ISSN 2186-5647

P-55

ゼミナールでの発言度合い調査のための UIS-RNN による話者識別

日大生産工(院)　○秋田智希　　日大生産工　吉田典正

1. まえがき

近年，双方向型授業が話題になっている．通常の講義では，受け身となってしまっている学生に対して，双方向型授業では，教員の方から何らかのコミュニケーションをとることで，授業への関心を引き起こし，学生自身に問題や課題を発見させ，自主学習へのきっかけを与えることができる[1]．ゼミナールでも，双方向型となることで，より良いゼミナールへと改善していける可能性がある．そのため，ゼミナールでの発言状態をリアルタイムで判別できるシステムが望まれる．本研究では，まず男女 4 人ずつそれぞれの 15 分程度の音声による話者識別とゼミナールの音声による話者識別を，UIS-RNN と深層学習による画像分類を用いて行い，正解率の比較を行う．

2. UIS-RNN[2]

Unbounded Interleaved-State Recurrent Neural Networks (UIS-RNN)は，話者識別のための深層学習によるネットワークである．従来の録音後行うクラスタリングベース手法による話者の識別エラー率は，8.8%であるのに対して，この手法ではリアルタイムの計算で，7.6%の結果を出している．

UIS-RNNは，話者ごとの特徴を抽出したd-vectorと呼ばれるニューラルネットワークの隠れ層の出力を入力データとして学習を行う．

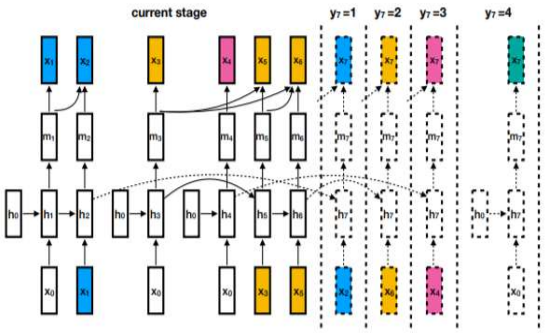

Fig.1 UIS-RNN の概要図

まず，UIS-RNNは，Fig.1のように対象ごとにネットワークを作成する．これによって初期条件として話者の人数を設定する必要をなくしている．その後，話者データとしてd-vectorを入力すると，そのデータから次のデータで話者が変わるのかどうかを予測する．予測の結果，話者が変わっていた場合には，違う話者のネットワークに変更し再度計算を行う．この作業を繰り返し，話者の識別を行う．

3. 実験方法および測定方法

男女 4 人ずつ，計 8 人の音声データと，研究室のゼミナールの音声を用意した．音声データは，サンプリング周波数は，44.1kHz，量子化ビット数は，16bit である．データを 1 秒ごとに分け，話者の正解ラベルを付けた後，スペクトログラムを作成した．スペクトログラムとは，横軸を時間，縦軸を周波数，周波数の強さを色で表したものである．作成したスペクトログラムを入力データ，正解ラベルを出力データとしてFig.2のネットワークを用いて学習を行った．その後，隠れ層を出力し，d-vector を作成した．作成した d-vector を入力データ，出力データをラベルとして UIS-RNN で学習を行った．学習後，学習に用いていないデータを使って，話者の予測を行い，正解率を計算した．また，Fig.2 のネットワークを用いて，スペクトログ

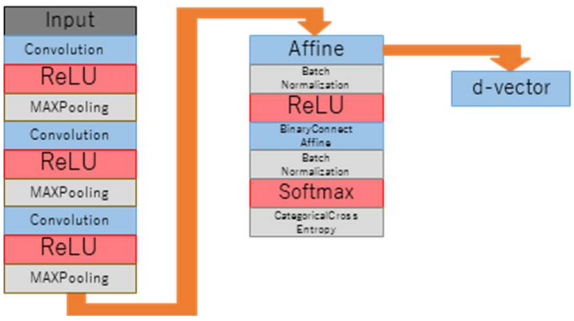

Fig.2 d-vector 作成に用いた画像分類のネットワーク図

Speaker Identification by UIS-RNN for Evaluating the Level of Speaking Activity in Seminar

