6R-01

携帯電話の電話音声画像を用いた兄弟間の判別について

財満敦也† 菱田隆彰‡
愛知工業大学 情報科学部†‡

1 はじめに

近年，スマートフォンの普及により，携帯電話でのコミュニケーションが増えてきている．携帯電話の通話の音声は通信しやすいように加工されるため，似た声の人物の音声は，電話越しの聞き手にとって個人を判別することが難しい場合がある．

本研究では，携帯電話の通話時に筆者の声とその声に似た2人の弟の声を判別するため機械学習を活用した判別方法を検討する．電話音声画像を学習に用い，学習させるデータや学習済みモデルの出力を変化させて，精度比較を行う．

2 本研究の概要

ディープラーニングを用いた声による人物判別は，様々な手法が提案されている.文献[1]では学習データを拡張処理することで約 20 倍のデータを用意している．ディープラーニングによる識別において有用な学習データをいかに用意するかは重要な課題と言える．

本研究では，携帯電話音声という条件下で声の似た人物をより判別しやすくなるような条件を調べるために，いくつかの検証を行う．具体的には1つ目は母音の組み合わせによる学習データの拡張とその有用性，2 つ目は音声データのコーデックのパラメータの違いによる精度の違いを調べる．判別方法は文献[2]を参考に，録音した音声は携帯音声用のデータ形式に変換した後，画像に加工し学習データとして用いる．

2.1 母音を組み合わた学習データの拡張

我々は様々なフレーズの音声を用意するコストを低減させるために，母音の音声の組み合わせで様々なフレーズの音声に近い学習データを作成することで学習データの拡張を行う．

本研究では，判別対象となる筆者及び筆者の二人の弟に対し，“あ”，“い”，“う”，“え”，“お”の 5 つの音をそれぞれ 10 秒間録音する．録音した音声を先頭から2秒の長さに切り取り，全体が10秒になるようにそれぞれの音声を組み合わせることで音声データを作成する．全ての音の並べ替えを行った場合，一人あたり 120 通りの音声データを作成できる．

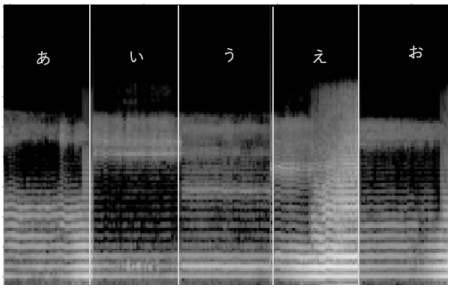

図 1:変換した音声画像

2.2 携帯電話の音声コーデック

本研究では録音データに携帯電話の音声と同じ加工を施すことで，より適切な学習データの作成を行う．

近年の携帯電話では AMR-WB (Adaptive Multi-Rate Wide Band)という音声コーデックが用いられる．AMR-WB はデータの圧縮率を選択可能で，6.60， 8.85， 12.65， 14.25， 15.85， 18.25， 19.85， 23.05， 23.85kbps の 9 種のビットレートが存在する．本研究では小さい方から 8 種の設定を用い PC で録音した音声から学習用の音声データを作成する．

2.3 音声のスペクトログラム画像への変換

音声の判別にはディープラーニングによる画像分類の技術を応用する．音声は音の周波数を縦軸，時間を横軸，値の大きさを色で表すスペクトログラム画像に変換し，その画像データを学習データとして用いる．実際に変換した例を図 1 に示す．値の大きさは明るい色ほど大きい．また，縦に 5 つの領域に分かれているように見えるのは，それぞれ異なる母音が組み合わさっているからである．今回，学習用に 8（ビットレート）×3(人)×120(通り)で 2880 枚，検証用には別途録音したデータから 312 枚の画像を用意する．

2.4 学習の概要

学習はファインチューニングという既存の学習済みのモデルを新たな学習に利用する方法を用いる．本研究では文献[3]を参考に画像分類で有名な学習済みモデル VGG16，VGG19 に音声画像

A study of a discrimination method between siblings using mobile phone voices
†ATSUYA ZAIMA,Aichi Institute of Technology
‡TAKAAKI HISHIDA, Aichi Institute of Technology

