4C-03

画像情報と音声情報を用いた発話者判別手法における汎用性に関する検討

中村　悦郎†　　景山　陽一†　　廣瀬　聡‡
秋田大学†　　日本ビジネスシステムズ‡

1.　背景・目的

近年，働き方改革の実現に向けて，業務の効率化や労働環境の見直しが行われている．特に，職場における労働の改善策として，会議の効率化が重要視されている[1]．例えば，音声認識技術を応用して構築された議事録自動作成システムは，議事録作成におけるヒューマンエラーの低減や，議事録作成に要する工数の削減が可能である．特に，議事録自動作成システムにおいて，発言ごとに発話者を自動判別する技術は，会議および業務の効率化に寄与すると考える．

現在，実用化されている議事録自動作成システムには，発話者判別手法として，マイクを発話者に割り振る方法や，会議参加者の声紋を機械学習モデルに学習させる手法[2]が用いられている．しかしながら，これらの手法は，設備を整える必要があることや，声紋登録が必要であるなど，事前準備を必要とする点が課題である．これらの課題は，画像情報から取得可能な口唇の動きと音声情報に基づいて発話者を推定することで解決可能であると考える．

本研究では，利便性の高い議事録自動作成システムの構築を目的とし，口唇の動きと音声情報を対象とした，発話者判別手法に関して検討を加えた．具体的には，音声の特徴量と Long short-term memory (以降，LSTM と略記) [3]を用いて発話者の口唇の動きを推定し，推定値と最も類似した口唇の動きを有する人物を発話者として判別した．本稿では，被験者 14 名を対象として，発話者判別手法における特徴量の汎用性に関する検討を加えた．

2.　使用データ

一般的な蛍光灯下(照度 400～900lx)において，全方位カメラ(THETA V：RICOH 社製)およびマイクロフォン(TA-1：RICOH 社製)を用い，被験者 14 名(以降，A～N と表記)が 11 種類の文章(以降，文章 1～文章 11 と表記)を 1 回ずつ発話する様子を正面約 50cm から撮影・録音し，発話動画データを取得した．11 種類の文章は，Web のニュース記事から抜粋したものを使用した．なお，本研究は「秋田大学手形地区におけるヒトを対象とした研究に関する倫理規程第 6 条第 2 項」に基づき，被験者の同意を得た上でデータを取得している．

3.　提案手法

3.1.　口唇の動き特徴量

提案手法の概要を図 1 に示す．はじめに，オープンソースライブラリの Dlib[4,5]に搭載されている顔検出機能を使用し，発話動画データから 68 点の顔の特徴点を抽出した．次に，顔の特徴点のうち口唇部分に位置する 20 点を使用して，口唇の縦幅と横幅を取得した．被験者とカメラ間距離が変化した場合，発話に伴う口唇の動き以外の要因で，画像中の口唇の縦幅と横幅は変化する．このため，鼻の縦幅に対する口唇の縦幅の割合(以降，口唇縦幅の特徴量と表記)と，鼻の横幅に対する口唇の横幅の割合(以降，口唇横幅の特徴量と表記)を算出した．最後に，口唇縦幅の特徴量と口唇横幅の特徴量の積を口唇の動き特徴量として算出し，検討に使用した．

3.2.　音声の特徴量

Mel-Frequency Cepstrum Coefficient（以降，MFCC と略記) [6]を取得し，音声の特徴量として使用した．MFCC は一般的に，音声認識のための特徴量として使用される．MFCC に含まれている複数の特徴量のうち，最も低い次元(0 次元目)から 10～15 次元程度までが音声認識に使用される[6]．本検討では，音声の特徴量として，MFCC の 0～19 次元目の特徴量をそれぞれ検討に使用した．

3.3.　前処理

動画のフレームレートと音声のサンプリングレートが異なるため，1 つの発話動画データにおける口唇の動き特徴量と音声の特徴量のデータ量は異なる．両特徴量におけるデータ量を等しくするために，口唇の動き特徴量を線形補間して使用した．

3.4.　LSTM の学習処理

音声の特徴量から口唇の動き特徴量を推定するために，LSTM [3]を使用した．具体的には，音声の特徴量 100 フレーム(約 1.0 秒間)を入力し，50 フレー

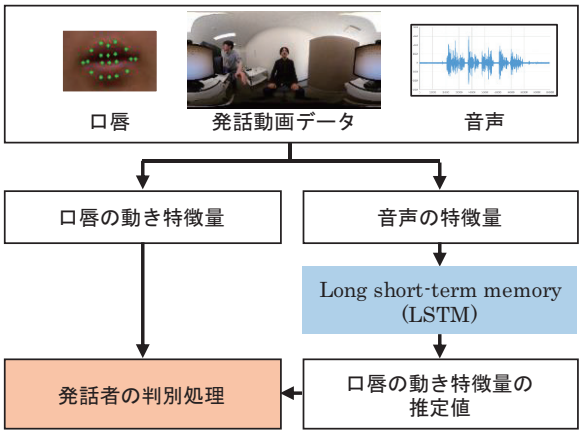

図 1　提案手法の概要

Versatility of Speaker Identification Method Using Image and Voice Information
Etsuro Nakamura†, Yoichi Kageyama†, Satoshi Hirose‡
†Akita University, ‡Japan Business Systems

ム目の口唇の動き特徴量を出力するように LSTM の学習を行った．このとき，モデルの入力層と出力層を 1 次元，LSTM 層を 100 次元に設定し，勾配法には Adam[3]を使用した．なお，各特徴量は発話動画データごとに標準化して入力し，音声の特徴量は 3.2 節に前述した MFCC のうち，任意の 1 次元を使用した．

