The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

機械学習を用いた構音障害分類における音声特徴量間の耐ノイズ性能の比較

An anti-noise performance comparison between acoustic features in detecting voice pathology using machine learning

鈴木 康大[*1*2] 篠原 修二[*2] 馬目 信人[*1*2] 朝長 康介[*1*2] 光吉 俊二[*2]
Kouta Suzuki Shuji Shinohara Nobuhito Manome Kosuke Tomonaga Shunji Mitsuyoshi

[*1] ソフトバンクロボティクス株式会社 SoftBank Robotics Corp.
[*2] 東京大学大学院工学系研究科 Graduate School of Engineering, The University of Tokyo

Developing communication robots requires to analyze human voice including various kinds of human biological information because the nonverbal information plays an important role in smooth communications between humans and robots. To analyze numerous voices available via the robot by using machine learning, we should take consideration of the existence of noises added to the voices. However, some acoustic features used for sensing human biological information is not designed for the noises. To validate the variation of the accuracy of classification when the voices includes the noises, we compare the classifications using voice indexes proposed for voice pathology estimation and using Mel-Frequency Cepstrum Coefficients(MFCC) in the classification problem of voice pathology as an early study. Experimental results show that classification using MFCC can detect voice pathology more precisely despite the noises while other voice indexes are adversely affected by the noises.

1. はじめに

人とのコミュニケーションを目的としたロボットが円滑に行動するためには，非言語情報を読み取ることが重要である．人から発せられた音声には多彩な生体情報が含まれており，健康状態や精神衛生状態を分類する試みが存在することから[光吉 2011] [Tokuno 2015] [Shinohara 2017]，声はコミュニケーションの際に重要な役割を果たすと考えられる．従って，コミュニケーションロボットの開発の際には，人から発せられた音声の解析が重要となる．

ロボットから取得される音声を解析する手法として，機械学習が挙げられる．機械学習による音声分類の大まかな流れは以下の通りである．まず分類したい状態に対応する音声をある程度用意し，それぞれの状態における音声の特徴量を抽出する（前処理段階）．次に同じ状態の音声に現れる特徴量のパターンを見つける（学習段階）．このようにして状態ごとに得られたパターンを使って，新たに得られた音声に対してパターンマッチングを行い，最も適合するパターンに対応する状態を，その音声の状態と分類する（分類段階）．

ここで学習させる際の音声と分類したい音声の取得環境が異なる場合，人の状態が同じでも異なる特徴のパターンが見られる場合があり，分類性能に影響を及ぼす可能性がある．特に，ロボットなどを介して取得された音声の場合，音声を取得する周囲の環境を制御することが困難なため，周囲のノイズが音声に含まれることを考慮する必要がある．音声を用いて人の状態を分類することを目的とした機械学習において，いくつかの音声指標を導出して状態を分類することが提案されているが，これらは音声にノイズが含まれることを考慮していないものが多く，音声の一部の特徴しか用いないため，含まれるノイズ次第では指標に大きな影響を与える可能性がある[髙野 2017]．

一方，近年計算機資源が増大することにより，深層学習をはじめとして膨大なデータを使った機械学習手法による分類が大きな成果を出している．これらは，得られた音声からヒューリスティックに指標を導出するのではなく，音声の様々な特徴を捉える分類を可能にする．ここで音声の特徴を表す普遍的な特徴量として，メル周波数ケプストラム係数(MFCC)が挙げられる．MFCCは音声認識でよく使われる特徴量であり，従来の提案されてきた指標よりも広い範囲で音声の特徴が含まれている．従って，音声にノイズが含まれている場合でも正確な分類が行えることが期待される．

本稿では，初期検討として，ノイズが含まれる環境下で音声から構音障害の有無を分類する問題を考え，音声からヒューリスティックに導出される指標と MFCC による分類を比較し，ノイズの強さ又は種類による分類性能の変動を検証した．

連絡先：鈴木康大，ソフトバンクロボティクス株式会社/東京大学大学院工学研究科，suzuki@coi.t.u-tokyo.ac.jp

2. 関連研究

音声から構音障害の有無を分類するために，健常者の音声と比較して構音障害と診断された人のみに現れる音声指標がいくつか提案されている．Teixeira らは，音声長母音から導出される jitter，shimmer，HNR(Harmonic to Noise Ratio)が構音障害を持つ人と健常者で異なる分布を持つことを明らかにし，これを用いて構音障害を分類することを提案している[Teixeira 2014]．また，篠原らは，構音障害を持つ人の長母音からピッチを検出しづらいことを踏まえ，ピッチレートを用いて構音障害を分類することを提案している[Shinohara 2016]．

