音声の時間周波数表現に対する嗄声判定の試み - Transformerベースの画像認識の応用 -*

○石原一樹，荒井隆行（上智大）

1　はじめに

GRBAS 尺度は，嗄声の重症度を表すものであり，Grade, Rough, Breathy, Asthenic, Strained で構成されている．それぞれの指標は4段階で評価され，0は正常または症状が見られない，1は少し症状がある，2は中程度の症状，3は重度の症状である．Grade は嗄声の重症度の全体を表す指標である．GRBAS 尺度は持続的な音声の特徴を表すものであり，持続母音に対しての使用が適切である．これは，比較的均一な音声の性質が一定時間得られ，評価しやすいためである[1].

しかし，GRBAS 尺度の評価は専門のトレーニングを積んだ人でないと評価できない．また，その評価は主観的であり，聴者のバイアスや習熟度に左右されるという課題もある．この課題を解決するために，機械学習を用いて GRBAS 尺度の G の推定の研究が行われてきた．Sáenz-Lechón ら[2]は，Mel-frequency Cepstral Coefficients (MFCCs)を特徴量とし，学習ベクトル量子化で分類を試み，68%の正解率であった．Xie ら[3]は，MFCCs, Smoothed Cepstral Peak Prominence (CPPS) and Long-Term Average Spectrum (LTAS) を特徴量として使用し，Deep Belief Nets で，最大 81.53%の正解率であった．

本研究では，画像認識のモデルとして Vision Transformer (ViT) [4]を使用し，GRBAS 尺度の自動推定を試みた．ViT は，画像を小さなパッチに分け，各パッチをベクトルに変換したものを入力情報として使用する．本研究では，GRBAS 尺度を付与されている各音声サンプルからスペクトログラムを生成し，画像認識による G 尺度の自動推定を試みた．

2　実験方法

2.1 音声データ

本実験では，長母音の/a/の音声サンプルのみを使用した．データセットとして，動画で見る音声障害[5]の音声データを使用した．このデータセットは，GRBAS 尺度がラベル付けされた 65 人分の音声データが収録されており，/a/, /i/, /u/, /e/, /o/の発話がそれぞれ少なくとも 2 回分含まれ，日本語の連続発話も収録されている．音声データは，すべてサンプリング周波数 44.1 kHz, 16 bit である．本研究では，G 尺度の推定を試みた．内訳は，G 0 が 2 人，G1 が 17 人，G2 が 26 人，G3 が 20 人であり，各被験者の/a/の発話を 2 回分使用した．本研究では，G0 の音声が少なかったため，Saarbrücken voice database (SVD)[6]の健常音声の中で，Cepstral Peak Prominence の値が大きい 30 サンプルを G0 として使用した．

表 1. 使用した音声サンプルの内訳

G class	動画で見る音声障害	SVD	total
0	4	30	34
1	34	0	34
2	52	0	52
3	40	0	40
total	130	30	160

2.2 スペクトログラムの生成

本実験で使用する音声データは，サンプルごとに音声の長さが異なるため，各音声サンプルの中央部分から前後 250 ms を切り出し，中央部分の 500 ms を使用した．

窓のパラメータを変化させて実験を行い，オーバーラップ率を 50%として，窓の長さを 10 ms, 20 ms, 40 ms, 100 ms のパラメータで，それぞれスペクトログラムを生成した．これらのスペクトログラムは，学習前にバイリニア補間法を用いて 224×224 にリサイズした．図 1 は，GRBAS 尺度の各 G のスペクトログラムの一例である．G の値が大きいほど，高周波数帯域でのパワーが大きくなり，また，時間方向に周波数成分が細かく揺らいでいることも確認できる．

2.3 機械学習の手法

本実験では，ViT のモデルとして

* An attempt to determine hoarseness for time-frequency representation of speech: Applications of Transformer-based image recognition, by Kazuki ISHIHARA and Takayuki ARAI (Sophia University).

