# 大規模対話音声コーパス作成を目的とする振幅情報と位相情報に着目した複数話者と単数話者の区間分類

Section Classification of Multiple Speakers and Single Speaker Focusing on Amplitude Information and Phase Information for Creating Large-Scale Dialogue Speech Corpus

羽原 俊輔
Shunsuke HABARA

広島市立大学 言語音声メディア工学研究室
Language and Speech Research Laboratory, Hiroshima City University

**概要** 近年 End-to-End による合成音声システムの発達により，大規模音声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，このタスクは非常に負担がかかるため本研究では，音声コーパス作成支援として対話音声から複数話者と単数話者の区間分類を行う．特に，音声特徴量として従来手法として用いられる振幅だけでなく，音声の位相情報に基づいた音声特徴量を用い分類を行うことを提案する．加えて，従来の RNN による分類手法と CNN による分類手法を提案手法として挙げる．結果，複数話者と単数話者の分類は，音声位相情報を用いることで分類可能であり，CNN を用いることで分類精度が向上するという結果が得られた．

## 1 はじめに

近年，End-to-End の音声合成システム技術が著しく発達している．しかし，これらの合成システム構築に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，様々な要素で構成されている．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．従来手法では，話者認識によってアノテーションを行う研究や音声から目的の声のみを抽出する研究がある．しかし，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，上記の問題を解消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話者と単数話者の区間分類を行う．手法，音声特徴量に位相情報を用い CNN によって区間分類を行う．これにより深層学習に適用可能な大規模音声コーパス構築の基礎とする．

## 2 関連研究

本章では関連研究の紹介を行う．2.1 節では，音声の区間分類課題の先行研究について述べ．2.2 節では，分類に使用する DNN について述べる．

### 2.1 音声分割課題

従来研究では，DNN を使用した音声分割技術が[Garcia-Romero 17]によって発表されている．[Garcia-Romero 17]の実験では，電話音声を 2 秒ごとに分割し，40 次元のメル周波数特徴量を用いて Embedding を行う．[Garcia-Romero 17]は，i-vector を使用した従来手法による，2 段階の分析手順を用いることなく，DNN を用いた一連の学習によって作業の簡略化が行われたこと，Error rate の結果より従来の i-vector を用いた分類手法と同等程度もしくは超える性能を示したことが実験により述べている．

### 2.2 ネットワーク

従来研究では，音声を分析する際時間軸を考慮した処理が行われてきた．特にニューラルネットワークでは，RNN が用いられてきた，代表的な RNN として，LSTM[Hochreiter 97]や GRU[Cho 14]が挙げられる．

しかし，近年では主に画像分類課題で使われる技術として CNN が大きな成果を挙げている．CNN は，画像をあるピクセルの周囲を畳み込み特徴量抽出を行う DNN である．代表的な CNN として，2015 年 ImageNet の分類課題のコンペティションにおいて高い成果を挙げた Resnet[He 15]や ResNext[Xie 16]が挙げられる．

## 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区間分類を行う．そのために，音声特徴量に位相情報を用い CNN によって区間分類を行うことを提案する．

本研究の目的として，1 つ目に音声特徴量として従来の振幅を用いた手法と提案手法として位相を用いた手法を比較する．

2 つ目に，ネットワークによる比較である．音声特徴量を 2 次元画像として示した事により，画像分類の手法として用いられる CNN を用いて分類を行うことが可能であると考え，従来音声など時系列データの分類に用いる RNN との分類精度の違いについても比較，考察を行う．

### 3.1 振幅情報

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ている．従来このような音声特徴量表現をスペクトログラムと呼び．声紋分析等で使われている音声特徴量である．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，波紋が消えている箇所がある．本研究では，このような特徴が単数話者と複数話者の分類に用いることが出来るのではないかと考える．