構音障害分類用に音声指標を設計するのではなく，音声特徴量を上手く使って機械学習を行い，構音障害を分類する試みもなされている．Londono らは，MFCC に加えて，音声波形から得られるリアプノフ指数や HNR を特徴量として，ガウス混合分布(GMM)及びサポートベクターマシン(SVM)で構音障害を分類する手法を提案し，マサチューセッツ工科大学が提供する構音障害データベース(MEEI)で 98.23%の分類性能を出した[Londoño 2011]．Fang らは，MFCC のみを用いて深層学習で構音障害を分類することを提案し，Far Eastern Memorial Hospital が取得した構音障害データベースで，94.26%, 90.52%の分類性能を男性，女性の音声に対して出した[Fang 2018]．ALHUSSEIN らは音声をスペクトログラムに変換し，これを画像として扱い既存の画像分類用モデルである VGG16 や Caffenet の一部分を使って学習し，先述の MEEI や saarbrücker が提供する構音障害データベース(SVD)で 93.9%の分類性能を出した[Alhussein 2018]．いずれの場合も音声にノイズが含まれていな

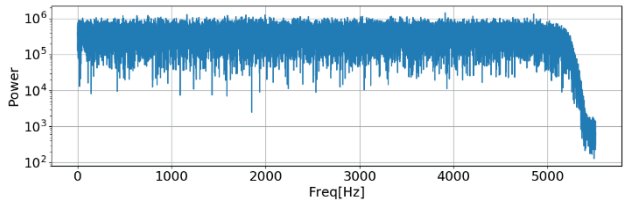

図 1. ホワイトノイズの周波数領域のパワー成分の分布

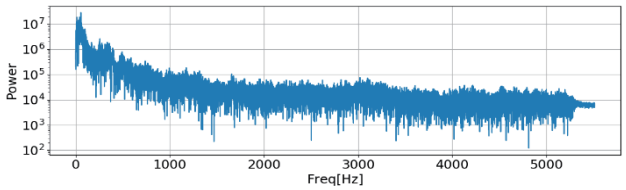

図 2. 車の走行音の周波数領域のパワー成分の分布

い場合に対しての性能であり，これらが外部からノイズが混入した場合の性能については改めて検討する必要がある．

3. 実験概要

3.1 使用した音声データセット

今回，構音障害データベースとして，先述の MEEI を利用した．音声のサンプリングレートは 11kHz で，モノラルチャンネルで録音されている．このデータベースから，[Llorente 2006]に挙げられている，健常者 53 人，構音障害者 173 人の長母音（ah の発音）を実験に利用するデータセットとして使用した．このデータセットは，健常者の発音は 3 秒間，構音障害者の発音が 1 秒間であり，年齢，性別などの偏りが健常者と構音障害者で同一であるため，健常者と構音障害者のデータ分布の偏りを最小限に抑えられている．このデータセットに含まれる音声を，ffmpeg-normalize を用いて音量の正規化を行った．ここでは，欧州放送連合が勧告している，EBU R128 に 従った方法で行った．

3.2 付加するノイズ

先述のデータセットを 1:1 に分け，一方の音声にはノイズを付加させず，MFCC を用いた機械学習による訓練用のデータセットとして使用した．もう一方の音声は，ノイズを付加させて分類性能を評価するためのテスト用のデータセットとして使用した．付加するノイズの種類は 2 種類用意した．一つはホワイトノイズ（全周波数帯におけるレベルが均一なノイズ）で，もう一つは車の走行音（低周波におけるレベルが高いノイズ）である．各ノイズの周波数領域でのレベルを図 1 図 2 に表す．ノイズの音源に対しても音声と同様の手法で正規化を行った．この正規化されたノイズ音源のレベルを変化させ，音声に付加させた．付加するノイズのレベルは 5 種類用意し，2×5=10 種類の音声を用意した．

