大規模対話音声コーパス作成を目的とするアンサンブル学習による人的コスト削減手法
a human cost reduction method by ensemble learning for the purpose of creating large-scale dialogue speech corpus

髙市 晃佑
Kosuke Takaichi .

広島市立大学大学院 言語音声メディア工学研究室
Language and Speech Research Laboratory, Graduate School of Information Sciences, Hiroshima City University

概要 近年, 深層学習の発展により大規模コーパスの必要性が高まっている. 対話音声から音声コーパスを作成する場合, 対話音声を話者ごとに切り出しアノテーションを行う. しかし, このタスクを人手で行う場合, 非常に負担がかかる. そのため, 音声コーパス作成支援の基礎研究とする. 本研究では, 音声コーパス作成支援として入力された音声が単数話者か否かを判定する. まず音声をスペクトログラムに変換する. その後, そのスペクトログラムが単数話者か否かの二値分類を行う. その判別にアンサンブル学習を用いることを提案する.

1. はじめに

近年, 深層学習に音声を用いる技術が発達している. しかし, これらの技術では大規模な音声コーパスが必要となる. 大規模音声コーパスは, 多種多様であることが求められる. 例えば, 年齢. 性別・使用言語等である. しかし現状の音声コーパスは, 特定の目的のために要素を選定しており, 全ての要素を満たす日本語音声コーパスは存在しない.

対話音声コーパスの作成に注目すると, 単数話者区間, 複数話者区間, その他周りの騒音等の区間がある. コーパスを作成する場合, それぞれの区間にアノテーションを行う必要がある. しかし, このアノテーション作業はコストが高い. コーパス作成には音声分割とともに分割音声の話者分類が必要であり, この作業が, 手動となれば前述の通りコストが高い.

本研究では, 上記の問題を解決するためコーパス作成支援のために入力された音声を複数話者と単数話者に分類する. この分類にアンサンブル学習を用いる.

2. 関連研究

本稿では, 関連研究の紹介を行う. 2.1 節では, アンサンブル学習の紹介を行う. 2.2 節ではアンサンブル学習を用いたシーン分類の先行研究について述べる.

2.1 アンサンブル学習

本研究で用いるアンサンブル学習の紹介を行う. アンサンブル学習とは, 複数の異なるモデルを独立して学習させ, 各モデルの最終的な結果を統合して評価する. 例えば単純な二値分類を行うと仮定する. この時, 通常の分類器で分類を行った場合, その分類器が誤分類したならば誤った結果が返されることとなる. しかし, アンサンブル学習の場合は複数の学習器を統合しているため, 複数の学習器が誤判定をしない限り誤分類することはない. このためアンサンブル学習が有効であると考えられる.

以下にアンサンブル学習の分類を示す.

・バギング

バギングとは, 学習データからブートストラップ法により弱学習器を複数構築し, それらを統合して最終的な分類器を構築する方法である.

・ブースティング

ブースティングとは, 逐次的に弱学習器を構築するアンサンブル学習アルゴリズムである. バギングと同様に学習データの一部を使用し, 弱学習器を複数構築し, それらを統合する. 違いとしてブースティングはバイアスを下げる特徴がある.

2.2 音響シーン分類

関連研究として, 音響シーン分類アンサンブル学習を用いた音響シーン分類が[Sakashita 18]によって発表されている. [Sakashita 18]は 9 つのネットワークをアンサンブルし実験を行った. それぞれの学習器で分類を行った場合よりアンサンブル学習を用いた分類がより良い性能を示した. しかし, 9 つものネットワークを学習する必要があった.

3. 提案手法

本研究では, 大規模対話音声コーパス作成を目的とした音声の分類を行う. そのために, 音声の振幅情報を用いアンサンブル学習することで分類することを提案する. この分類を行う際に使用する分類器は, 複数の弱分類器を組み合わせたアンサンブル学習を用いる. 画像分類を行うモデルには CNN や SVM, 数学的なアプローチとしては NMF などがある. これらで分類を行うと高い精度では分類を行うことができるが, 完全とは言えない. そこでこれらのモデルを組み合わせてアンサンブルすることでより高い精度が得ることができると考えられる.

アンサンブルするアルゴリズムについては検討している最中である. 現在は SVM と [Sakashita 18]が用いているランダムフォレストに注目している. SVM の採用理由としては, 本研究が二値分類を目的としているため高い精度が期待できることが挙げられる. ランダムフォレストに関しては, ノイズに強く精度についても期待できる. また, データ量が多くなっても高速で動くため採用している.

