H-020

転移学習を用いた音源識別における画像変換手法とその性能比較
Image Transformation Methods for Sound Classification using Transfer Learning and Comparison of their Performances

藍原 直大† 川喜田 佑介† 宮崎 剛† 田中 博†
Naohiro Aihara Yuusuke Kawakita Tsuyoshi Miyazaki Hiroshi Tanaka

1. はじめに

話者識別や室内環境の音源識別において，音を画像に変換し，深層学習を用いて識別モデルを作成する方法がある[1],[2]．画像への変換によって音の特徴抽出を行うが，適用される変換手法は各種あるものの、それぞれの特徴が異なり，識別に最適な手法は識別対象の音源によって異なると思われる．本検討では各種の音源に対して，特徴抽出に多く用いられているスペクトログラム，メルスペクトログラム，スカログラムへの変換を行い，各画像を学習データとした識別モデルを作成する．本モデルによる精度評価を行うことで，最適な変換方法やそのパラメータについて検討した結果を述べる．

2. 識別対象音源

識別対象の音源として，表 1 に示す 3 種のカテゴリ（室内音，アラーム音，話者音）を用いた．機械的な音や人の声といった特徴の異なる音源を識別することで，画像変換手法の比較を行う．室内音は室内環境音推定の検討[1]で収録した，代表的な家電音 4 種とその混合音を用いた．2 音混合音が 6 種（$={}_4C_2$），3 音混合音が 4（$={}_4C_3$）で計 14 クラスである．アラーム音は聴覚障がい者支援を目的とした識別実験[3]で対象としたアラーム機器を用いた．また，識別対象とする日常を想定したアラーム音の出力されていない空間の環境音，会話音も同様に収録した．話者音は男性 10 名の新聞記事読み上げ音声を用いた．

表 1　識別カテゴリと音源取得

カテゴリ	対象音源	
室内音	空調機器	ドライヤー
	髭剃り	掃除機
	2 音混合音（6 クラス）	3 音混合音（4 クラス）
アラーム音	窓アラーム	煙感知器
	熱感知器	ガス漏れ報知器
	インターホン	キッチンタイマー
	目覚まし時計	会話音
	環境音	
話者音	10 人分の新聞記事読み上げ音声	

3. 画像変換手法と変換画像

3.1 用いる画像変換手法

音を画像に変換することで，時系列の音の強度情報を時間－周波数の 2 次元の画像情報に変換できる．特徴抽出に多く用いられている画像変換手法として，図 1 に示す 3 種の手法を比較する．

スペクトログラムは短時間フーリエ変換（STFT）を行い，パワースペクトル密度を描画する．窓幅が固定されているため時間分解能が一定である．

メルスペクトログラムはフーリエ変換後にメルフィルタバンクを適用した画像である．周波数をメル尺度に変換することで，人間の聴覚特性に近い特徴を抽出できる．図 1 ではスペクトログラムの特徴に対して，メルスペクトログラムの低周波成分である横線の縞模様の特徴が明確に表れていることがわかる．

スカログラムは連続ウェーブレット変換（CWT）の結果を描画する．短時間フーリエ変換とは異なり，周波数に合わせて適用する窓幅を変える．高周波成分では時間分解能が高く，低周波成分では周波数分解能が高い．画像では 100Hz 以下の周波数成分まで特徴が現れているのがわかる．また，低周波ほど大きくなる窓幅による計算結果の歪みを考慮し，円錐状影響圏の輝度値を 0（黒色）に設定する．

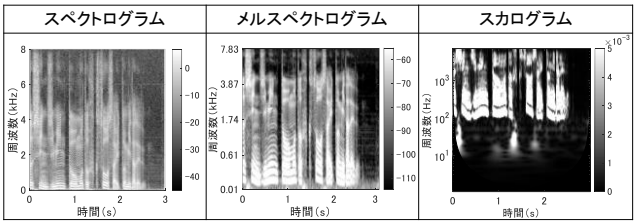

図 1　各画像変換結果の例

3.2 スケールによる画像変化

周波数のスケールやレンジ設定など，画像変換時のパラメータ変更によって画像が変化する．変換手法の違いを明確にするために，まず各パラメータを確定させ，その後スペクトログラムとスカログラムの周波数スケールを一致させた．変換結果の例を図 2 に示す．メルスペクトログラムに関しては，デフォルトのメルスケール画像として比較を行った．

