G-7

CNN を用いた生活音識別に関する検討

A Study on Daily Life Sound Identification Using Convolutional Neural Network

○井上翔貴[1]，　関弘翔[2]，　細野裕行[2]
*Shoki Inoue[1], Hiroto Seki[2], Hiroyuki Hosono[2]

Abstract: The purpose of this study is to support the lives of the hearing impaired by visually transmitting the sounds of life. In this paper, we studied the method of identifying daily life sounds by CNN focusing on the spectrogram of sound, and further studied the visualization of the factors of identification.

１．まえがき

厚生労働省が行った平成 28 年度の調査結果[1]において，聴覚障害者の人数は 65 歳未満は 6 万人，65 歳以上及び年齢不詳は 23 万 7 千人であり，約 31 万人が聴覚障害を患っていると報告されている．少なくない聴覚障害者の生活支援を目的とし，本研究では生活音識別システムの実現を目指している．

先行研究[2]では NN(Neural Network)を用いて 18 種類の生活音識別の検討を行い，全体の識別率が約 99%，さらに 3%程度のノイズへの耐性を得る事にも成功した．開き戸を開ける音と閉める音などよく似た音の誤識別があったが，NN がどのような特徴に基づいて識別を行うか不明確であり，誤識別の原因の解明は困難であった．

本研究では，新たに CNN(Convolutional Neural Network)を用いたスペクトログラムを入力とする手法による生活音識別を検討するとともに，XAI(eXplainable AI)の一つである SHAP(Shapley Additive exPlanations)[3]を用い，誤識別を含む識別要因の可視化について基礎的検討を行う．

２．データセット

データセットには先行研究[2]で使用した 18 種類の生活音を用いる．CNN への入力を前提に，スペクトログラム化したものをデータセットとしており，学習 2080 サンプル，検証 520 サンプルの合計 2600 サンプルで検討を行う．

３．CNN モデル

CNN モデルには，画像認識において成果を挙げているモデルのなかでも，SHAP の適用を前提として比較的シンプルな構造を持つ VGG16[4]を参考に構築した．Fig.1 に構築した CNN モデルを示す．入力画像については 224×224 に正規化した後，入力している．一般的な画像と異なりスペクトログラムはクラスごとの違いが少ないため，Pooling により VGG16 と同様に入力の 32 分の 1 まで特徴マップを小さくして全結合層に入力した場合に精度が得られなかった．本研究では GPU メモリと精度の兼ね合いより，16 分の 1 で全結合層に入力する構造とした．

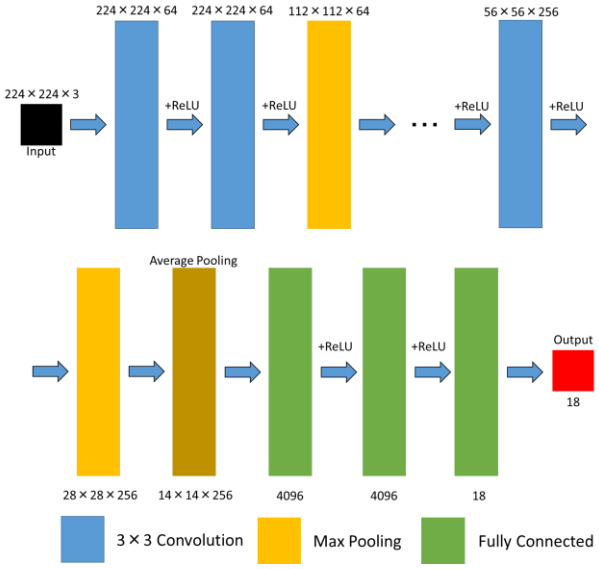

Figure 1. CNN model

４．結果

Table 1 の条件で学習用データを用いて CNN を学習し，検証用データを用いて識別精度を検証した．また，SHAP を用いて識別要因を検証した．

SHAP とは，出力結果に対する入力の寄与を可視化する XAI の手法であり，様々な従来の XAI の手法に協力ゲーム理論の考え方を導入したものである．CNN を用いた識別に対しては，入力画像の各画素が識別結果に対してポジティブに貢献したかネガティブに貢献したかを知ることができる．

１：日大理工・院（前）・情報，　２：日大理工・教員・情報

Table 1. Learning conditions

GPU	NVIDIA GTX 1080 8GB
Epoch	20
Batch size	4
Optimization function	Adam
Learning Rate	0.0001
Loss function	Cross Entropy

４．１．　識別精度の検証

Table 2 に識別精度を示す．全体で約 94%の精度となったが，先行研究と同様に開き戸を開ける音や冷蔵庫の開け閉め，引き戸を閉める音において誤った識別をしていることが分かる．

