E-047

会議録自動作成システムに向けた話者識別技術の検討

A Study on Speaker Identification for Automatic Meeting Transcription System

水野寛之† 竹内 伸一‡ 田村 哲嗣¶ 速水 悟¶

Hiroyuki Mizuno Sinichi Takeuchi Satoshi Tamura Satoru Hayamizu

1. はじめに

近年,会議など多人数で行われる会話を対象とした話者識別の研究・開発が広く行われている[1][2][3].自動的に誰が話したか識別することで,自動的な字幕の生成や情報検索が容易になる.話者識別する際の問題点として,音響状況,未知の話者,クロストーク(以下:重畳音声)時などが考えられる.本稿では,会議録自動作成を目的とし,短い発話や発話が重なった場合の話者識別の検討を行った.

話者識別には,音声認識で使用されている,ガウス混合分布モデル(GMM)を用い,重畳音声時の対応として話者混合モデルを作成することで解決を行った.併せて,話者識別実験の結果を報告する.

2. GMM を用いた話者識別

本稿では,HMM を基に,GMM を作成し話者識別を行った.特徴量の抽出,モデルの学習,認識には HTK[4]を用いた.以下に処理の流れを示す.

- 音響特徴量の抽出(39 次元の特徴量,MFCC など)
- 話者別初期モデル(GMM)の作成・初期値の設定
  (VAD 結果によるラベルの付与)
- 話者別モデル(GMM)の学習(混合数の増加処理)
- 認識・評価

2.1 話者別モデル(GMM)の作成

話者別モデルを GMM を用いて作成した. 識別の仕組みは,各モデルにおいて観測系列を生起する確率を計算し,最大確率(尤度)を与えるモデルを発話者とする.図.1 に,GMM の構成と認識のためのネットワーク表現を記す.

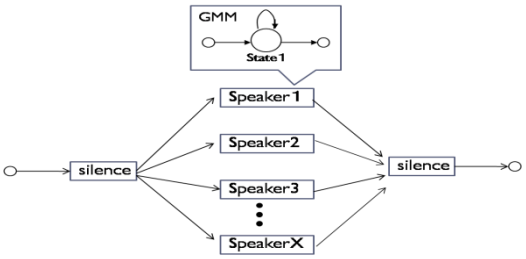

図. 1 GMM の構成とネットワーク

GMM の構成は, 中心の 1 状態に出力確率を持つ. また,認識のためのネットワークとして,silence からいずれかの Speaker を通り,silence に接続される処理である.これは,3.1 で述べる実験データに基づいている.

2.2 VAD(Voice Activity Detection)

VAD は音声区間検出のことで,音声/非音声区間を分ける技術である.これは,雑音環境下などでの認識率低下を抑える処理として利用され,音声認識する前処理として有効な手段である.さらに,時間情報も付与されるため識別率の向上が見込まれる.詳しくは,[5]を参照されたい.

3. 実験

上記に述べた技術を用い,重畳していない音声,重畳した音声,部分的に重畳した音声の場合にどれだけ話者識別できるのか実験をおこなった.

3.1 実験データと実験条件

実験データを,表.1 に記す.実験に用いる音声は,雑音なしの環境下であり,全話者共通発話,1 発話 3～6 秒,1 モデルあたり 50 発話を用いた.

重畳音声は,2 人の音声が同時に発話し始めるように合成した重畳音声(重畳モデル)と,重畳する片方の話者を 1 秒遅延し,部分的に重畳されるように合成した部分重畳音声(部分重畳モデル)を使用した. 36 モデルに含まれる重畳 28 モデルは,男女計 8 名から 2 名を選ぶすべての組合わせをした.また,男女 3 名ずつを増加させた,60 モデルの重畳 46 モデルと,部分重畳 46 モデルは任意の話者を組合わせた.

部分重畳モデルで識別実験をする際は,重畳している部分としていない(1 人のみ発声している)部分を識別できるように,図.1 のネットワークを書き換える必要がある. open 環境下での学習には分割学習法を用いた.

音響特徴量は,MFCC と,音声信号の対数パワーそれぞれの⊿,⊿⊿成分の計 39 次元を用いた.これは,VAD や音声認識する際に有効な特徴量である.

表.1 実験データ

音声データ	ASJ 音素バランス文 (A セット 50 発話) ・36 モデル(男 4,女 4,重畳 28) ・60 モデル(男 7,女 7,重畳 46) ・部分重畳 46 モデル 学習 : 25 発話 , 評価 : 25 発話
標本化・量子化	16kHz , 16bit
フレーム長・間隔	25ms , 10ms (ハミング窓)
特徴量	39 次元 MFCC(12),⊿MFCC(12),⊿⊿MFCC(12) Pow(1),⊿Pow(1),⊿⊿Pow(1)

†岐阜大学大学院工学研究科
Graduate School of Engineering, Gifu University
‡岐阜大学 VSL
Virtual System Laboratory, Gifu University
¶岐阜大学工学部
Faculty of Engineering, Gifu University

実験条件を,表.2 に記す.

表.2　実験条件

条件	モデル/数	VAD 有無	環境
i	重畳なし/8	あり	close
ii	重畳/28	あり	close
iii	重畳含む/36	なし	close
iv	重畳含む/36	あり	close
v	重畳含む/36	あり	open
vi	重畳なし/14	あり	close
vii	重畳含む/60	あり	open
viii	部分重畳/46	あり	close

3.2 結果

3.1 の実験データと実験条件を基におこなった話者識別結果を以下に記す.ここで,話者 a と話者 b の重畳モデルを識別する際,識別結果が話者 a + b となる場合に正解とした.また,部分重畳モデルの識別時には,話者 a,話者 a+b,話者 b と識別されるのが完全な正解である.

