6ZE-09

クラウドソーシング収集発話を用いた深層学習若年話者判別

奥本 佑哉 †
和歌山大学大学院システム工学研究科 †

西村 竜一 ‡
和歌山大学データ・インテリジェンス教育研究部門 ‡

1 はじめに

自動音声認識の能力を有する対話型システムでは，対話相手に合わせてシステムの応対を柔軟に変更することでユーザ体験の向上が期待できる．また，これまでも，話者の年齢を自動推定する研究は行われているが，高齢化社会の課題を想定して対象が高齢者であることが多く，子どもに着目した事例は限られている．我々の研究では，音声信号の音響的特徴から，その話者が若年者であることを判断する自動若年話者判別の手法を検討している．

先行研究 [1] では，クラウドソーシングによって2,361 個の年齢・性別ラベル付き実環境発話を収集している．さらに，収集した音声信号から抽出したMFCC(メル周波数ケプストラム係数) を特徴量にして，若年者・大人（男性）・大人（女性）のクラスに話者を分類する GMM-HMM 音響モデルを構築した．

本研究では，アルゴリズムに深層学習を導入した．本稿では，先行研究と比較した実験結果を示す．

2 若年話者判別タスクの定義

本研究の若年話者判別は，入力となる発話の音響信号から，話者が若年者（子ども）と大人のどちらに属するか推定する自動クラス分類のタスクである．

音声から判別する際，人によって時期が異なる「声変わり（変声期）」によって，大人と子どもを区分する定義そのものが難しくなることが予想される．そこで，本研究では，境界が変動する「年齢閾値」という概念を設けている．年齢閾値は，話者を若年者と大人に区分する境界の年齢である．話者の年齢が，年齢閾値未満であれば若年者，それ以上であれば大人であるとみなす．本研究では，年齢閾値を一つに固定せず，値を変えて検証することで，機械が判別することが可能な境界の年齢値を調査することも目的としている．そのため，あらかじめ法令や社会

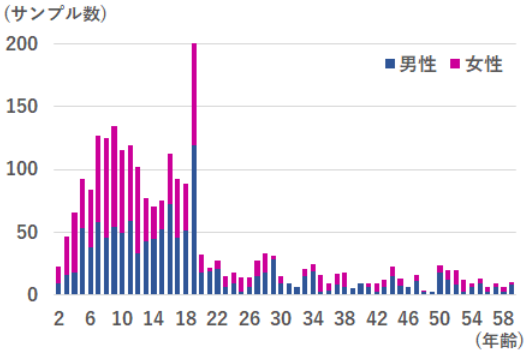

図 1 話者年齢ごとの発話サンプル数

ルールによって定められた未成年や成人を区分することとは異なるタスクとなっている．

また，男女で異なる声変わりの特徴を考慮するため，実験では，若年者・大人（男性）・大人（女性）の 3 クラス分類を行っている．

3 使用データセット

先行研究では，2 歳から 59 歳までの一般ウェブ利用者による発話のデータ 2,361 個を収集している．年齢別のデータ数の内訳を図 1 に示す．発話の内容は比較的短い一言の発言であり，時間では平均 3.6 秒，合計 2.4 時間となっている．各データは，サンプリング周波数 16kHz，量子化ビット数 16bit，モノラルの音響信号（WAV ファイル）である．

また，本研究では，環境変化に頑強なモデルを構築するためにデータ拡張を適用している．音声処理ツールキット Kaldi[3] を用いて，前述の音響信号に音楽やノイズ，話し声の雑音信号を加算し，ランダムに残響を重畳することで，合計 6,109 個にデータを増やしている．

後述の深層学習による実験では，拡張後の音響信号から抽出した振幅スペクトルを時間–周波数の 2 次元画像に展開したスペクトログラム（計 257 次元）を入力とする．

4 比較評価実験

4.1 実験条件・評価指標

実験では，年齢閾値を 9~18 歳に設定した場合の比較を行った．また，データ拡張の適用有無の影響

Identifying young speakers of crowdsourced speech based on deep learning
† Yuya Okumoto, Graduate School of Systems Engineering, Wakayama University
‡ Ryuichi Nisimura, Data Intelligence Education Research Division, Wakayama University

表 1　モデルごとの学習条件

学習条件	CNN モデル	LSTM モデル
epochs	20	
early stopping	3 epoch	5 epoch
batch size	16	8
learning rate	0.0001	
optimizer	Adam	
loss	Categorycal Crossentropy	