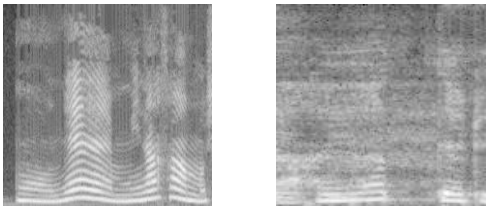


**図 1 振幅情報**

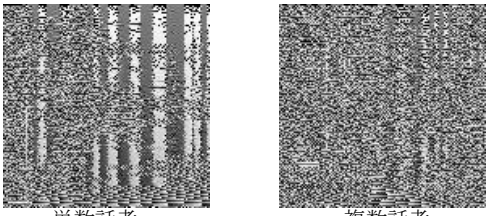


**図 2 位相情報**

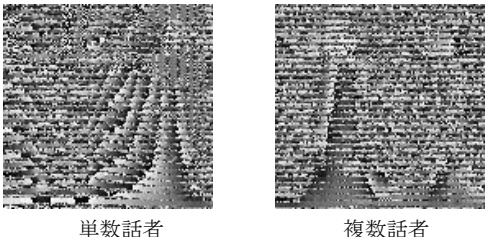


**図 3 位相の差分情報**

## 3.2 位相情報

### 3.2.1 位相

単位円範囲で示された位相は，-πからπの範囲で各値を表現することが出来る． 図 2 で示す 2 次元音声特徴量では，横軸が時間軸方向，縦軸に周波数方向を示している．位相情報では，図 2 の左側に示す単数話者では，縦縞がはっきり出ている箇所がある．しかし，図 2 の右側に示す複数話者では，縦縞が消えている箇所がある．このような特徴的な部分によって分類が行われるのではないかと考える．

### 3.2.2 位相の時間変換による差

先程（1）で示した位相情報では，基準点がランダムであり，圧一定の基準からの位相を表現しているとは言えない．また，[王 08]の研究では，異なる位相を揃える手法について述べられている．しかし本実験では，位相をアンラップすることによりこの問題を対処する．

単位円の中で位相は連続している，本特徴量ではある時間 t において t+1 の時間との位相の差分の大きさを縦軸に周波数，横軸に時間として表現した．図 3 の左の単数話者では，山なりである箇所があるが，図 3 の右の複数話者では，山なりである箇所が崩れている．よって，このような画像の崩れについて画像より検出出来るのではないかと考える

# 4 実験

本実験では，音声を一定の長さに切り出し，それぞれの区間がどの話者であるかを分類する課題を行う．4.1 節では実験条件について述べ．4.2 節では CNN を用いた分類結果を示す．4.3 節では，RNN を用いた分類結果を示す．

## 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認識を行う．音声特徴量として，2 次元音声特徴量を画像分類課題として使用し分類を行う．本実験の実験条件を以下で述べる

・ ネットワーク

本実験では，音声特徴量を 2 次元音声特徴量として扱う．よって CNN では，画像分類課題で大きな成果を上げている Resnet34 を用いた実験 RNN では， LSTM を用いた実験を行った．

・ 使用データ

本実験で使用した音声データは，2 人の女性話者が単体で話しているラジオ放送を切り出し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．

**表 1 データ数の詳細**

| | データ数 |
|---|---|
| train | 75,000（2,500） |
| validation | 1,000（3,000） |
| test | 1,000（3,000） |

## 4.2 CNN

・ 振幅情報

5 epoch ごとに，間引いた validation の結果を図 4 に示す．本研究の目的として音声コーパス作成を目的としているため precision の結果のみを記載する．図 4 より，epoch による変動が大きく分類精度にばらつきが出た． しかし，200epoch ほどで分類結果の最高値を示した．

・ 位相情報

同様に validation の結果を図 5 に示す．学習が進むごとに値が向上している．特に，振幅情報に比べ，epoch による変動が少なく安定している．

・位相差分情報

同様に validation 結果を図 6 に示す．200 epoch ですでに最高値を示しており学習が早く収束したといえる．また，epoch による変動も大きくなく安定した結果を示した．