3.3 音声指標による構音障害分類

今回使用する音声指標として，jitter, shimmer, HNR を採用した．構音障害を持つ人の音声の特徴として，音声波形に乱れが生じることがある．この乱れ具合を定量的に評価できる指標として，jitter, shimmer, HNR が挙げられる．jitter は時間軸方向に発生する非常に短い変動の成分を表し，shimmer は音声の振幅方向に発生する変動の成分を表す．また，HNR は音声波形の基本周波数のパワーと，ノイズ成分のパワーの比を表す．今回この 3 つのパラメータを，音声特徴量抽出ソフトである OpenSMILE[Eyben 2010]を用いて音声ファイルから 3 つの特徴量を抽出した．プリセットとして，avec2011.conf を使用し，音声ファイルごとに得られたパラメータのうち，中央値を表す値を採用した． この値を用いて健常者と構音障害者のラベル及び各指標の値を対応付け，ROC 曲線を描きその AUC の値を見ることで分類性能を評価した．

3.4 MFCC を用いた機械学習による構音障害分類

今回 MFCC を用いた分類を行うため，ニューラルネットワーク(DNN)および SVM の二つのモデルを構築した．二つのモデルの構築のために，ノイズが付加されていないセットで学習を行った．MFCC の導出にあたっては，Python のライブラリである librosa を利用した．音声から MFCC を抽出するにあたって，音声ファイルの先頭からウィンドウ幅 46 ms (512 サンプル)フレームシフト幅 12 ms (128 サンプル)と指定して切り出した．この切り出したウィンドウに対して，MFCC の 13 次元の係数と，その一次微分項の計 26 次元のベクトルを導出した．この 26 次元のベクトルを，DNN 及び SVM の入力として扱った．DNN の構築には chainer を利用した．DNN は，中間層として 300 ノードの全結合を 2 層用意し，出力層は 2 ノードの全結合を用意した．SVM は，python のライブラリである sckit-learn を用いた．パラメータとして，コストパラメータは 1.0，基底関数は RBF(動径基底関数)を指定した．この 2 つの学習済みモデルに対して，テスト用のデータセットから音声を入力すると，各ウィンドウに対して健常者又は構音障害者の分類が得られる．今回は全体のウィンドウ数に対して構音障害と分類されたウィンドウの数の割合を，その音声における構音障害のスコアとして利用し，このスコアを用いて健常者と構音障害者のラベル及び各指標の値を対応付け，ROC 曲線を描きその AUC の値を見ることで分類性能を評価した．

4. 実験結果と考察

図 3 図 4 にノイズレベルを変化させたときの AUC の値の推移，表 1 表 2 に AUC の数値を示す．音声指標を利用した分類において，jitter, shimmer, HNR などの音声指標を用いた場合の分類性能は，ノイズレベルを上昇させるといずれも低下した．ホワイトノイズを付加させた場合では，レベルが上昇するにつれて均一な下がり方をしており，車の走行音を付加した場合では，ノイズレベルが比較的高い時に性能が急激に悪化した．これは，この 3 つの音声指標はいずれも音の乱れが激しいほど高くなるため，ノイズの影響が無視できる際は健常者と構音障害者の声の乱れを捉えることが出来る指標となるが，ノイズレベルが大きい際は，ノイズから生じる音の乱れが指標に及ぼす影響が強くなるため，健常者と構音障害者の差が指標に含まれなくなる．そのためノイズレベルが高くなるにつれて，構音障害の分類性能が悪くなったと考えられる．

一方 MFCC を用いて DNN や SVM を用いた場合は，音声指標を用いた場合と比較して，ノイズレベルが上昇した場合でも分類性能が低下しなかった．DNN と SVM の間で性能及び性能の下がり具合に顕著な差は見られなかった．これらの結果は，MFCC から構音障害を持つ人に現れる音声特有の特徴を抽出し，テストデータにノイズが含まれていても読み取ることが出来る特徴を学習した結果が現れたものだと考えることが出来る．MFCC は音声特徴量として汎用的に用いることができるため，構音障害以外の状態分類に応用出来る可能性がある．

