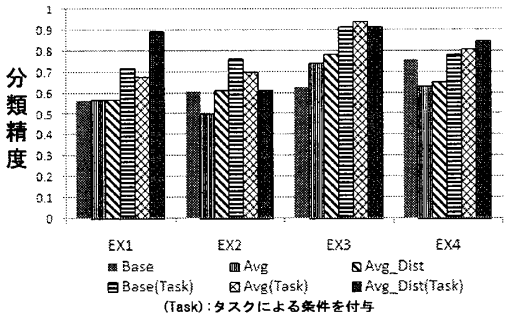

図 3　分類手法と分類精度の関係

以上の結果より，提案手法が対話音声に対する話者分類にある程度効果があることが分かった．

4　おわりに

本稿では，対話音声を対象とした，母音の音響的特徴を利用した自動話者分類について述べた．本稿で述べた分類手法により，対話音声に対してもある程度の分類精度を得ることができた．しかし話者の組み合わせにより精度にやや差があることから，今後さらに分類手法や音響的特徴の検討が必要だと言える．

参考文献

[1] 小林 恵太，西崎 博光，関口 芳廣，"音響的特徴を利用した自動話者分類"，情報処理学会，第 70 回全国大会講演論文集，3U-2，Vol.2，pp.137 - 138，2008.3

[2] 小林 恵太，西崎 博光，関口 芳廣，"母音中の音響的特徴を利用した自動話者分類の検討"，日本音響学会，2008 年秋季研究発表会講演論文集，3-Q-16，pp.187 - 188，2008.9

[3] 河原 達也，李 晃伸，「連続音声認識ソフトウエア Julius」，人工知能学会誌，Vol.20，No.1，pp.41-49，2005.

[4] 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

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