を入力し，その出力を画像の特徴量として線形分類器の SVM へ入力し学習させる．

3 精度と検証結果

本研究の精度比較は使用する学習済みモデルの種類，出力に使用するモデルの層，学習させる音声データのビットレートの組み合わせの 3 点を変化させて比較を行う．

使用する学習済みモデル VGG16，VGG19 はニューラルネットワークの層の数が異なる．一般的に層が深いほど分類の精度が高いと言われており，本研究における精度への影響を確認する．

モデルの層については，文献[3]によると，使用する層が後半になるほど，VGG が元々判別していた問題に適応しているため，他の問題については最終出力より途中の層の出力を利用した方が，精度が高くなる可能性が示唆されている.VGG は 5 つの畳み込みブロック(block_1～block_5)で構成されている．本研究では，ファインチューニングの際に block_4，block_5 のプーリング層の出力を利用する．

ビットレートの組み合わせについては，用意した 8 つのビットレートから 1 つから 3 つを選択し学習させて傾向を確認する．

表 1 にビットレート 8.85kbps に着目したモデルと利用したブロックの層の組み合わせごとの精度を示す．ビットレートは昇順に A，B，C，D，E，F，G，H と表記している．また，表 2 には 8 つのビットレートの中から 1 から 3 つのビットレートを組み合わせた学習データ，総組み合わせ数 92 パターンについての学習のモデルとブロックの組み合わせごとの全ての精度の平均，最大，中央値を示す．

表 1 から 8.85kbps(表 1 では B と表記)を含む 3 つのビットレートの組み合わせの学習は B のみの学習比べて良い精度を示した．表 2 から本研究の場合，VGG19 の block_4 の出力を利用した時に精度が最も精度が良くなり，VGG19 の block_5 の出力を利用した時に精度が最も精度が悪くなった．

4 まとめ

精度比較の結果から，本研究の機械学習の分類の精度は 1 つのビットレートだけの学習データでは精度が悪くても，複数のビットレートの画像をまとめて学習に用いると精度が良くなることが分かった．また，今回の結果においては，モデルの最終出力よりも途中の出力を用いた学習の方が良い精度が得られた．モデルの層の多い VGG19 は良い結果が残ったが，他の条件によっては大きな差にならないことがわかった．

本研究では携帯電話の電話音声を学習データとして VGG という学習済みモデルを使用した音声判別の方法を検討し，いくつかの学習方法について声の似ている兄弟の声の判別の精度を比較した．判別精度はおよそ 7 割であったが，学習データの組み合わせやモデルの使用方法によって精度が変化する事がわかった．学習データの効率的準備のために有効な結果が得られた．

表 1:8.85kbps に着目した精度

ビットレート	モデル　block			
	VGG16,4	VGG16,5	VGG19,4	VGG19,5
B	0.313	0.318	0.273	0.308
BCD	0.721	0.611	0.786	0.666
BCE	0.711	0.671	0.796	0.696
BCF	0.696	0.686	0.741	0.701
BCG	0.706	0.681	0.746	0.686
BCH	0.706	0.666	0.736	0.651
BDE	0.726	0.696	0.805	0.651
BDF	0.681	0.686	0.776	0.696
BDG	0.686	0.691	0.781	0.691
BDH	0.701	0.651	0.766	0.681
BEF	0.706	0.726	0.776	0.676
BEG	0.711	0.766	0.781	0.676
BEH	0.711	0.671	0.751	0.646
BFG	0.676	0.706	0.761	0.716
BFH	0.696	0.661	0.756	0.681
BGH	0.696	0.661	0.756	0.681

表 2:学習の精度の平均，最大，中央値

モデル block	VGG16,4	VGG16,5	VGG19,4	VGG19,5
平均	0.637	0.633	0.707	0.628
最大	0.726	0.766	0.805	0.716
中央値	0.646	0.631	0.706	0.631

謝辞

本研究の一部は，JSPS 科研費 JP19K12073 の助成を受けたものです．

参考文献

[1]NEC、声認証技術を強化、5 秒で個人を認識可能に,https://jpn.nec.com/press/20190219_01.html　アクセス日 2020-12-11.

[2]小林啓悟，松井孝典，福井大，町村尚，”CNN を用いたエコーロケーションコールによる日本産コウモリ類の種判別システムの開発”，情報処理学会，Vol.2019-MUS-123,No.62,pp.1-2,2019.

[3]画像分類の機械学習モデルを作成する(3)転移学習で精度 100%,https://techblog.nhn-techorus.com/archives/8352，アクセス日 2020-12-03.