スペクトログラムは対数スケールにすることで低周波領域の特徴が出るが，スカログラムと比較してデータのプロットされていない周波数帯がある．また，スカログラムでも同様に高周波領域の上限が異なる．本検討では，各画像の周波数レンジを表 2 に示すように，プロットデータの最大値，最小値を変換手法ごと固定し，比較を行った．

†神奈川工科大学情報学部情報工学科　Dept. of Information and Computer Sciences, Kanagawa Institute of Technology

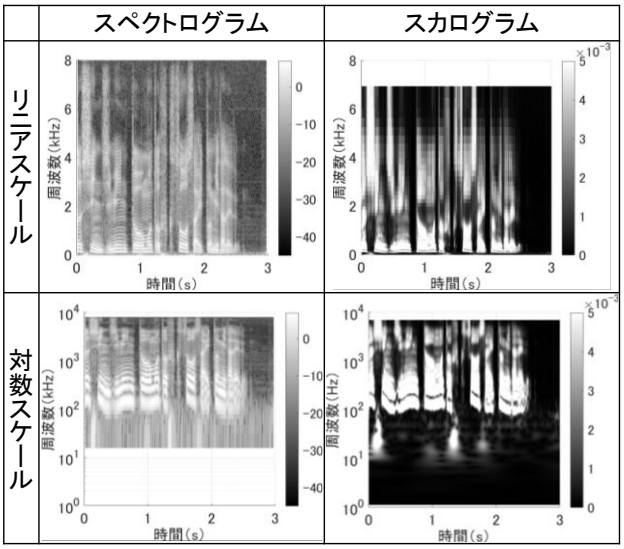

図 2　周波数スケールの違いによる画像変化の例

表 2　変換手法ごとの周波数レンジ

	リニア	対数	メル
スペクトログラム	[0.0 8000.0]	[15.6 8000.0]	
メルスペクトログラム			[13.8 7831.7]
スカログラム	[1.1 6949.0]	[1.1 6946.0]	

周波数レンジ(Hz)[最大値 最小値]

4. 識別手法

変換方法とスケール設定による画像の変化と識別性能を評価するため，図 3 に示す 5 種の画像それぞれの学習データを作成し，後述する識別モデルの入力とする． 3 カテゴリの識別対象音源に対して同様の手法を行い，その識別精度を比較する． また，同一のデータを用いた学習を 3 回行い，学習精度の変動を確認する．

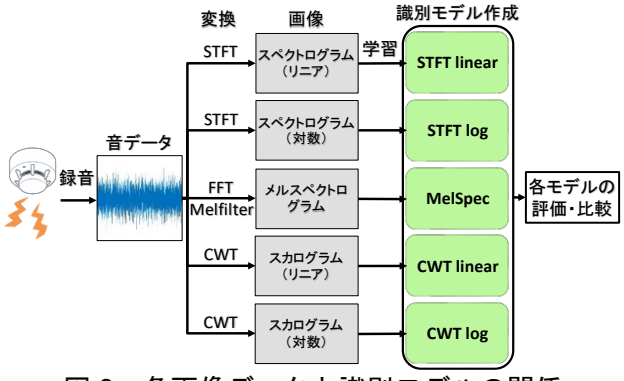

図 3　各画像データと識別モデルの関係

4.1 入力データの作成

画像変換手法ごとの違いを調べるため，表 3 に示す学習・評価データは同条件で収録した音源データを使用した．室内音とアラーム音は，マイクと対象機器の距離を 1m と 2m に設定し録音を行った．話者音で使用した音源データは「日本音響学会新聞記事読み上げ音声コーパス」[4]に収録されたものを使用した．

画像の各パラメータは変換手法ごとに画像全体の特徴が明瞭に表示され，良好な識別結果が出るよう調整を行い，最終的にスペクトログラム，メルスペクトログラムは窓関数 hann，フレームサイズ 512，オーバーラップ 256，ポイント数 1024 に設定した．メルスペクトログラムのメルバンドパスフィルター数は 128 である．各音源のサンプリング周波数は 16kHz で統一している．

