J-040

簡易な教師無し会話活発度分類器作成のための話者交代時音響特徴量の抽出

Extraction of acoustic features at the time of speaker change for a simple unsupervised conversation activity level classifier

西田 悠 † Haruka Nishida
中平 勝子 † Katsuko T. Nakahira
北島 宗雄 † Muneo Kitajima

1 はじめに

1.1 背景

近年，大学教育においてアクティブ・ラーニングの重要性が増している．また，小・中学校などにおいても，「主体的・対話的で深い学び」をさせることが必要であるとされている．

アクティブ・ラーニングとは，「教員による一方向的な講義形式の教育と異なり，学修者への能動的な学修への参加を取り入れた教授・学習法の総称」であるとされ，教室内でのグループ・ディスカッション，ディベート，グループ・ワーク等が有効なアクティブ・ラーニングの方法であると言われている．このような状況に置いて，いわゆるグループ学習の評価をどのように行うかが重要となってくる．

そこで筆者らは，学習中の会話音声から，グループ学習におけるコミュニケーションの巧拙やその場の雰囲気を評価し，グループ評価を行うこと検討している．

このようなグループにおける会話についての研究は盛んに行われている．

張ら [1] はグループ内での会話コンテキストを推定し，そのグループの状態に応じた適切なアドバイスを，会話エージェントを通して提示するというグループ内会話における意思決定支援を検討している．河野ら [2] らは音声認識において発話意図を分析し，テキストから推定できなかった部分については発話の音響特徴量を分析して発話意図を推定するシステムを検討している．また西村ら [3] は，2 者間の会話を調べ，その同調傾向や会話の盛り上がりなどが音圧レベルや基本周波数，発話速度などの変化に現れることを分析している．横森ら [4] は，女性を対象として女性の発話好感度に影響与える音響特徴量の分析を行っている．

1.2 目的

本稿では，コミュニケーション場の雰囲気判定のための一歩として，コミュニケーション場の活発度の判定指標に必要な音響特徴量抽出を行う．そのために，会話音声から話者交代を自動的に推定し，発話頻度や参加人数などから会話の活発度を簡易に算出することを考える．

話者認識については，一般的に混合ガウス分布やサポートベクタマシンなどで話者音声データを学習して話者ごとのモデルを作成し，モデルのパラメータを推定，最適化するという流れが一般的である．[5] しかしこれには多量の話者教師データが必要となり，ディスカッションや会話への参加者から下準備として教師データを集める必要がある．

集音用のマイク 1 本のみから簡易に活発度判定を行うことを想定し，多量のデータを必要としない教師無しの分類器を用いて会話音声中の話者ごとの特徴量を分類し，個人を特定しない簡易な話者分類を行う．

2 手法

2.1 活発な会話

本稿においての「活発な会話」とは，話者の交代が頻繁に起こり，特定の話者だけが話し続ける状況になく，グループ内の全員が会話に参加して意見を述べることができる状況であると定義する．また，会話が活発であるかを示す「活発度」を規定し，これを測定することを考える．

グループ内での会話の活発度が測定できることで，グループ・ディスカッションやグループ学習等におけるグループ評価に寄与できる．

ここで，ある会話が活発か否かということについて，水上ら [6] は表 1 に示した項目を検討すると良いのでは

† 長岡技術科学大学

表 1　活発な会話

見るべき項目	関連するパラメータ候補
誰かの発言中，傍観者となっていないか	応答者のバランス，音声的同意表現の有無
誰かの発言に対して，応答しようとしているか	発話量，ポーズ量，発言-応答ペアの成立数
発言者が偏っていないか	各人のフロア数
発言の受け止め役やまとめ役が固定されていないか	各人のフロア数，メタ議論，発言者バランス