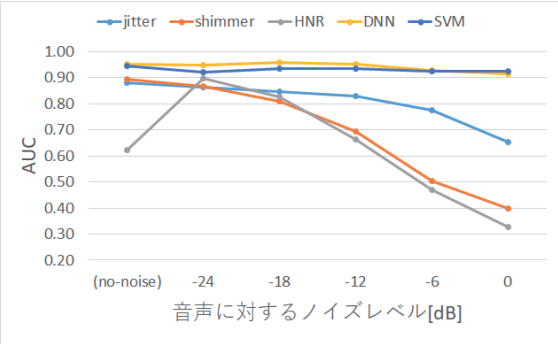

図 3. ホワイトノイズを付加させた際の AUC 推移

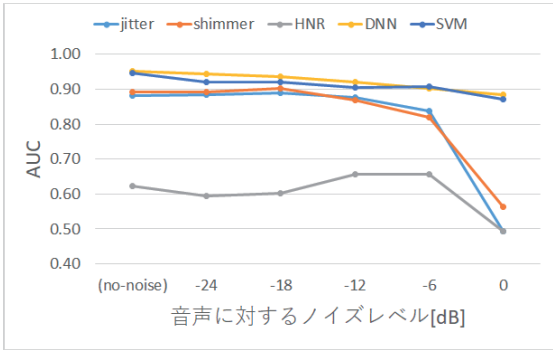

図 4. 車の走行音を付加させた際の AUC 推移

表 1. ホワイトノイズを付加させた際の AUC

音声に対するノイズレベル[dB]	jitter	shimmer	HNR	DNN	SVM
(no-noise)	0.88	0.89	0.62	0.95	0.95
-24	0.86	0.87	0.90	0.95	0.92
-18	0.85	0.81	0.83	0.96	0.94
-12	0.83	0.69	0.66	0.95	0.93
-6	0.78	0.50	0.47	0.93	0.93
0	0.65	0.40	0.33	0.92	0.93

表 2. 車の走行音を付加させた際の AUC

音声に対するノイズレベル[dB]	jitter	shimmer	HNR	DNN	SVM
(no-noise)	0.88	0.89	0.62	0.95	0.95
-24	0.88	0.89	0.59	0.94	0.92
-18	0.89	0.90	0.60	0.94	0.92
-12	0.88	0.87	0.66	0.92	0.91
-6	0.84	0.82	0.66	0.90	0.91
0	0.49	0.56	0.49	0.88	0.87

5. まとめ

ノイズが混在するような環境下で音声を取得する際に，MFCC を用いた分類を行うことで，既存の音声指標を用いた場合よりも分類性能を向上させる可能性が示唆された．今後の展望としては，今回は既存のデータセットに対してノイズを付加して評価したが，実際の録音時にノイズが含まれた場合の耐ノイズ性能の推移を評価することや，構音障害分類以外の問題での耐ノイズ性能の検証を行い，音声から人の内部状態をセンシングする際により強固な手法を確立していくことが挙げられる．

参考文献

[光吉 2011] 光吉俊二, 徳野慎一, 田中靖人: “音声感情技術 ST を使ったストレスへの応用”, 日本疲労学会誌, 第 6 巻, 第 2号, pp.641-644, 2011.

[Tokuno 2015] Tokuno, Shinichi: Stress evaluation by voice: from prevention to treatment in mental health care, Econophysics, Sociophysics & other Multidisciplinary Sciences Journal, vol.5, pp30-35, 2015.

[Shinohara 2017] Shinohara, Shuji, et al.: Case studies of utilization of the mind monitoring system (MIMOSYS) using voice and its future prospects, Econophysics, Sociophysics, and Multidisciplinary Sciences Journal, vol.7, pp7-12, 2017

[髙野 2017] 髙野 毅 ほか, “走行中の自動車騒音が感情やストレス状態の分類に有用な音声構造解析に与える影響”, HCG シンポジウム 2017, 2017.

[Teixeira 2014] Teixeira, João Paulo, and Paula Odete Fernandes: "Jitter, Shimmer and HNR classification within gender, tones and vowels in healthy voices." Procedia Technology, vol.16 pp.1228-1237, 2014.

[Shinohara 2016] Shinohara, Shuji, et al. "Voice disability index using pitch rate." Biomedical Engineering and Sciences (IECBES), 2016 IEEE EMBS Conference on. IEEE, 2016.

[Londoño 2011] Arias-Londoño, Julián D., et al. "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients." IEEE Transactions on Biomedical Engineering, vol.58, pp.370-379, 2011

[Fang 2018] Fang, Shih-Hau, et al. "Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach." Journal of Voice, In press, 2018.

[Alhussein 2018] Alhussein, Musaed, and Ghulam Muhammad. "Voice pathology detection using deep learning on mobile healthcare framework." IEEE Access, vol.6, pp. 41034-41041, 2018.

[Llorente 2006] Godino-Llorente, Juan Ignacio, Pedro Gomez-Vilda, and Manuel Blanco-Velasco. "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters." IEEE transactions on biomedical engineering vol.53, pp.1943-1953, 2006

[Eyben 2010] Eyben, Florian, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor." Proceedings of the 18th ACM international conference on Multimedia, 2010.