1 画像に十分な音の特徴が含まれるように各画像 3 秒間の音データを変換した．後述する学習モデルの入力データとするため，表 2 の周波数レンジで固定した各画像を 703×703 の正方形画像として出力する．各カテゴリの 1 クラスの学習データ数は 30，評価データ数は 20 で行った．

表 3　学習と評価に使用したデータ

カテゴリ	1 クラスのデータ数	クラス数	取得方法
室内音	50(30, 20)	14	1m の距離で録音
アラーム音	50(30, 20)	9	の距離で録音
話者音	50(30, 20)	10	新聞記事読み上げデータ

(学習用, 評価用)

4.2 識別モデルの作成

識別のための学習モデル作成は，転移学習を用いて行った．図 4 に示す事前学習済みのネットワークに対し，全結合層と分類層を置き換えたうえで再学習を行う．必要となるデータ数や学習時間から，GoogLeNet を適用した．従来 1000 クラスの物体の分類を行うネットワークの識別対象を，本実験で使用する室内音 14，アラーム音 9，話者音 10 クラスへと変更した．学習パラメータは，バッチサイズ 24，エポック数 300，学習率 1e-5，最適化アルゴリズム SGDM，損失関数は交差エントロピーに設定した．各学習パラメータは画像比較を目的としているため，全て固定してモデル作成を行った．

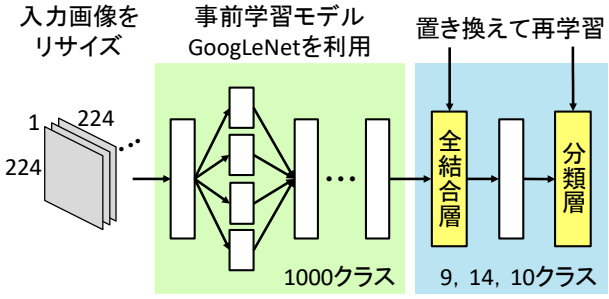

図 4　識別モデル作成手法

4.3 識別モデルの学習状況

識別モデル作成時の学習曲線の一例として，図 5 の話者音の学習状況を示す．学習パラメータを統一したことで，学習曲線の収束の違いを比較することができる．話者識別においては，スカログラムの対数スケール画像を用いた識別モデル CWT log が，ほかのスペクトログラムやメルスペクトログラムのモデルに比べて，収束状況が早いことが確

認できる．一方，同じ対数スケールに設定したスペクトログラムの識別モデル STFT log ではスカログラムほどの収束はしていない．このことから話者識別では，スカログラムの対数スケール画像が識別しやすい特徴であると考えられる．