ないかということを提示している．

これらの項目を参考に，会話の活発度を規定するために必要なパラメータを考える．

2.2　会話の活発度

活発な会話においては少なくとも，各参加者の発言量が同等である必要があると考えられる．そこで，本稿ではこれを会話音声から検証するために以下の 2 点を測定する．

- 会話音声中での話者交代を検知し，その数を測定する
- 各話者間の発言時間の比を計算することで話者の偏りを測定する

これらを測定するために会話音声を解析し，話者分類器を作成する．各時間における話者を特定することで話者交代を検知し，交代までの発言時間を測定することを考える．話者を分類するにあたり，話者を規定する音響的特徴量を会話音声から抽出し，分類器に利用する特徴量として利用する．

2.3　音響特徴量

音声の特徴を示すパラメータとして，以下ようなの特徴量が挙げられる．

- 音圧レベル
- ピッチ
- フォルマント周波数
- Mel-Frequency Cepstral Coefficients (MFCC)
- ポーズ長
- 発話速度

このうち本稿では，個人差を表しやすいとされる「ピッチ」，「フォルマント周波数」，「MFCC」を個人の特徴として利用し，分類器にかける特徴量を作成する [7].

2.4　話者分類器作成の手順

会話音声データから，前節で挙げた 3 種類の音響特徴量をフレーム分割処理を行って抽出する．フレーム長は $FL = 0.25\,[\mathrm{ms}]$ とし，フレームシフトは $FS = 0.01\,[\mathrm{ms}]$ とする．

MFCC は多次元の特徴量で構成され，1 つのフレームから n 次元が抽出されるが，音声認識などにおいて一般的に利用されている $n = 12$ 次元を採用することとし，$m_1, \ldots, m_{12}$ とする．フォルマント周波数は，母音を区別するのに足る第 1 フォルマント $(F1)$，第 2 フォルマント $(F2)$，第 3 フォルマント $(F3)$ を利用することとし，それぞれの対数を取る．これは MFCC に比べて $F1, F2, F3$ の数値が大きく，これらのみが強調されることによる分類の誤りを少なくするためである．同様の理由からピッチも対数をとり，p とする．これらから以下のような特徴量ベクトル A を作成する．

$$A = (m_1, \ldots, m_{12}, f1, f2, f3, p) \tag{1}$$

特徴量ベクトル A は 1 つのフレームごとに抽出される．

さらに，各フレームを s 個集めて「セクション」A' を作成する．

$$A' = \begin{pmatrix} m_{1,1} & \ldots & m_{1,12} & f1_1 & f2_1 & f3_1 & p_1 \\ m_{2,1} & \ldots & m_{2,12} & f1_2 & f2_2 & f3_2 & p_2 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{s,1} & \ldots & m_{s,12} & f1_s & f2_s & f3_s & p_s \end{pmatrix} \tag{2}$$

集めた各ベクトルの列方向の平均を取り，「セクション」特徴量 A'' を作成する．

$$A'' = \left(\bar{m}_1, \ldots, \bar{m}_{12}, \bar{f1}, \bar{f2}, \bar{f3}, \bar{p}\right) \tag{3}$$

特徴量の分類は $k-$means 法によって行う．会話に参加している人数は事前情報 h として与えられていると仮定し，これをクラスタ数として利用する．

3　会話の活発度判定手法評価

前章で規定した分類器を利用し，会話音声分析を行う．

3.1　音声データ

ある会話例文 [8] を 2 人の男性に読んでもらい，活発な会話を演じてもらった音声データを利用した．会話音声データは 44.1 [kHz] で録音され，16 [kHz] にダウンサンプリングした．簡単のため，話者らは間をおかずに発言を行い 2 者間の交代があるのみの音声とした．4 組のペアからデータを収集し，各音声は平均 2 分の音声データである．取得データに対し，第 2.4 節において示した手順によって特徴量を抽出し，話者の交代を検知する．