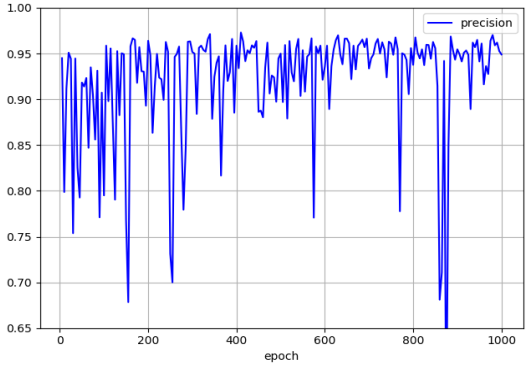


**図 4 CNN を用いた振幅情報の validation 結果**

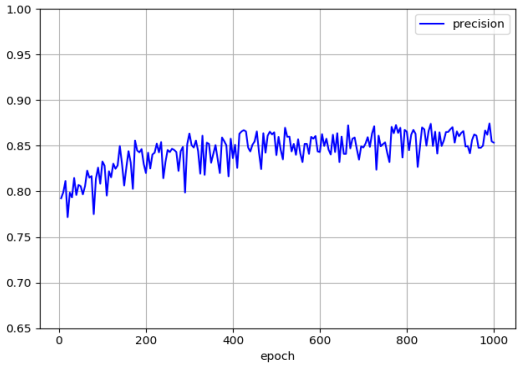


図 5 CNN を用いた位相情報の validation 結果

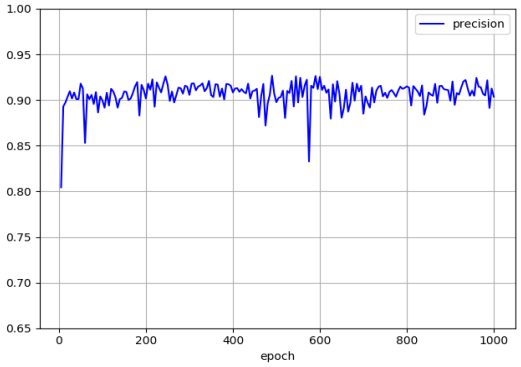


図 6 CNN を用いた位相差分情報の validation 結果

### 4.3 RNN

・ 振幅情報

validation 結果を図 7 に示す．こちらも 5 epoch ごとに間引いた結果を記載し，precision の結果のみ記載する．図 4 の CNN と比べ epoch による変動が少なく，安定した結果を示した．