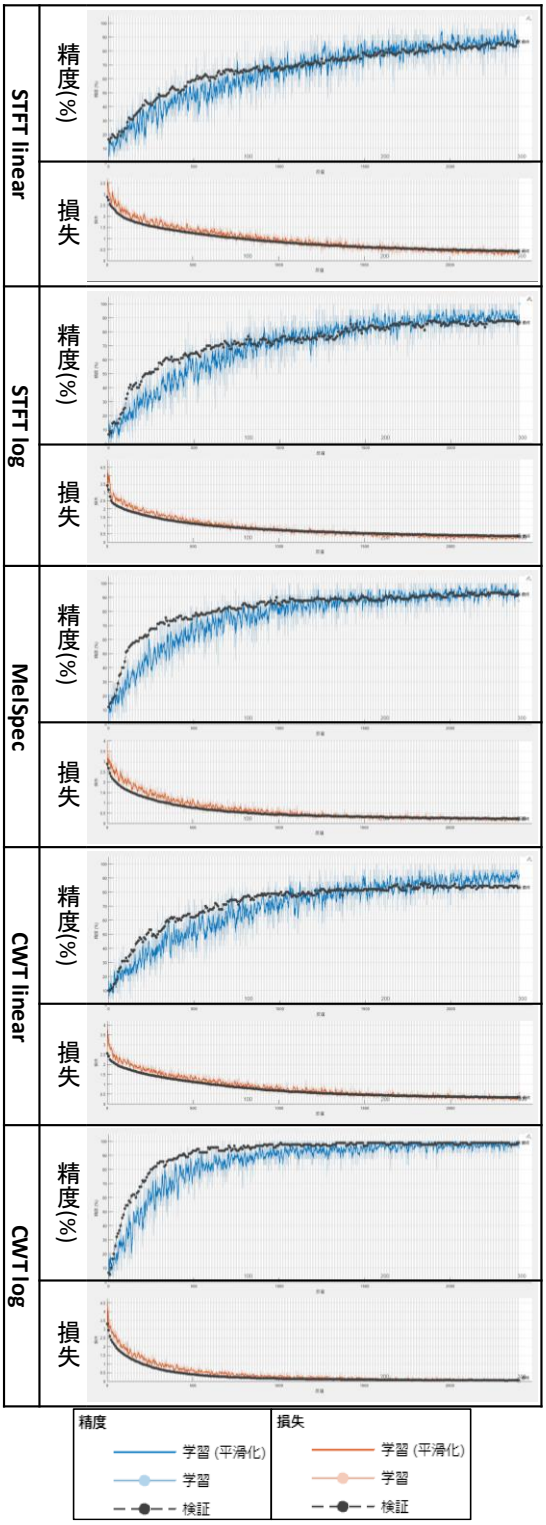

図 5　識別モデル作成時の学習曲線の一例（話者識別）

5. 各識別モデル評価

評価用データを用いて，作成した各種識別モデルの混同行列を作成した．話者音の混同行列の一例を図 6 に示す．

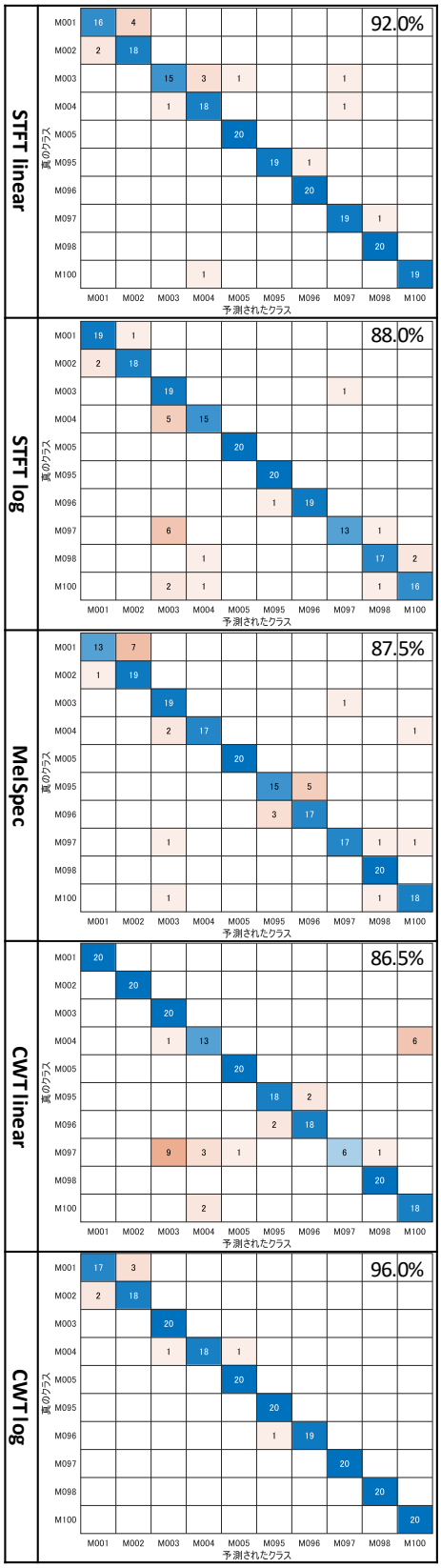

図 6　混同行列の一例（話者識別）

各クラスの略字は引用元[4]と同一の音源の人物を識別とする M001，M002…とし，雑音の含まれていない音源を選出した．各混同行列の縦軸は真のクラス，横軸は予測されたクラスを示す．識別の正解率は混同行列の右上に示す．低周波領域の特徴が出やすいスカログラムの対数スケール画像を学習したモデル CWT log が 96.0%と最も高い識別精度となった．一方，スカログラムのリニアスケール画像を学習したモデル CWT linear では 86.5%と対数スケールに比べて精度が落ちている．特に声の低周波成分が多い M097 の話者音源の誤認識数が多くなっている．