また抽出には音声解析ツールである「SPTK」を利用した．

3.2　試験結果

話者交代の例を図 1 に示す．

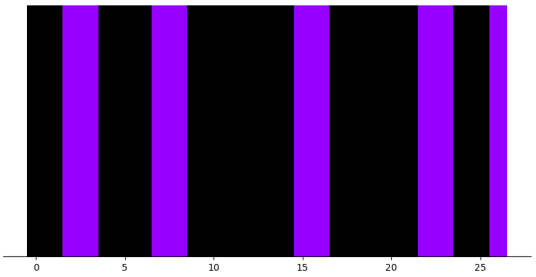

図 1　話者遷移例

紫色の部分が「話者 1」，黒色の部分が「話者 2」となっている．色の変化した部分が，話者が遷移したタイミングである．

分析に使用した会話音声では 2 人間の会話の遷移は 9 回行われていたが，図 1 においても 9 回の遷移が確認できる．

2 人の男性のペアは 4 組あり，4 パターンの会話音声の分析を同様に行った．

3.3　集積フレーム数

話者分類器の構築に必要な「セクション」の計算には，いくつかのフレームの集積が必要である．集積フレーム数は，セクションを構成する時間と等価であるため，短すぎるフレーム数では分類精度が悪くなると考えられる．そこでこれを変数として話者交代数が最も真値と近くなる集積数を調べた．分析単位としたい秒数 i を変数として分類を行い，話者交代数の真値を n_{hcT}，話者分類器によって判定された話者交代数を n_{hcL} としたとき，

$$\Delta N = n_{hcT} - n_{hcL} \tag{4}$$

が最も 0 に近くなる i を求めた．集積フレーム数と分類状況の変遷を図 2 に示す．上から $m = 2$, $m = 6$, $m = 10$ の場合の分類状況である．

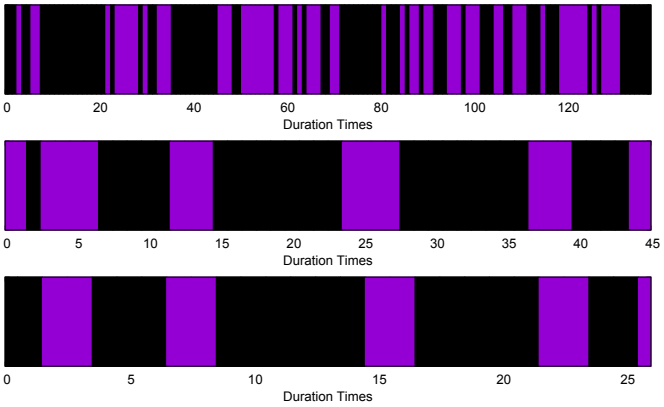

図 2　集積フレーム数と分類

また，図 3 に i と ΔN の関係を示す．

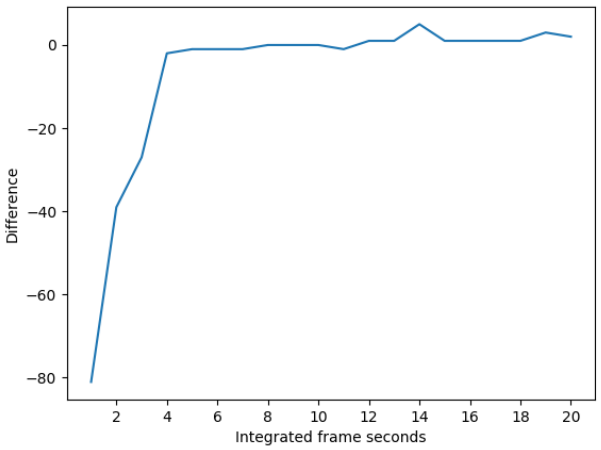

図 3　集積フレーム数と判定された話者交代数と真値との差

図 3 において，$i < 5$ くらいでは細かく分類されすぎて ΔN が負に傾き真値とかけ離れた話者交代数となる．$i = 5$ くらいから交代数が真値に近づき，一部を除いて 0 近傍で安定している．

本稿では，$\Delta N = 0$ となり，かつ各話者の発言時間が真値と最も近くなっている $i = 10$ を採用した．発言時間については 3.4 節において説明する．

