This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.

第 27 回日本バーチャルリアリティ学会大会論文集（2022 年 9 月）

WEB 会議システム使用下における不快音の除去
― 不快音の特徴と検出 ―

Remove discomfort sound during videoconference
- Analysis and detection of Discomfort sound -

松山潔 [1),高橋秀智 [2)
Kiyoshi MATSUYAMA and Hidetomo TAKAHASHI

1) 東京工科大学 工学研究科 （〒192-0982 東京都八王子市片倉町 1404-1）
2) 東京工科大学 工学研究科 （〒192-0982 東京都八王子市片倉町 1404-1,takahashihdtm@stf.teu.ac.jp）

概要：本研究では, WEB 会議システム使用下における不快音を, 機械学習を用いて判別, 除去をすること目的とする． 本報では, まず WEB 会議システム使用時に生じる不快音を抽出・分析する. ここでは不快音を飲食音とする. 次いで, 通常の音声データと不快音データに分けラベル付けを行う. これらのデータの音圧波形や, 音圧波形を MFCC など画像化した物を使用し, 機械学習を用い検出器を作成し, 不快音の検出特性の評価を行う.
キーワード：聴覚, 不快音,

1. 背景

新型コロナウィルスの感染拡大により対面での接触を減らすため,急速に利用が拡大したのが「Web 会議システム」である.従来あまり使用してこなかった用途での利用も拡大し,対面では気にならなかった小さな音が大きく聞こえ,不快に感じる場面が増加した.「Web 会議システム」使用時に生じる音に対して分析を行い人が不快に感じる音を簡単に除去することが出来ないかと考えた.

2. 目的

「Web 会議システム」使用時に生じる音に対して分析を行い,人が不快に感じる音を機械学習を用いて簡単に除去するフィルタの作成を目標とする.本論文では人が食べ物を食べる飲食音を不快音として扱い,会話音との識別を目指した.

3. 研究手法

3.1 使用するデータ

学習に使用する音声データとして,YouTube を用いて男女の会話音,飲食音を 1 秒ごとに切り出した音を使用する.

今回使用したデータ数は会話 45,飲食音 80,それぞれにノイズを付加し,会話 90,飲食音 160 に増やして使用した.

3.2 手法

入力した音声データを二次元画像に変換し， CNN の学習および画像を用いた分類を行う.

南ら(2020)の研究では,畳み込みニューラルネットワーク（CNN：convolutional neural network）を用いた呼吸音の自動分類手法の提案を行っており． 呼吸音データに対して短時間フーリエ変換と連続ウェーブレット変換を適用し， スペクトログラム画像およびスカログラム画像を生成し,その画像を用いて CNN による呼吸音の識別を行っている.[1]

本論文では短時間フーリエ変換を用い， スペクトログラムに変換し,周波数領域のスペクトルを時間信号とみなし,さらにフーリエ変換を施したケプストラムと呼ばれる画像に変換する方法,更に人間の聴覚,音の聞こえ方に基づいたメル尺度と呼ばれる尺度に変換する方法を用いる.

(1)メル尺度

メル尺度は,人間の聴覚,音の聞こえ方に基づいた尺度である.人間の聴覚には,周波数の低い音に対して敏感で,周波数の高い音に対して鈍感であることから,.1000 [Hz]の音を 1000 [mel]の音高と定め,以下のような関係式(1)がある. ここでは mel はメル尺度を表し,f は周波数を表す.

$$mel = 2595.0 \times \log_{10}(1.0 + \frac{f}{700.0}) \qquad (1)$$

ある信号をサンプリングする際,そのサンプリング周波数の 1/2 に相当する周波数のことをナイキスト周波数と呼

ⓒ 2022 日本バーチャルリアリティ学会

ぶ,ナイキスト周波数に対応するメル尺度を算出し,このメル尺度の値を上限として,0 [mel]と上限値との間を等分割する.等分割して得られたメル尺度の値を対応する周波数へ戻すことにより,人間の聴覚に即した周波数スケールを作ることが出来る.