Tomoki AKITA, Norimasa YOSHIDA

ラムを入力, ラベルを出力として話者識別を行った．そして，UIS-RNN での結果と画像分類での結果を比較した．

8 人の音声データによる話者の識別では，話し合っている状況の再現を行うため, 同じ話者が長時間続かないように順序をランダムに入れ替えた．また，スペクトログラムを作成する間隔を 3 秒，5 秒にし，同様に計算を行った．

ゼミナールでの音声では，3 秒や 5 秒の場合，その時間の中で，複数の話者が話している場合が多くみられた．そのため，1 秒のみで計算を行った．

4. 実験結果および検討

8人で話者識別を行った結果を，表1に示す．

表 1　8 人での話者識別の正解率

	1 データ当たりの秒数		
手法	1秒	3秒	5秒
UIS-RNN	75.0%	70.5%	85.7%
画像分類	84.4%	89.3%	98.2%

表1でUIS-RNNと画像分類の結果を比べると，どの秒数でも画像分類の結果の方が，正解率が高い．これは，実験を行うときに，同じ話者が長時間続くのを防ぐために行った, 順序をランダムに入れ替えたことが問題だったのではないかと考えられる. RNNは, 時系列データの特徴をうまく学習できるようになっている．そのため, 順序をランダムに入れ替えたことで系列データでなくなり, 学習がうまくいかなかったのではないかと考えられる．

表1より，どちらの結果でも，スペクトログラムを作成するのに用いる秒数が，5秒のときが一番高い正解率になっていることがわかる．5秒のときでは, 1秒や3秒のときと比べて, 1データ当たりの情報量が多く学習がしやすいと考えられる．また，5秒と長いため，1枚のスペクトログラムの中に系列データとしての特徴が含まれていた可能性がある．

ゼミナールの音声による話者識別についての結果を表2に示す．

表 2 ゼミナールの音声での話者識別の正解率

手法	正解率
UIS-RNN	38.0%
画像分類	50.0%

ゼミナールの音声での話者識別では，UIS-RNN, 画像分類両方とも, 正解率が低いことがわかる. ゼミナールでの音声は, 人が多いため, 声のほかに雑音が含まれることや, ゼミナール内での話す量によってデータ数に差があることが問題だと考えられる．また，d-vectorを作成する際に使用したニューラルネットワークは，画像分類であり，順序は変えていないが, 系列データとしての特徴を踏まえての学習はしていない. そのため，系列データとしての情報が失われた状態で，UIS-RNNの学習を行ってしまい, 学習がうまくいかなかったのではないかと考えられる．

5. まとめ

今回の実験では, 男女4人ずつ計8人のそれぞれの15分程度の音声を用いて，UIS-RNNと画像分類で話者認識の結果の比較を行った. また, ゼミナールの音声を用いて，UIS-RNNと画像分類での話者認識の結果の比較を行った．

男女4人ずつの計8人で行った話者認識では, 画像分類による話者認識のほうが正解率は高くなったが，5秒では85.7%と高い正解率を出せた．そこで，より系列データの特徴を踏まえたほうが良いと考え，ゼミナールの音声では, 順序を入れ替えずに，話者識別を行った．画像分類では, 正解率が50%以下となり, UIS-RNNにおいては，38%と低い結果となった．しかしながら，系列データを使用し，系列データとしての特徴をより多く学習させることにより, 精度を高められる可能性を見出した．

今後は, 話者識別の精度を上げていきたいと考えている．そこで，d-vectorを作成する際に使用するネットワークをLSTMやRNNなどの系列データの特徴を学習できるものを使い, 話者識別を行う．また，画像認識の方でも，ネットワークの変更や, パラメータの調整を行い精度の向上を行っていきたいと考えている．

参考文献

1) 木野茂 京都大学高等教育研究,「教員と学生による双方向型授業 ―多人数講義系授業のパラダイムの転換を求めて―」, (2009) https://core.ac.uk/download/pdf/39229474.pdf (参照 2019-10-9)

2) Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, Chong Wang: “Fully Supervised Speaker Diarization” , (2019) https://arxiv.org/pdf/1810.04719 Pdf （参照 2019-10-09）