条件 i ～ v の実験結果を,図.2 に記す.

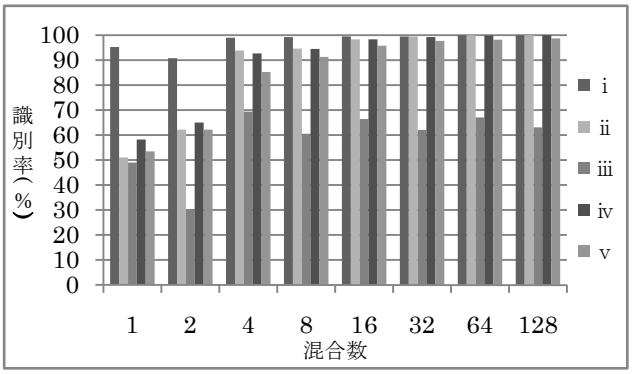

図.2　条件 i ～ v での話者識別結果

図.2 から, 条件iii と i , ii , iv , v から VAD の有無で大きく識別率の差がでた.これは,音声区間検出の有効性を表している.条件 i と ii ～ v から,重畳音声を含めることにより識別が困難なことを表している.しかし,混合数 16 以上でほぼ誤識別がない結果になった.また,条件iv と v から,僅かな差ではあるが open 条件の方が識別率が低くなっている.全体的に混合数が増えるにつれて識別率が向上している.

条件vi ～viiiの実験結果を,図.3 に記す.

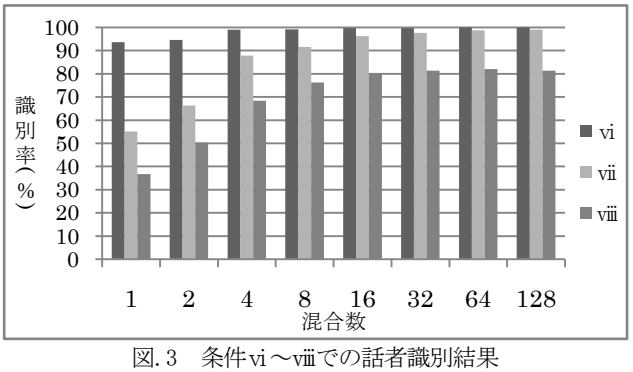

図.3　条件vi ～viiiでの話者識別結果

図.3 より, 条件vi と図.2 の i から重畳なしでの話者識別は,高い識別率となっている.また,条件vii と図.2 の v からモデル数を増加させても高い識別率になった.条件viiiの結果から,部分重畳モデルの識別が他に比べて難しくなることが分かった.

3.3 考察

重畳音声でないモデルの場合と,重畳音声モデルの場合も,全体的に高い識別率を得た.しかし,実環境などの雑音がある場合には識別率が下がると考えられる.また,close 環境時と open 環境時に識別率にあまり差がなかったのは,学習発話数を全体の半分としたため,各話者の特徴量を十分に取得でき,僅かな差になったと考えられる.

部分重畳の識別結果については,その結果の内枠を考察する.今回,合成された部分重畳音声は,一人の話者のみが発声している時間帯は約 0.1～1 秒で,非常に短かったため誤識別が多く確認された.一方,音声が重畳されている時間帯は約 2～3 秒あり,うまく識別されていた.話者識別を正確におこなうためには,少なくとも 1 秒は音声のデータが必要なのではないかと考えられる. 今回とは逆に,重畳している時間帯を短くすると,重畳している 2 話者を特定しなければならないので,1 話者を特定する場合よりも,識別率は低下すると考えられる.

4.　まとめ

本稿では,会議などで起こる重畳音声に対し,GMM モデルの作成や VAD を用いることで,どれだけ話者の識別が可能であるか実験した.実験により,VAD をすることで大幅な識別率の向上を確認した. 重畳音声の識別は,重畳していない音声の識別よりも識別率が低下した.また,部分的な重畳音声の場合には,重畳時間が短くなると識別が困難になることを確認した.

今後は,雑音環境下での話者識別や,部分重畳している時間の短縮,発話や話者比率の変更などを施し,より会議の状況に近い条件で実験を試みていく.将来的には,研究室内で取り組まれている,字幕生成システムや,マイクロフォンアレイを使用した音源方位推定技術と組み合わせ,会議録自動作成システムを構築していく.

参考文献

[1] 石塚健太郎,荒木章子,大塚和弘,藤本雅清,中谷智弘,”音響情報と映像情報の統合による多人数会話における話者決定技術”, 音声研究会, SP2008-83, pp. 25-30, 2008.

[2] 藤原 弘将, 北原 鉄朗, 後藤 真孝, 駒谷 和範, 尾形 哲也, 奥乃 博,”調波構造抽出と高信頼度フレーム選択を用いた雑音下での話者識別”, 日本音響学会 2006 年春季研究発表会, 1-11-17, 2006.

[3] 西田昌史,秋田裕哉,河原達也,”討論音声を対象とした話者モデル選択による話者インデキシングと自動書き起こし”, 電子情報通信学会技術研究報告, pp.55-60 2002

[4] The HTK BOOK, http://htk.eng.cam.ac.uk/

[5] 羽柴 隆志, 竹内 伸一, 田村 哲嗣, 速水 悟,”マルチストリーム HMM を用いた音声と画像による音声区間検出”,日本音響学会 2009 年春季講演論文集, 1-P-5, pp.131-132