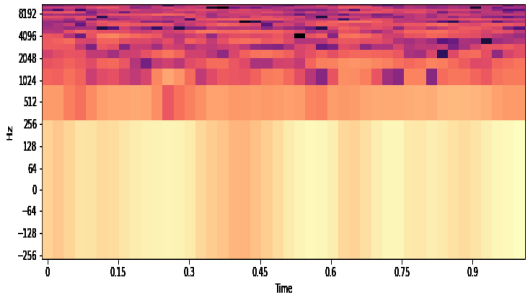
(2) ケプストラム[5-6]

音声信号のフーリエ変換の絶対値に対数をかけ,さらに逆フーリエ変換したもの.信号 x(n)のフーリエ変換を $X(e^{j\omega})$ とすると,ケプストラム c(m)の式(2)を示す.

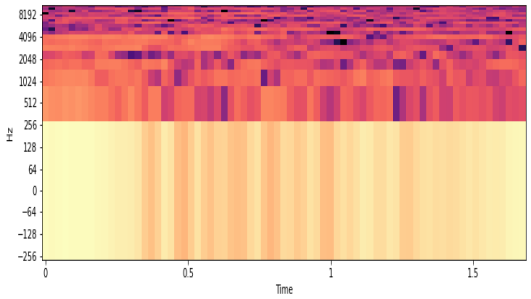
$$c(m) = e^{j\omega} = \frac{1}{2\pi}\int_{\pi}^{\pi} \log|X(e^{j\omega})|\, e^{j\omega n} d\omega \quad (2)$$

ケプストラムの横軸 m は時間になるが,通常の時間波形の時間と異なることから,ケフレンシ(quefrency)と呼ぶ.高ケフレンシ領域は微細振動,低ケフレンシ領域はスペクトル包絡を表す.ケプストラムの一部のケフレンシ成分を取り出す処理をリフタリング（liftering）と呼び,リフタリングを施すことによってこれらを分離することができる.

本論文では,CNN に使用する学習データとして,ケプストラムをリフタリングしてスペクトル包絡を切り出した画像データと,スペクトルをメル尺度に変換したメルスペクトルに対してケプストラムを求めた MFCC（メル周波数ケプストラム係数）を使用する.生成した画像の例を**図 1,2** に示す.作成した画像は，ケプストラムのスペクトル包絡,MFCC ともにフレーム長 1 [s]である.

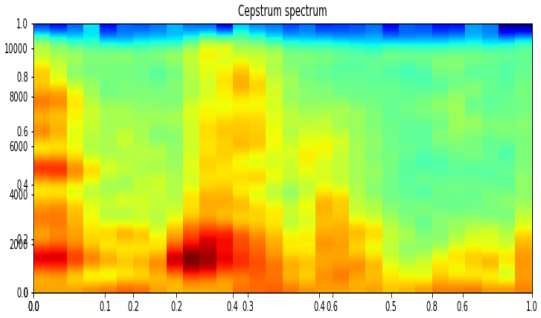


(a) **会話音**

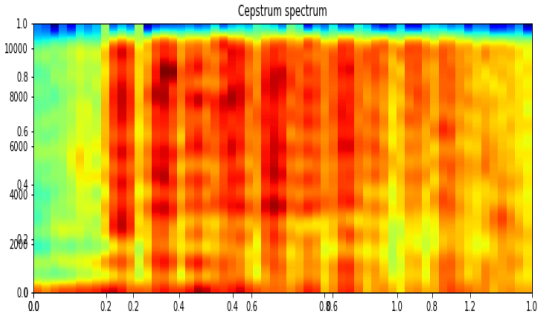


(b) 飲食音

図 1:MFCC 化画像



(a) 会話音



(b) 飲食音

図 2:ケプストラムのスペクトル包絡化画像

4. 学習システム

4.1 概要

本研究では CNN を用いて機械学習を実装していく．学習には機械学習用のオープンソースライブラリである Keras を使用する.

