幼児の声色に基づく感情・行動の判別と可視化

樽見理花 (指導教員：伊藤貴之)

1　はじめに

2 歳程度の幼児は目を離すと高いところに登ろうとしていたり、おもちゃなどを食べようとしていたりと危険がつきものだ. そのため彼らの育児はひとときも目を離すことが出来ない. しかし育児をする中で別の家事を行ったり, 自分の娯楽を楽しんだりする時間も必要である. 近年, 乳児の泣き声から乳児の感情を予測するアプリ [1] がリリースされ, 言葉を話すことが出来ない乳児の泣き声による音声感情分類が可能となった. しかし, 2 歳程度の泣き声だけでなく言葉にならない喃語を話すようになった幼児の音声感情分類については未だ研究が少ない. そこで本稿では, 研究数が少ない 2 歳程度の幼児に着目し, 幼児の声色から幼児の感情・行動の推定を行い, 特徴量と機械学習モデルの比較を行うことで最適な特徴量と機械学習モデルを検討する.

2　関連研究

幼児の音声感情認識は近年発達してきた分野であり, 幼児の音声データを取得することが難しいことから, 古典的な機械学習手法が採用される場合が多い. 田中ら [3] は, 生後 2 日目から生後 5 日目までの新生児の乳児の泣き声を収録した DB を作り, それぞれの泣き声に乳児の状態を記したラベルを添付することで, 泣き声の音響的情報を分析できるツールを作成した. しかし, この研究はデータ数が少ないことから泣き声を用いた感情分類までには至っていない. Ashwini ら [4] は, 生後 1 日から生後 10 日までの新生児を対象とし, 空腹, 痛み, 眠気の 3 つの感情ラベルで計 300 の音声データを使用して, CNN-2D で特徴量を抽出し, SVM や RDF で判別する研究である. この研究により, 少ないデータ数でも深層学習を使用した特徴量抽出が可能であることが示された. 本稿では, 実験対象とする幼児の年齢が違うものの, この研究を参考に実験を行っている. 本稿で取り入れた特徴量については, 門谷ら [2] の音の 3 要素を考慮した特徴量を参考に選定している. この研究では, 機械学習を使用せず隠れマルコフモデルを採用することで成人の音声感情認識の実験を行っている.

3　提案手法

次の手順で実験を行う.

1. iPhone 搭載のカメラで 1 歳半程度の男児 1 名の行動の撮影を行い, 音声を抽出するなど前処理を行う.
2. 音響特徴量を抽出する.
3. SVM やランファムフォレストを使用し, 行動・感情を判別する.
4. 判別した結果を t-SNE を用いて可視化する.

3.1　音声データの収集と前処理

1 歳 10ヶ月から 2 歳 6ヶ月までの期間の男児 1 名を対象とし, 幼児の自宅で幼児が生後約半年から関わりのある大人が幼児と 2 人きりの状態で iPhone 搭載のカメラを用いて幼児の行動の撮影を行う. この時, 幼児とビデオカメラの距離は 2 メートル以内である. 撮影した動画を見ながら, 幼児の感情を推定する. アノテーションがついた動画 (mov) を音声データ (wav) に変換し, 2 秒ごとにデータを切り取る. 本稿では笑い声 (happy) と泣き声 (sad) の実験を行っている. 今後ラベルについては追加していく予定である.

3.2　特徴量の抽出

本研究は, 2 種類の手法で特徴量の抽出を行う. 1 つ目は成人を対象とした音声感情分類の研究と乳児の泣き声を対象とした音声感情分類の研究の論文調査を行い, それぞれで使用されている音響特徴量を独自に選定し, 組み合わせたものを使用する手法である. 本稿では, 文献 [2][3] より, 成人と乳児の研究の両者で採用されている mel と mfcc の抽出を行った. 2 つ目は音声データをスペクトログラムに変換し, 特徴量抽出器として CNN-2D モデルを用いることで特徴量の抽出を行った.

3.3　行動・感情の判別

本稿は, 笑い声 (happy) と泣き声 (sad) の分類を行い, 特徴量には前述にある通り, mel のみ抽出したもの, mfcc のみ抽出したもの, mel と mfcc を抽出したもの, そして CNN-2D モデルで抽出した特徴量, 以上の 4 パターンについて実験を行った. 使用した機械学習については, CNN-1D モデル, SVM(線型カーネル), SVM(ガウスカーネル), ランダムフォレストの 4 パターンについて実験を行った. 特徴量と機械学習の組み合わせのそれぞれの精度 (accuracy) についての比較を行った. また交差検証については, SVM(線型), SVM(ガウス), ランダムフォレストの 3 パターンについて実験を行った.

3.4　可視化

前章で判別した結果を基に, 正解データと不正解データに分けて t-SNE で可視化する. 以下に可視化する際のラベルを示す.

可視化ラベル

1：笑い声 (happy) 正解データ
2：泣き声 (sad) 正解データ
3：笑い声 (happy) 不正解データ
4：泣き声 (sad) 不正解データ

4　実行結果・考察

本章では実行結果から得られる知見について述べる.