また，同一のデータを用いた学習を 3 回行い，学習精度の変動を調査した結果を図 7 にまとめた．縦軸は精度(%)，横軸は比較に用いた変換画像 5 種を示している．傾向として室内音，話者音はスカログラムの対数スケール CWT log が高精度であり，アラーム音はスペクトログラムのリニアスケール STFT linear が高精度であることがわかる．3 つの識別音源カテゴリのうち，高周波領域で音源が特徴的なものが多いアラーム音の識別精度が全体的に高く，スカログラムの対数スケールの 90%前後の識別を除いて，100%に近い精度が出ている．

これらの結果から，高精度な識別のためには，識別対象の周波数領域をもとに変換手法を選択する必要があると考えられる．一方，学習回数に応じた分布では精度のばらつきが大きい変換画像もあることから，学習データ数や学習のエポック数などのパラメータを調整し，各モデルの結果の安定性を確認していく必要があると考えられる．

表 4　各モデルの識別精度[%]

	STFT		LMS	CWT	
	linear	log	Mel	linear	log
室内音	61.6	67.9	69.9	72.5	74.0
	60.3	68.2	69.1	75.0	78.4
	57.9	71.4	67.1	75.7	77.0
話者音	92.0	88.0	87.5	83.5	96.0
	85.5	84.5	86.0	86.5	96.5
	93.5	90.5	89.0	90.0	94.5
アラーム音	100.0	98.9	97.8	99.4	86.7
	100.0	99.4	99.4	99.4	90.6
	100.0	98.9	100.0	99.4	91.1

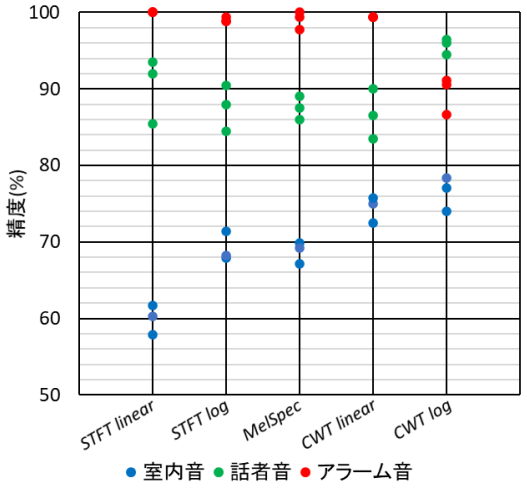

図 7　各モデルの識別精度のグラフ

6. まとめと今後の課題

スペクトログラム，メルスペクトログラム，スカログラムによる音源識別性能の比較を目的として，各スケールによる特徴変化を示し，音源識別における各変換方法の精度比較を行った．

室内音，話者音の識別では，低周波成分の分解能が高いスカログラムの対数スケール，アラーム音ではスペクトログラムのリニアスケールを用いた識別が良好であると確認することができた．

音源に含まれる周波数成分によって，各変換方法の識別性能に差が出ることから，識別対象音源の周波数帯に応じて，変換手法を選択していく必要があることが分かった．

今後の課題として，高周波と低周波が共に含まれた場合の音源を識別する場合を検討していく．

参考文献

[1] 佐野 将太, 川喜田 佑介, 宮崎 剛, 田中 博, “スペクトログラム画像を用いた転移学習の適用による室内音識別”, 画像電子学会誌, Vol.52, No.2, pp.357-363 (2023)
[2] Jianrui Lu, Ruofei Ma, Gongliang Liu, Zhiliang Qin, “Deep Convolutional Neural Network with Transfer Learning for Environmental Sound Classification”, 2021 International Conference on Computer, Control and Robotics, 21-8-10, pp.242-245 (2021).
[3] 藍原 直大, 佐野 将太, 田中 博, “スペクトログラム画像と GoogLeNet を用いたアラーム音の識別実験”, 情報・システムソサイエティ特別企画 ジュニア&学生ポスターセッション予稿集, ISS-P-20, p.72 (2023).
[4] JNAS 音声資源コンソーシアム新聞記事読み上げ音声コーパス, http://research.nii.ac.jp/src/JNAS.html, 参照 2022/4/27.