さらに結果として，各パターンで $i = 10$ が最も適した集積数となった．このため，図 1，図 2 においては $i = 10$ での結果を示している．

3.4 発言時間の比

分類結果から各話者の発言時間を計算することで，話者らの発言時間の比を求めることを考える．計算方法を式 5 に示す．

SL を 1 セクションに集積されたフレーム数とし，NoC をある話者が紐付けられたセクションの数とすると，

$$SL = i/FL$$
$$ST = (FL + FS \times (SL - 1)) \times NoC \qquad (5)$$

によって各話者の発言時間 ST を求めることができる．

図 1 を例として，各話者の発言時間を表 2 に示す．

表 2　各話者のおおよその発言時間

	話者 A [sec]	話者 B [sec]
真値	70	40
分類結果	66	44

実測において真値とほぼ同様の時間が得られている．

話者 A と話者 B の発言時間の比は $66/44 = 1.5$ となって偏りが見られるため，活発な会話の定義に当てはまらず，活発な会話とは言えない．

4 まとめと今後の課題

本稿では，2 者間の会話音声から抽出した音響特徴量を利用して話者を分類し，話者交代数と各話者の発言時間を測定した．音響特徴量は「MFCC」,「フォルマント周波数」,「ピッチ」の 3 種類を利用し，フレーム分析を行い，最適な分類に適した集積フレーム数を推定した．

集積フレーム数 $i = 10$ とした時に，話者交代数と各話者の発言時間が真値と最も近くなり，i が小さい場合には交代数が真値とかけ離れた数値となった．これから，正確な分類を行うためにはある程度まとまった時間分の特徴量を用いることが必要と考えられる．

しかし，本稿で利用した会話音声は 2 分弱の短いものであり，各話者の発言時間も 1 回につき十数秒と短めであった．長時間であったり，1 回の発言時間が長い会話音声を利用した場合，i の最適値が変化する可能性が考えられる．

またこれは，扱うデータが短いことやデータ量の不足が懸念されるため今後より多くの会話音声において検証が必要と考える．抽出する特徴量の変更や追加なども行い，最適な特徴量の検証も必要である．

さらに，本稿で利用した会話音声には発言者が発言しない区間（ポーズ）が存在しないが，実際の会話においてポーズ時間は活発度を測定する上で重要なパラメータであるため，これを分析できるようにすることを考える．

活発度についても，話者交代数や発言時間バランスなどの各指標のしきい値を決定し，活発度を数段階表記で表せるようにすることも必要である．

謝辞

本研究の一部は科研費 16K01061, 16K00436 の助成を受けたものである．

参考文献

[1] 張, 黄宏軒, 木村清也, 岡田将吾, 大田直樹, 桑原和宏. グループディスカッション参加者の機能的役割とその変遷の分析．ヒューマンインタフェース学会論文誌, Vol. 20, No. 1, pp. 31–44, 2018.

[2] 河野進, 相良健郎. グループ会話における発話意図の推定システム．情報処理学会論文誌, Vol. 58, No. 5, pp. 1113–1123, 2017.

[3] 西村良太, 北岡教英, 中川聖一．音声対話における韻律変化をもたらす要因分析．音声研究, Vol. 13, No. 3, pp. 66–84, 2009.

[4] 横森文哉, 二宮大和, 森勢将雅, 田中章浩, 小澤賢司. 好感度評価の性差に着目した女性発話の音響特徴量分析．日本感性工学会論文誌, Vol. 15, No. 7, pp. 721–729, 2016.

[5] 小川哲司, 松井知子. 話者認識で用いる機械学習．日本音響学会誌, Vol. 69, No. 7, pp. 349–356, 2013.

[6] 悦雄水上, 郁代森本, 裕子大塚, 佳奈鈴本, 和広竹内, 学奥村, 秀紀柏岡. 話し合いを評価するための評価パラメータの検討 (1). 言語処理学会 第 15 回年次大会発表論文集, pp. 769–772, 2009.

[7] 山田大輔, 北岡教英, 中川聖一．音源情報の特徴量を用いた音声認識．電気学会論文誌. C, 電子・情報・システム部門誌, Vol. 122, No. 12, pp. 2028–2034, 2002.

[8] William M. Newman and Michael G. Lamming. インタラクティブシステムデザイン．ピアソンエデュケーション, 1999.