F-008

機械学習的手法による言葉に着目した音声データの感情分類の検討
A Study of Emotion Classification of Voice Data Focusing on Words Using Machine Learning

髙久雅史 †
Masashi Takaku

浦野昌一 †
Shoichi Urano

1. はじめに

近年、音声認識技術は音声翻訳や音声での検索サービスなど、生活の様々な場面で用いられており、作業の自動化や新しいサービスの提供に役立っている。その中でも、コールセンターの自動対応などの場面では、話者の感情を分析して行動の支援を行うための人工知能が搭載されている。しかし、感情の誤認識が発生し、話者の発言や感情を正しく認識することが出来ないといった課題がある。

筆者らはこれまでに、音声波形から得られるデータの活用方法の面での課題があることに着目し、特徴量であるフォルマントを出来るだけ多く用いて感情分類モデルの作成を行ってきた[1]。決定木とニューラルネットワークを用いて感情分類を行い、それらの精度の比較検討を行った。また、ニューラルネットワークの入力変数に対して決定木でモデル作成した際の変数重要度を考慮するなど、データの選択を工夫してきた。しかし、これまで作成した分類モデルでは、言葉の違いに着目しておらずモデル構築データと評価データの音声に込められた言葉によって精度が偏ってしまっていた可能性がある。

そこで今回は、音声データの言葉に着目して、それらの母音や言葉の長さなどを考慮した検討を行う。決定木とニューラルネットワークの各手法を用いて、どのようなデータの選択をすれば感情分類精度が向上するか検証を行う。

2. 分析手法

2.1 ケプストラム分析

音声認識には、音声に含まれる言葉や感情を分析する際に必要となるフォルマントと呼ばれる特徴量がある。音声の波形は声帯の振動の波形に、声道が持つ音響特性の波形が畳み込みされた形で形成されている。そのうち

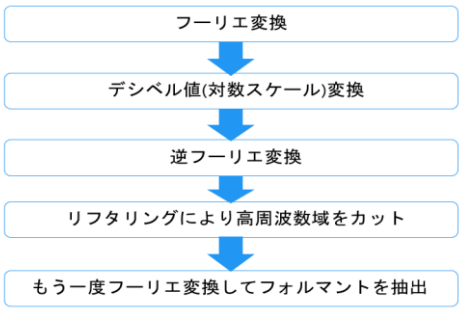

図 1 ケプストラム分析の手順

声道の音響特性の成分がフォルマントを決定づけており、二つの波の周波数の違いに着目して音声データから抽出するための分析手法がケプストラム分析である。音声データに対して図 1 に示す手順でケプストラム分析を用いてフォルマントの抽出を行う。

2.2 決定木

決定木とは機械学習的手法の 1 つであり、不純度が最小化される分割条件を決定し、同一クラスのデータで満たされるグループに分割できるようにする手法である。アルゴリズムは CART（ClassificationAndRegressionTrue）を用いる。このアルゴリズムでは、不純度と呼ばれるジニ係数などの評価関数を用いて目的関数の分岐を行う。決定木の処理の手順は以下の通りである。

(1) 不純度が最小となるデータの分割基準を決める
(2) (1)で決めた基準に基づいてデータを分割する
(3) 設定した基準になるまで(1)、(2)を繰り返す

2.3 ニューラルネットワーク

ニューラルネットワークは人間の脳の構造を数理モデル化した機械学習的手法の一つであり、パターン認識によく用いられるものである。入力層、中間層、出力層から成り立ち、それぞれの層に適当な数のニューロンが存在する。各ニューロン間での結合の強さが重みとして存在し、学習を繰り返す中で重みなどを更新し最適化する。今回は分類問題でのニューラルネットワークとして、出力層の活性化関数にはソフトマックス関数を適用し、重みの学習には AdaGrad を用いる。

3. シミュレーション

本稿では、6 つの感情(怒り、悲しみ、喜び、不安、落胆、平静)が込められた音声データをシミュレーション対象とする。それぞれの音声データから抽出した特徴量のフォルマントに対して決定木とニューラルネットワークを適用し、感情の分類モデルの作成と評価を行う。

3.1 データセット

本稿では、国立情報学研究所が設置する音声資源コンソーシアムが提供する慶應義塾大学研究用感情音声データベース（Keio-ESD）の音声データを用いる[2]。

今回は、6 種類の感情(怒り、喜び、悲しみ、落胆、不安、平静)のそれぞれ 20 単語ずつ計 120 個の音声データをシミュレーション対象とした。各感情 20 単語のうちそれぞれ 16 単語を使用してモデルの作成を行い、残りの 4 単語をモデルの評価に使用した。音声データの単語のリストを表 1 に示す。

† 明治大学先端数理科学研究科
MeijiUniversityAdvancedMathematicalSciences

表１音声データの単語リスト

あまがえるわ	あらわに	ながめ	おぼろずきよ
あまみず	あらゆる	なま	おもなが
あまみずわ	えもいわれぬ	なみ	おもうままに
あまのがわ	いわずもがな	ななめ	わらわれもの
あまりもの	みどり	なによりも	やわらげる