4.1　判別結果 1-サーベイにより特徴量を抽出-

mel と mfcc を特徴量として抽出した際のそれぞれの判別結果を以下に記す.

表 1 の平均精度より, 実験した特徴量の中で mel と

表 1: 特徴量と機械学習モデルの比較.（単位：%）

	CNN-1D	SVM(線型)	SVM(ガウス)	RDF	平均精度
mel	70.9	63.64	67.27	81.82	70.91
mfcc	83.64	76.36	67.28	80.00	76.82
mel+mfcc	87.27	78.18	67.27	76.36	77.27
CNN-2D	-	90.91	-	-	90.91

表 2: 交差検証を行った結果の比較.（単位：%）

SVM(線型)	SVM(ガウス)	RDF
88.91	69.57	84.37

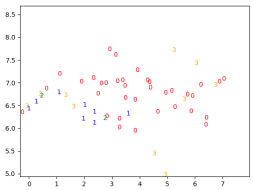

図 1: SVM(線型)

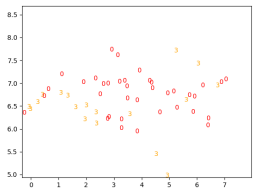

図 2: RDF

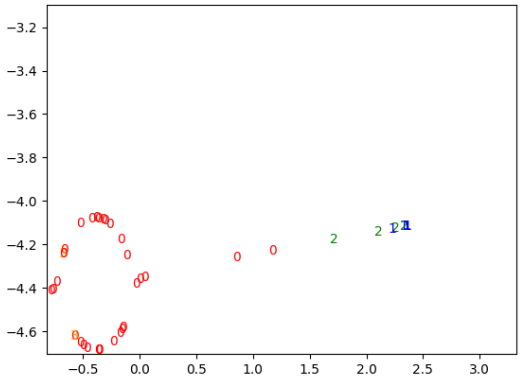

図 3: CNN-2D で特徴量を抽出, 機械学習モデルに SVM(線型) を使用した実験結果.

mfcc 両者を抽出したものが 1 番精度が良い結果となった. 機械学習手法については, 次元数が 1 番大きい mel+mfcc を使用した場合, CNN-1D モデルが最も精度の良い結果となった. しかし, CNN を適用するにしてはデータ数が 217 と少なく, 特徴量次元数も 168 と少ないため, 過学習の可能性を想定して, 最適な機械学習手法と断定は出来ないと考察した. また, 表 1 よりランダムフォレストも精度が高いと言えるが, 図 2 に示すように, 全ての推定結果が笑い声 (happy) となってしまい過学習の可能性が高いため, 最適な機械学習手法ではないと判断した. 一方で図 1 に示すように, SVM(線型) はテストデータに対して各ラベルが散布図上でほどよく分離されており, かつ表 2 に示すように交差検証の精度が 88.91 %と最も高かった. そこで現状では, SVM(線型) が最適な機械学習手法であるとする.

4.2 分析結果 2-CNN-2D により特徴量を抽出-

1 より, 前章のサーベイして選定・抽出した特徴量を使用する手法よりも, 高い精度が得られた. また t-SNE で可視化を行った際に, 各ラベル毎にまとまって可視化されているため, 本稿の実験の中では最も最適な音声感情分類手法といえる.

5 まとめと今後の展望

本研究では, 喃語を話す幼児を対象とし, 幼児の声色から感情を判別する際に最適な特徴量と機械学習について議論した. その結果, 笑い声と泣き声の判別において, CNN-2D モデルで特徴量を抽出し, SVM で感情を判別する手法が最も精度が高くなることを発見した.

今後の課題としては, 成人・乳児それぞれの音声感情認識に関する研究を参考に候補となる特徴量を抽出し, 機械学習モデルに SVM(線型) を使用して最も精度の良くなる特徴量を検討する. また, 笑い声・泣き声だけに限らず他の感情においても判別出来るかを検討するためラベルの追加を行いたいと考えている. さらに, 本研究に適用可能な学習済モデルを入手できたら, それを用いた転移学習についても検討する.

謝辞

本研究はお茶の水女子大学の生物医学的研究の倫理審査を受けたものです.

研究にあたり貴重な助言を下さったお茶の水女子大学心理学科の山田美穂准教授と砂川芽吹助教, ソニー CSL の大和田茂研究員に, またデータ収集でお世話になった対象児のご家族の皆様に感謝致します.

参考文献

[1] 株式会社ファーストアセント. パパっと育児@赤 ちゃん手帳. https://papaikuji.info/.

[2] 門谷信愛希 al. 特徴量の抽出. 音声による感情認識システムに関する研究, pp.18–20, 2000.

[3] 田中宏和 al. 乳児の泣き声の収集とその分析ツール. 第 66 回全国大会講演論文集, pp.281–282, 2004.

[4] Ashwini K, P.M. Durai Raj Vincent, Kathiravan Srinivasan, and Chuan-Yu Chang. (2021). Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. Front Public Health. 2021;9:670352. Published online 2021 Jun 10. doi:10.3389/fpubh.2021.670352