250 個の音声データを画像化,CNN の入力サイズ(256×256 [pixel])にリサイズを行い,学習データとして使用し,学習データの 10%を精度検証データとして使用する.ランダムで活性を抑えるドロップアウトを設け,2 層目と出力層に 30%のドロップアウト率を設定した.学習には活性化関数として ReLU 関数を使用した.

本論文で用いたネットワーク構造を**表 1** に示す.

表 1:ネットワーク構成

Layer	Filter_size	Stride	Output_size	Remarks
Input	—	—	256×256× 3	—
Conv.1	3×3	1×1	256×256×32	ReLU
Pool.1	2×2	2×2	126×126×32	Dropout
Conv.2	3×3	1×1	126×126×64	ReLU
Pool.2	2×2	2×2	62×62×64	—
FC1	—	—	1×1×512	ReLU,Dropout
FC2	—	—	1×1×2	Softmax

ⓒ 2022 日本バーチャルリアリティ学会

4.2 学習結果

MFCC 化画像とケプストラムのスペクトル包絡化画像について学習を行った結果が**図 3,4** である.

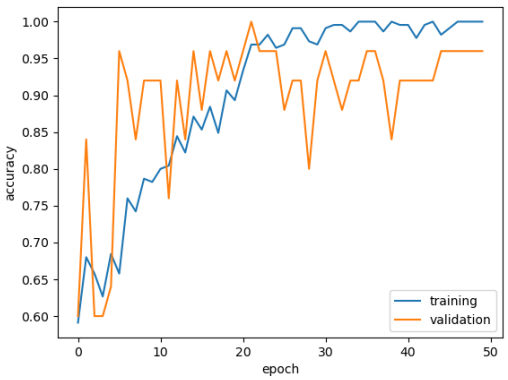


図 3:MFCC 化画像による学習曲線

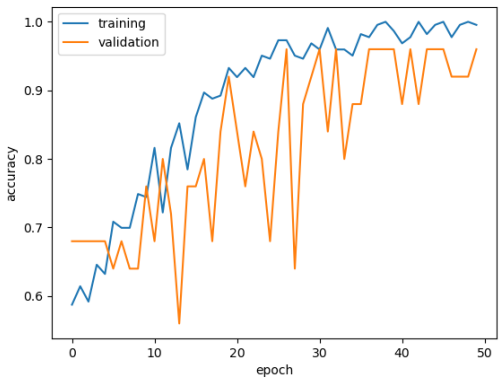


図 4:ケプストラムのスペクトル包絡による学習曲線

5. 今後の課題

機械学習について,使用する画像データは適切か,機械学機械学習について,使用する画像データは適切か,機械学習のモデルは適切であるか,男女や食べ物の種類やその他の不快音との識別は可能か,学習用データを増やすことで正答率の向上が図る.さらに,リアルタイム処理化についても検討する.

参考文献

[1] 南弘毅, 陸慧敏, 金亨燮, 平野靖, 間普真吾, 木戸尚治：時間-周波数解析と畳み込みニューラルネットワークを用いた呼吸音の自動分類．Medial Imaging Technology 38(1),pp.40,47,(2020)

[2] 和田成夫：よくわかる信号処理ーフーリエ変換 からウェーブレット変換まで．森北出版，(2009)

[3] 谷萩隆嗣：ディジタル信号処理の理論 1 基礎・ システム・制御．コロナ社，(1985)

[4] 小野順貴：短時間フーリエ変換の基礎と応用．日本音響学会誌 72(12),pp.764,769, (2016)

[5] 板橋秀一：音声工学.森北出版，(2005)

[6] 古井貞熙:新音響・音声工学.近代科学社，(2006)

[7] 美濃羽靖,和田誠,田中：紡深層学習を用いた樹幹からの打撃音に基づく樹高および材積の推定.日本森林学会誌 103(5),pp.351,360,(2021)

ⓒ 2022 日本バーチャルリアリティ学会