ViT-Large-32 を使用した．全データの中で30%をテスト用とし，残りのデータの中で，20%を検証用データとし，ホールドアウト検証を行った．学習はバッチサイズ 2 のミニバッチで行い，エポック数を 20 とした．モデルの最適化手法は Adam を使用し，学習率は 0.00005 とした．

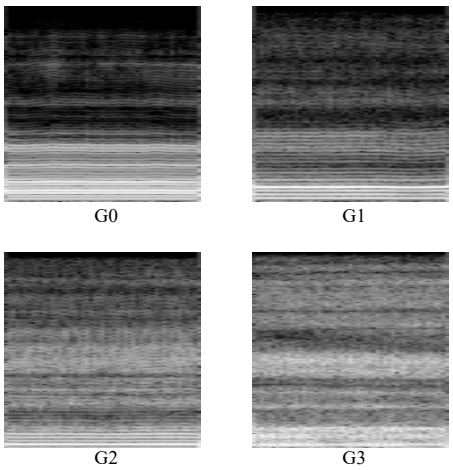

図 1. G 尺度のクラス毎のスペクトログラム

3　実験結果

3.1 実験 1

窓の長さが 10, 20, 40, 100 ms での正解率を求めた．窓の長さが 20 ms のときに，正解率が 75%と最も高くなった．

表 2. 実験 1 の分類結果

窓の長さ (ms)	10	20	40	100
Accuracy	0.625	0.750	0.688	0.719

3.2 実験 2

窓の長さが 20 ms と 100 ms のスペクトログラムをどちらも学習させ，学習するデータの数を増やした．この時，窓の長さが 20 ms, 100 ms, どちらも組み合わせた場合の 3 つのテストデータでテストを行った．結果としては，20 ms のテスト用のスペクトログラムに対する正解率は 71.9%となり，100 ms のテスト用スペクトログラムに対して 75%となった．図 2 の通り，全ての G のクラスで精度が 70%ほどとなり，更なる分類性能の向上が必要であると考えられる．

表 3. 実験 2 の分類結果

窓の長さ (ms)	20	100	20 & 100
Accuracy	0.719	0.750	0.734

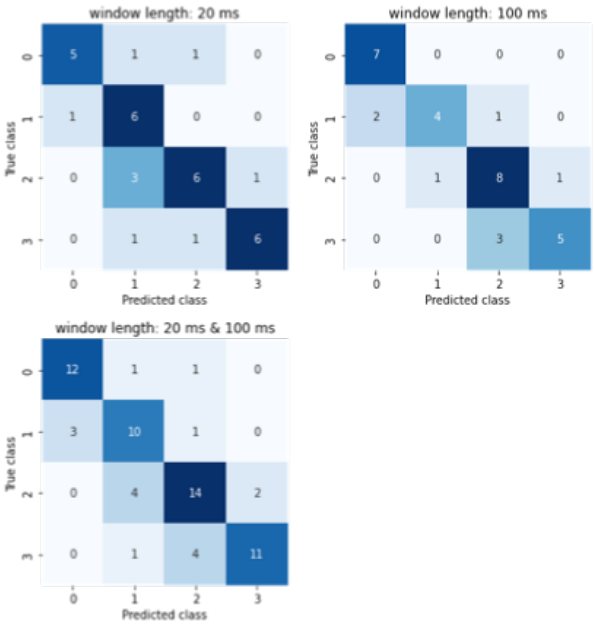

図 2. 実験 2 の分類結果の混同行列

4　おわりに

本研究では，最大 75%の正解率で G 尺度のクラスを推定できた．しかし，GRBAS 尺度が付与された音声データの数に限りがあるという課題があった．そのため，学習データが少ない場合にも有効な半教師あり学習を用いた手法を検討したい．

謝辞　上智大学重点領域研究の助成を得た．

参考文献

[1] 日本音声言語医学会，声の検査法 第 1 版，医歯薬出版，1979.
[2] N. Sáenz-Lechón et al., EMBS'06, 28th Annual International Conference of the IEEE, pp. 2478–2481, 2006.
[3] S. Xie et al., Interspeech, San Francisco, pp. 2656–2660, CA, USA, Sep. 2016.
[4] L. Yuan et al., arXiv preprint arXiv:2101.11986, 2021.
[5] 日本音声言語医学会，動画で見る音声障害 ver.1.0(DVD)，インテルナ出版，2005.
[6] B. William and P. Manfred, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/, 2007.