3.5. 発話者の判別処理

実際の口唇の動き特徴量と LSTM を用いて推定された口唇の動き特徴量の推定値が最も類似している人物を発話者として判別した．はじめに，発話動画データ内の同じ時刻間から，実際の口唇の動き特徴量および推定値の時系列データをそれぞれ 50 フレームずつ取得した．次に，これら 2 つの時系列データ間の相関係数[7]を算出した．この相関係数を，発話動画データ内における全てのフレームを対象として算出した．さらに，相関係数が 0.4 以上の事例(弱い正の相関～強い正の相関)の割合を算出し，これを類似度として記録した．最後に，類似度が最も高い被験者を発話者として判別した．

4. 評価実験

4.1. 発話区間の設定

本検討では，発話区間のフレームのみを検討に使用するために，手動で発話区間を設定した．はじめに，発話開始時のフレームにおいて，最後に口を閉じていたフレームを発話開始フレームに設定した．次に，発話終了時，口を閉じたフレームを発話終了フレームに設定した．最後に，発話開始フレームと終了フレーム間を発話フレームと定義し，これらのフレームを以降の検討に使用した．

4.2. データセットの作成

モデルの性能や汎用性を評価するためには，異なる被験者かつ異なる文章の発話動画データを教師データとテストデータにそれぞれ設定する必要がある．本検討では，表 1 に示すように 8 種類のテストデータを設定した．また，テストデータと被験者および文章が異なる教師データを，各テストデータに割り当てた．なお，発話区間におけるデータ量の平均値に基づいて，データ量の差が小さくなるように文章の組み合わせを選定した．

4.3. 実験内容

4.2 節で設定した教師データを入力して LSTM の学習を行い，テストデータを用いて発話者判別成功率の評価を行った．はじめに，テストデータにおける 1 名の被験者の 1 文を発話者データとして設定した．次に，テストデータから発話者データと異なる被験者，かつ異なる文章のデータ 1 つを非発話者データとして選定した．さらに，発話者データにおける音声の特徴量を用いて，口唇の動き特徴量の推定値を算出し，これと発話者データおよび非発話者データを用いて発話者を判別した．このとき，発話者データが発話者と判別された場合，判別成功と定義した．最後に，発話者データごとに判別成功率を算出した．なお，LSTM の学習回数を 10～100 まで 10 刻みで変更し，発話者判別成功率の平均値が最も高い学習回数の結果を比較した．

表 1　テストデータの設定

	文章	
被験者	1, 3, 7, 8, 10, 11	2, 4, 5, 6, 9
A～E	テストデータ 1-1	テストデータ 1-2
D～H	テストデータ 2-1	テストデータ 2-2
G～K	テストデータ 3-1	テストデータ 3-2
J～N	テストデータ 4-1	テストデータ 4-2

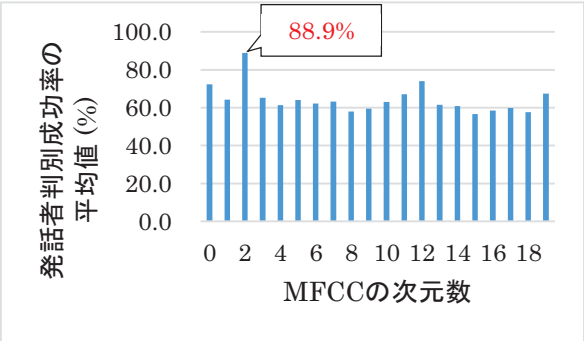

図 2　MFCC の次元数と発話者判別成功率の関係

4.4. 実験結果・考察

図 2 に音声の特徴量として使用した MFCC の次元数と発話者判別成功率の関係を示す．音声の特徴量として MFCC の 2 次元目を使用した場合における発話者判別成功率が最も高く，平均で 88.9%の結果を得た．本検討で算出した MFCC の 2 次元目は約 100Hz の周波数成分を含む特徴量である．したがって，これらの結果から，発話に伴う人物の口唇の動きと 100Hz 付近の音声の変化には関連があると考える．

以上の結果は，①提案手法が発話者の判別に有用である点，②MFCC の 2 次元目を音声の特徴量として使用することで高い発話者判別成功率が得られることを示唆している．

5. 謝辞

本研究は JSPS 科研費 JP19K12909 の助成を受けて行われた．

参考文献

[1] 日本経済団体連合会：2019 年労働時間等実態調査 集計結果, https://www.keidanren.or.jp/policy/2019/076.pdf (Accessed 2020/12/21)
[2] Microsoft: Azure Speaker Recognition, https://azure.microsoft.com/ja-jp/services/cognitive-services/speaker-recognition/ (Accessed 2020/12/21)
[3] 斎藤康毅：ゼロから作る DeepLearnimg②自然言語処理編, 株式会社オーム社 (2018)
[4] Dlib C++ Liblary, http://dlib.net/ (Accessed 2020/12/21)
[5] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou and Maja Pantic: 300 Faces In-The-Wild Challenge: Database and results, Image and Vision Computing (IMAVIS), Vol. 47, pp. 3–18 (2016)
[6] 篠田浩一：音声認識，講談社 (2017)
[7] 星野満博，西崎雅仁：数理統計の探求 -経営的問題解決能力の開発と論理的思考の展開-，晃洋書房 (2012)