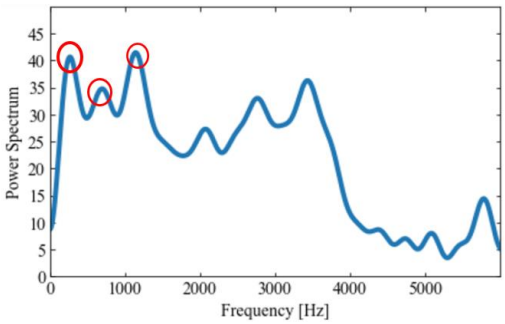

図２ケプストラム分析後の波形例

3.2 ケプストラム分析

音声データをケプストラム分析することによって、120 個全てのデータから特徴量であるフォルマントを抽出した。音声データのケプストラム分析後の波形例を図 2 に示す。図 2 の横軸がフォルマント周波数、縦軸が音の強さを表すパワースペクトルであり、波形の頂点が特徴量のフォルマントである。周波数の低いものから第 1 フォルマント、第 2 フォルマント、第 3 フォルマントと順に呼ぶ。今回、シミュレーションに用いた音声データにケプストラム分析を行った結果、全ての音声データに対して第 10 フォルマトまで抽出できたことから、10 個のフォルマント周波数とパワースペクトルを用いて感情分類を行った。

3.3 モデル作成

今回のモデル作成の際には、ケプストラム分析によって得られた特徴量を、120 単語全ての音声データから抽出された第10フォルマントまで用いてシミュレーションを行った。

決定木では木の深さを最大 6、データの分割指標をジニ係数とした。ニューラルネットワークでは 3 層の階層型とし学習係数は 0.01 に設定した。入力層のニューロン数は特徴量データの項目数である 20 とし、出力層のニューロン数は感情の種類の数である 6 とした。

3.4 シミュレーションフロー

筆者らはこれまでに、決定木とニューラルネットワークを用いて感情分類精度の比較検討を行うと共に、分類モデルの違いによる特徴を検証してきた。

今回は、表 1 に示す音声データの単語に着目して条件を変え、複数のパターンでのシミュレーションを行った。シミュレーションの手順を以下に示す。

(1) 120 単語分のデータセットを用意し、ケプストラム分析を用いてすべての音声データから特徴量であるフォルマントを抽出

表２シミュレーション結果例

	モデル１	モデル２	モデル３
分析手法	決定木	ニューラルネットワーク	ニューラルネットワーク
使用したデータ	抽出した全要素	抽出した全要素	重要度５%以上のフォルマント
正解率	６２.５%	５８.３%	６６.６%

(2) モデル作成用データの特徴量に対して決定木とニューラルネットワークを適用し、2 つの手法それぞれのモデルと、決定木の変数重要度でデータを絞ったモデルの合計 3 つのモデルを作成

(3) 評価用データを用いて 3 つのモデルの正解率を求め、それぞれの感情分類精度を確認

(4) (3)の結果から 3 つのモデルを比較検討し、分類モデルの違いによる特徴を検証

(5) (1)から(4)までの手順を、構築と評価のデータを入れ替えながら数パターンのシミュレーションを行う

尚、モデルの正解率は、作成したモデルに対して評価用データで正解した割合とした。

3.5 シミュレーション結果

評価用データを「あまのがわ」、「えもいわれぬ」、「おぼろづきよ」、「わらわれもの」の 4 単語で行ったシミュレーションの結果例を表 2 に示す。一文字目に含まれる母音や単語の長さなどの条件に注意し構築と評価のデータを入れ替えながら、他にも数パターンのシミュレーションを行ったうえで言葉と感情との関連性を検証する。

4. まとめ

本稿では、音声認識研究の課題である使用するデータの量や活用方法に注目し、音声データからケプストラム分析によって得た特徴量を第１０フォルマントまで活用した。また、決定木とニューラルネットワークによる分類モデルの作成と検証を行った。また、どのような言葉を話しているかという条件にも着目して、モデル構築と評価のデータを使い分けることで、言葉と感情の関連性の分析を行った。今後は、感情分類に適したデータの選択を行って精度向上を目指す。

謝辞

本研究では、国立情報学研究所音声資源コンソーシアムから提供を受けた「慶應義塾大学研究用感情音声データベース（Keio-ESD）」を利用した。

参考文献

[1]高久雅史・浦野昌一:音声データを用いたニューラルネットワークによる感情分類の検討,令和 4 年人工知能学会全国大会,1A4-GS-2-05 (2022)

[2]森山剛:慶應義塾大学研究用感情音声データベース(Keio-ESD),国立情報学研究所音声資源コンソーシアム(データセット), https://doi。org/10。32130/src。Keio-ESD (2011)