も比較している．

データ全体を 10 分割して，学習段階の訓練用 8 つ及び 1 つを検証用，残り 1 つを評価段階に使用した．10 分割交差検証によって，データを入れ替えながら 10 回学習段階と評価段階を繰り返すことで，すべてのデータを評価するようにした．結果に示す値は，10 回を平均したものである．なお，評価段階の使用データと同じ話者による発話は，学習段階（訓練・検証）からすべて除外した状態（話者オープン）で実験を行っている．また，検証・評価のデータには，拡張で追加されたデータは含まれない．

今回，比較の対象とした深層学習によるモデルは，CNN（畳み込みニューラルネットワーク）と LSTM（Long Short-Term Memory）である．各モデルの学習条件を表 1 に示す．

CNN は，畳み込み層 2 層と MaxPooling 層 1 層，全結合層 2 層で構成した．過学習対策のため，各全結合層の後に Dropout 層を追加した．

LSTM は，1 層のマスキング層，3 層の LSTM 層，2 層の全結合層で構成した．LSTM 層の 1，2 層目の後には Batch Normalization 層を追加した．

評価指標には，Accuracy(正解精度)，F 値 (F1-score, F-measure) を用いた．Accuracy は，推定結果のクラス正解率である．F 値は，Precision(適合率，若年者と推定した発話が正しく若年者による発話であった割合) と Recall(再現率，元のデータに含まれる若年者の発話を推定できた割合）の調和平均であり，式 (1) によって求められる．

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad (1)$$

4.2　実験結果

実験結果として，図 2（Accracy），図 3（F 値）を示す．各図中には，先行研究 [1] の GMM-HMM（特徴量 MFCC）による結果も示す．

GMM-HMM では，年齢閾値 13 歳のとき，Accuracy 0.69，F 値 0.77 であった．対して，本研究では，CNN（データ拡張なし），年齢閾値 18 歳のとき，Accuracy 0.74，F 値 0.84 となった．また，LSTM（データ拡張あり），年齢閾値 18 歳では，Accuracy 0.71，F 値 0.83 であった．

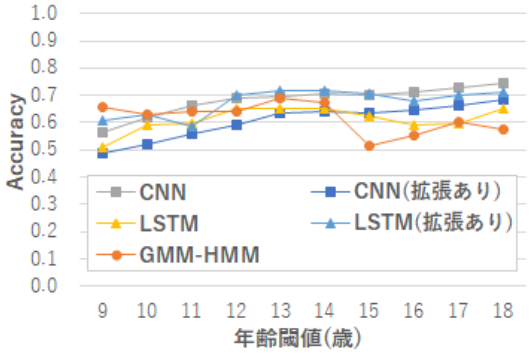

図 2　年齢閾値ごとの Accuracy

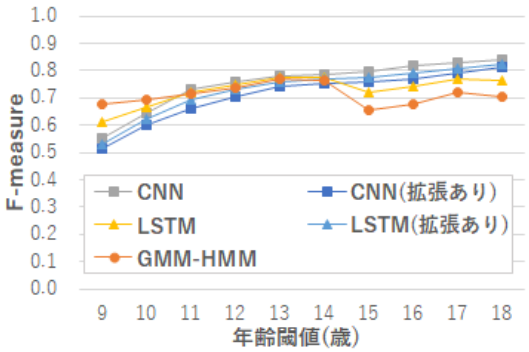

図 3　年齢閾値ごとの F 値

5　まとめと今後の課題

機械学習のアルゴリズムに深層学習を導入し，特徴量をスペクトログラムとした結果，CNN（年齢閾値 18 歳）のとき，Accuracy，F 値ともに，他と比べて高い値となった．深層学習を用いた場合，年齢閾値 15 歳以上で精度向上の傾向を示した．LSTM（年齢閾値 13 歳）の Accuracy が，GMM-HMM との比較で 0.03 高くなっており，本タスクでの深層学習の有用性を確認することができた．

今後は，話者認識等で使われており，話者固有の特徴を表現できると考える特徴量である i-vector や x-vector[2] 等の導入を検討する必要がある．

謝辞

本研究は JSPS 科研費 JP18K02862, JP21K12155 の助成を受けたものです．

参考文献

[1] 宮森ら, ちょっとした一言の音声認識による子ども利用者判別法の検討, 情報科学技術フォーラム講演論文集（FIT2010）, pp. 469–472, 2010.

[2] David Snyder, et al., X-vectors: robust DNN enbeddings for speaker recognition, Proc. ICASSP, 2018.

[3] 篠崎ら, Kaldi における CSJ レシピの利用法, 情報処理学会研究報告, vol. 2016-SLP-110, no. 8, pp. 1–6, 2016.