・ 位相情報

同様に validation 結果を図 8 に示す．本実験の中では最も値が低く epoch によるばらつきも大きいためグラフの変動も大きい．

・ 位相差分情報

同様に validation 結果を図 9 に示す．結果は安定しており epoch によるばらつきも少ないことがわかる．

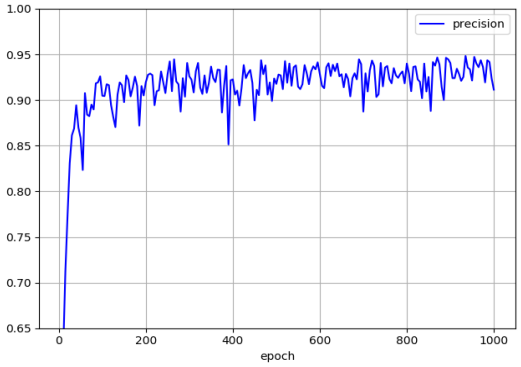


図 7 RNN を用いた振幅情報の validation 結果

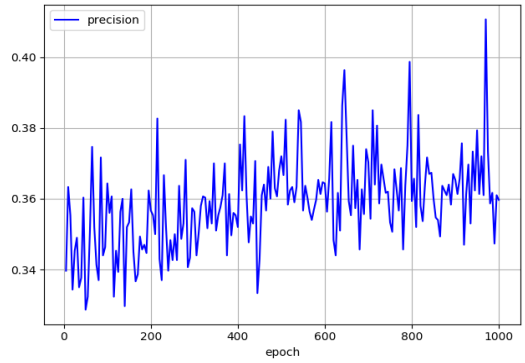


図 8 RNN を用いた位相情報の validation 結果

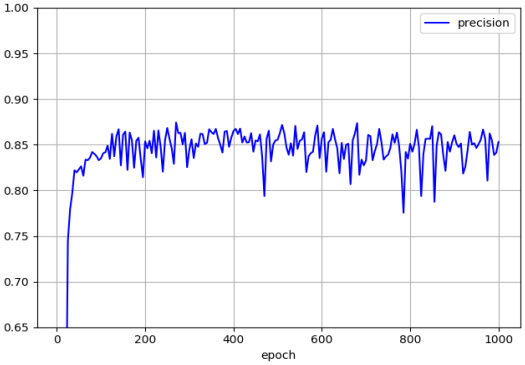


図 9 RNN を用いた位相差分情報の validation 結果

## 5 考察

実験結果より考察を行う．5.1 節では，特徴量による分析結果の差について 5.2 節では，ネットワークによる違いについて考察を行う．

### 5.1 特徴量による違い

音声コーパス作成のため単数話者と複数話者の分類課題を行った．各特徴量を使用し CNN によって分類した結果を表 2 に示す．また RNN によって分類した結果を表 3 に示す．どちらの結果も本実験での最終 epoch である 1000 epoch における Confusion Matrix の結果を示す．F1･F2 が各女性話者，Mix が 2 人の話者の音声を重ね合わせた音である．Precision を PRE として示す．

表 2 および表 3 より特徴量として振幅情報を用いた分類結果では，各予測クラスにおいて precision が 0.8 以上と高い数値を示した．これは，3.1 節で述べた特徴が大きく出ているためであると考える．

同様に表 2 および表 3 より特徴量として位相情報を用いたとき，振幅情報に比べ分類精度が良くない結果となった，これは，3.2 節で述べたような特徴が現れておらず分類が困難であったためと考える．しかし，特徴量を位相の差分とした実験においては，各クラスにおける precision は 0.8 以上と振幅と同等な高い数値をしめしており，位相の差分情報が十分話者分離課題として用いることのできる特徴量であることがわかる．

表 2 CNN による分類結果

| | | predicted | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 振幅 | | | 位相 | | | 位相差分 | | |
| | | Mix | F1 | F2 | Mix | F1 | F2 | Mix | F1 | F2 |
| actual | Mix | 919 | 21 | 60 | 885 | 59 | 56 | 907 | 28 | 65 |
| | F1 | 20 | 960 | 20 | 67 | 881 | 52 | 47 | 845 | 108 |
| | F2 | 27 | 20 | 953 | 105 | 125 | 770 | 39 | 44 | 917 |
| | sum | 966 | 1,001 | 1,033 | 1,057 | 1,065 | 878 | 993 | 917 | 1,090 |
| | PRE | 0.951 | 0.959 | 0.923 | 0.837 | 0.827 | 0.877 | 0.913 | 0.921 | 0.841 |

表 3 RNN による分類結果

| | | predicted | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 振幅 | | | 位相 | | | 位相差分 | | |
| | | Mix | F1 | F2 | Mix | F1 | F2 | Mix | F1 | F2 |
| actual | Mix | 796 | 121 | 83 | 162 | 736 | 102 | 841 | 81 | 78 |
| | F1 | 9 | 955 | 36 | 162 | 739 | 99 | 37 | 870 | 93 |
| | F2 | 11 | 84 | 905 | 164 | 672 | 164 | 28 | 143 | 829 |
| | sum | 816 | 1,160 | 1,024 | 488 | 2,147 | 365 | 906 | 1,094 | 1,000 |
| | PRE | 0.975 | 0.823 | 0.884 | 0.332 | 0.344 | 0.449 | 0.928 | 0.795 | 0.829 |