Table 2. Identification rate

Daily Life Sound	Accuracy [%]	Daily Life Sound	Accuracy [%]
1.Intercom	100	10.Water flowing(1)	100
2.LINE phone	100	11.Toaster	100
3.Fire alarm	100	12.Sliding door open	100
4.Microwave(1)	100	13.Sliding door close	95
5.Opening door open	50	14.Water flowing(2)	100
6.Opening door close	100	15.Microwave(2)	100
7.Refrigerator open	65	16.Black phone	100
8.Refrigerator close	85	17.Phone(1)	100
9.Gas stove	100	18.Phone(2)	100

４．２．　識別要因の検証

開き戸を開ける音について Fig.2 に正しい識別の要因を，Fig.3 に誤った識別の要因をそれぞれ可視化した例を示す．例えば Fig.2(b)は Fig.2(a)を入力した際に最も高い確率で予測した開き戸を開ける音という予測に対してどの画素がどのように貢献したかを示しており，SHAP value が高い（赤い）ほどポジティブ，低い（青い）ほどネガティブであることを意味している．

Fig. 2(b)と Fig.3(c)より，正解したサンプルと誤ったサンプルで同様の特徴を捉えて開き戸を開ける音を予測していることが分かる．しかし，冷蔵庫を開ける音を予測する際も同じような特徴を捉えているため，誤った識別結果になったことが推測できる．Fig.4 に冷蔵庫を開ける音及び開き戸を開ける音のスペクトログラムの例を示す．精度向上のためには，このように視覚的に似たクラスに対しても弁別性の高い特徴を捉えられるような CNN モデルを構築する必要がある．

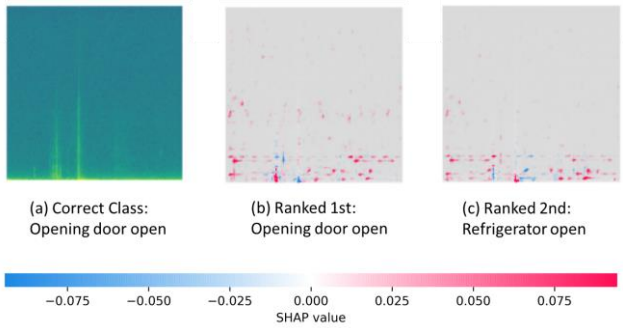

Figure 2. Correct result for Opening door open

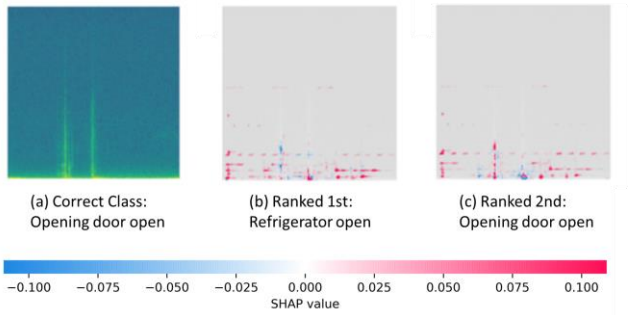

Figure 3. Erroneous result for Opening door open

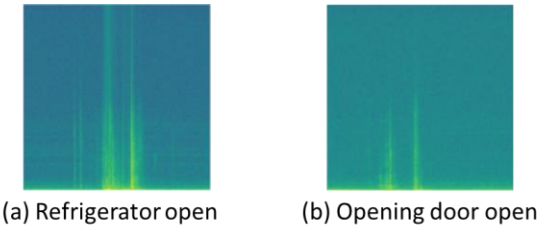

Figure 4. Example of Spectrogram

５．まとめ

本報告では，CNN を用いた手法及び識別結果要因の可視化について基礎的検討を行った．

今後は得られた識別要因を基に，異なるクラス間で同じような特徴を利用しないようなモデルの構築を模索すると共に，ノイズを付加した生活音に対する識別を CNN で検討していく．

６．参考文献

[1] 厚生労働省社会・援護局障害保健福祉部：「平成２８年生活のしづらさなどに関する調査（全国在宅障害児・者等実態調査）結果」，(2018-04)

[2] 佐々木駿，他：「ニューラルネットワークを用いた生活音識別」，日本大学理工学部学術講演会，G-18，(2019-12)

[3] S. M. Lundberg, et al. :“A Unified Approach to Interpreting Model Predictions,” NIPS2017, pp.4675-4774 (2017)

[4] K. Simonyan, et al. :“Very Deep Convolutional Networks for Large-Scale Image Recognition, ” arXiv preprint arXiv:1409.1556 (2014)