アルゴリズムに関してだけではなく, データセットについても変更していく. 複数話者のスペクトログラムは音声が重なっているが, どのタイミングで重なっているかはまちまちである. そのため前半部分が重なっている, 後半部分が重なっている, などに分けて学習を行いアンサンブルする. また, 音声が重なっているか否かで分類を行うと, 前半と後半で発話者が異なっているが, 音声は重なっていない場合に対応できなくなる可能性がある. 以上のような理由があり, データセットを内容により分割する. また, 使用する特徴量についても複数用意する. 後述する実験では Short Time Fourier Transform によって表現された音声波形から取り出された振幅情報を使用したが, openSMILE や WORLD からのデータセットを用意しこれらから作成されたモデルをアンサンブルしていく.

4. 実験

本稿では，音声を一定の長さに分割し，スペクトログラムに変換したそれぞれを分類する実験をする．入力された音声を Short Time Fourier Transform によって表現された音声波形より振幅のみを取り出し図 1 に示す．横軸は時間，縦軸は周波数を示し，この図をスペクトログラムと呼ぶ．このスペクトログラムが単数話者か否かを分類する．図 1 の左に示した単数話者の振幅情報と図 1 右に示した複数話者の図では，音声波紋の重なりに違いが見られる．このような特徴を用いることで分類ができるのではないかと本研究では考える．

4.1 節では実験条件について述べ，4.2 節ではアンサンブル学習を行わず，一つの学習器で分類した予備実験の結果を示す．

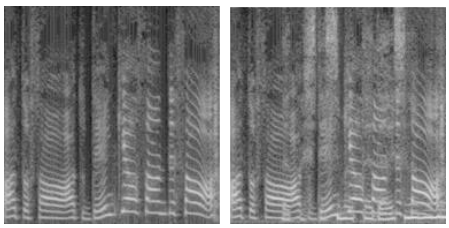

図 1　振幅情報　単数話者（左）と複数話者（右）

4.1 実験条件

本実験では，大規模対話音声コーパス作成を目的として，複数話者と単数話者の分類を行う．本実験の実験条件を以下で述べる．

・STFT のパラメータ

スペクトログラムを作成する際に使用した Short Time Fourier Transform のパラメータについて，本研究では，512 の長さに区切り，ステップ幅は 128 としている．窓関数はハミング窓を使用している．

・使用データ

本実験で使用した音声データは，二人の女性が話しているラジオ放送と別の女性が単独で話しているラジオ放送を使用している．それらの音声データを BGM や環境音が入らないように切り出し音声データとして使用した．また，切り出した音声を重ね合わせ複数話者音声として見立て使用した．表 1 で使用したデータ数を示す．すべてのデータは 256×256pixel，256 階調グレースケールによって表現されている．

・ネットワーク

本実験では CNN を用いて分類を行った．画像分類の手法として大きな成果が挙げられている．今回は深層学習と畳み込みニューラルネットワークの理解のため，AlexNet を使用した．

表 1　データ数の詳細

	データ数
train	9,500
validation	300
test	300

4.2 アンサンブル学習を行っていない学習器の結果

アンサンブル学習を試す前に一つの学習器で実験を行ってみた．accuracy と loss をそれぞれ図 2 と図 3 に示す．

5. まとめ

以上の結果より，単一の学習器であっても accuracy，loss 共に収束しつつあるという結果が得られた．しかし，図 2，図 3 をみると train よりも varidation の値がよくなっている．これはデータセットに偏りがある，もしくは CNN の重みに問題があったと考えられる．結果としては CNN によって分類を行うことができた．

6. 今後の課題

今後の課題として，アンサンブル学習を行った分類器を実装する．その際どのようなデータで学習を行うのかについて検討を行う．また，その結果との比較及び，本実験で用いていない音声データを実際に入力することも検討する．

アンサンブルするアルゴリズムやデータセットについて他に有用なデータがないか今後も検討していく．

参考文献

[Sakashita 18] Yuma Sakashita， Masaki Aono：acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions，Detection and Classification of Acoustic Scenes and Events 2018，(2018)

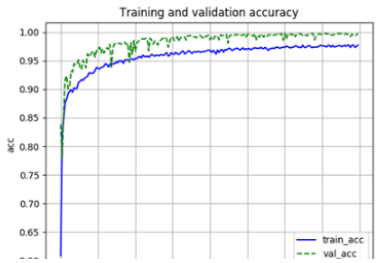

図 2　train,varidation 結果(accuracy)

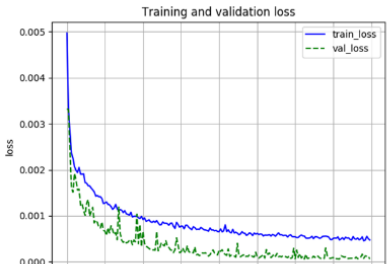

図 3　train,varidation 結果(loss)