## 5.2 ネットワークによる違い

本実験では，空間的特徴量を用いる CNN と時系列を考慮するネットワークである RNN を用いて実験を行った．

まず，2 つのネットワークに対して振幅情報を用いた validation 結果を図 10 に示す．図中では，10 epoch ごとに間引いた accuracy の結果のみを示す．CNN では，epoch における変動が大きい．しかし，RNN では epoch によるばらつきが CNN よりは少なく安定した結果が得られた．validation 結果の最大値を見ると CNN が RNN より良い値を示していることがわかる．これは，振幅情報の特徴を CNN であれば RNN より捉えることができることを示している．

また，2 つのネットワークに対して位相の差分情報を用いた validation 結果を図 11 に示す．こちらも同様に 10 epoch ごとに間引いた accuracy の結果のみを示す．こちらは 2 つのネットワークにおいて epoch による大きなばらつきがなかった．しかし，accuracy の値より，CNN が RNN より高い値を示していることがわかる．これより，単数話者と複数話者の分類において，提案手法によって分類が行われることによって精度が向上することがわかった．

## 5.3 まとめ

以上の結果より，音声特徴量について位相の差分を示した情報で，複数話者と単数話者の分類は可能である．また．同一のネットワークで結果を比較すると，従来の振幅を用いた音声特徴量と同等に結果を示すことができた．また，分類を行うネットワークにおいて RNN と CNN を比較すると，3 種類の音声特徴量使用した実験全てにおいて CNN が RNN の分類精度を超えており，CNN を用いるほうが話者分類を行うにはより良いという結果が示された．

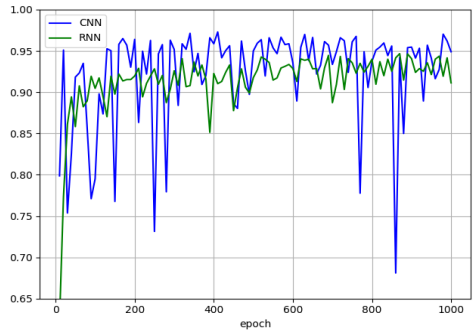


図 10 振幅情報を用いた RNN と CNN の比較

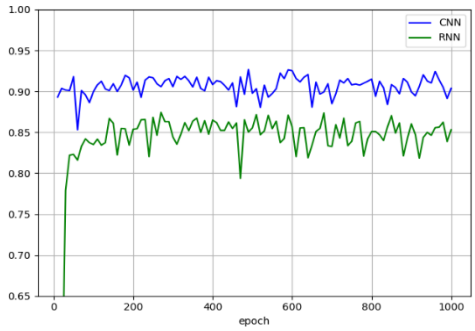


図 11 位相差分情報を用いた RNN と CNN の比較

## 6 今後の課題

今後の課題として，i-vector など既存の分類手法との比較や，今回使用した音声特徴量を用いた様々なデータを使用した話者分類を検討している．また，本実験で用いた分類システムを実際に構築することも検討する．

## 7 参考文献

[Cho 14] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio: Learning:Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, EMNLP 2014, (2014)

[Garcia-Romero 17] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, A. McCree: Speaker diarization using deep neural network embeddings, IEEE International Conference on Acoustics, Speech and Signal Processing, (2017)

[He 15] K. He, X. Zhang, S. Ren, J. Sun: Deep Residual Learning for Image Recognition, arXiv preprint, 1512.03385, (2015)

[Hochreiter 97] S. Hochreiter, J. Schmidhuber: Long Short-Term Memory, Neural Computation Vol. 9, pp. 1735-1780, (1997)

[王 08] 王 龍標，南 和江，山本一公，中川聖一: 位相情報を利用した話者識別・照合法の評価，第 10 回 音声言語シンポジウム，信学技報, vol. 108, no. 338, SP2008-108, pp. 173-178, （2008）

[Xie 16] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He: Aggregated Residual Transformations for Deep Neural Networks, CVPR 2017, (2016)