

《様式B》

研究テーマ	「CMOS カメラで撮影された動画像からの音響信号復元とその打音・振動計測への応用」		
研究責任者	所属機関名	常葉大学・社会環境学部（応募時: 静岡大学大学院工学領域）	
	官職又は役職	教授	
共同研究者	氏 名	立蔵 洋介	メールアドレス tatekura@sz.tokoha-u.ac.jp
	所属機関名	株式会社エヌエスティー	
共同研究者	官職又は役職	システム事業部システム部ソフトウェア開発室第 4G リーダー	
	氏 名	大橋 英之	

（令和 4 年度募集）第 35 回 助成研究 完了報告書

上記様式記載後

1. 実施内容および成果ならびに今後予想される効果の概要（1, 0 0 0 字程度）

※産業技術として実用化の可能性や特許出願（予定も含む）の有無についてもご記載ください。

本研究は、安価なローリングシャッター方式の CMOS カメラを用い、無音映像から物体表面のごく微細な振動を捉えて音響信号を復元する「ビジュアルマイクロホン」の実用化と実環境適応を目指して実施したものである。本技術は、従来の物理マイクでは困難な遠隔からの特定箇所の音の取得や、騒音下における特定音源の分離を可能にする新しい計測手法として期待される。研究では、信号復元性能の向上、打音識別への応用、振動源の検知・可視化の 3 点を中心に検証を進め、基盤技術としての有効性と応用可能性を多面的に明らかにした。

第一に、撮影原理により生じる周期的な信号欠損に対し、深層学習を用いて失われた音の情報を補う手法を導入し、数値評価と聴取評価の両面で改善を確認した。さらに、複数台の CMOS カメラを用いて欠損を補う方法も検討し、模擬信号では性能向上を確認した。第二に、外部雑音の影響を受けにくい特性を活かし、インフラ点検や異常音検知を想定した打音識別に応用したところ、鉄琴を対象とする実験で極めて高い識別精度を達成し、本技術の有効性を示した。第三に、撮影画像から物体表面の複数個所の振動の大きさを抽出して比較することで、振動源の位置の検知と可視化を検証し、1 台のカメラで振動源位置を概ね特定できること、さらに 2 台のカメラを用いることで振動源の方向推定にも可能性があることを示した。これらの成果は、単なる音響復元にとどまらず、非接触センシング技術としての展開可能性を示すものである。

一方で、実環境では高周波成分の減衰や記録時に生じる信号の粗さの影響により、欠損補間や高音域識別の改善が限定的となる課題も明らかとなった。しかし、これらは実用化に向けて克服すべき課題を具体化し、今後の改良方針を明確にした成果でもある。本技術は、監視カメラやドローン等への搭載が見込まれ、工場騒音下での異常音検知やインフラの打音検査など、産業技術として幅広い応用が期待される。特許出願は現時点では予定していないが、今後、信号抽出段階での情報劣化の抑制と欠損補間性能の向上を進め、実環境で安定に動作するシステムへ発展させることで、知財化も視野に入れている。

1. 実施内容および成果の説明（A 4 で、5 ページ以内）

2.1 欠損補間手法の構築

(1) 単一カメラにおける推定補間手法の検討

まず、単一カメラ環境下において、既存の信号処理ベースの補間手法（線形補間、後方信号の基本周波数を用いた補間、自己回帰モデルを用いた補間など）の適用性を検証した。その結果、欠損区間が長い安価なカメラの抽出信号においては、位相の不整合による歪みや予測信号の発散が生じ、十分な音質改善が困難であった。これに対する改善策の 1 つとして、欠損前後の信号のパワースペクトルから最大ピーク周波数成分を抽出し、重み付け加算により信号を生成・接続する手法を構築した。この手法により、純音に対しては滑らかな接続と SNR の改善が見られたが、参照する周波数が 1 つに限られるため、倍音を含む複雑な音響信号への適用には限界があった。

そこで、長期的なオーディオ信号の欠損を埋める深層学習タスク Audio Inpainting に着目し、拡散モデルである CQT-Diffusion を用いた生成

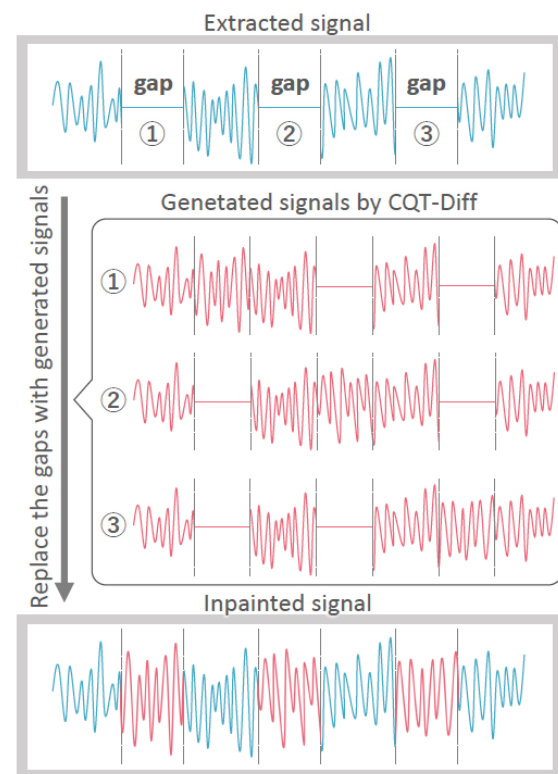


図 1 深層学習を用いた欠損補間手法の概要

的補間手法を提案した（図 1 参照）．本手法は，定 Q 変換（CQT）を前処理に用い，ピアノ音などのように強い調波構造を持つ信号の分析に適している．さらに，カメラの撮像特性による高周波数帯域の減衰を考慮し，帯域制限を施したデータセットを用いてモデルの再学習を行った．模擬欠損信号および実際のカメラ抽出信号を用いた評価実験の結果，従来手法と比較して客観評価指標 ODG スコアが改善し，主観評価においても聴感上の音質の向上が確認された．ただし，周囲の有効な信号自体が大きく情報劣化している実環境下では，ピークの完全な再現が難しくなる課題も明らかとなった．

(2) 複数カメラ（カメラアレイ）を用いた実測値補間手法の検討

単一カメラでの推定補間の限界を克服する別のアプローチとして，2 台の CMOS カメラで同一音源を同時に撮影し，カメラ間で欠損タイミングがずれている場合に，図 2 のように一方の欠損区間を他方の実測信号で置換するカメラアレイ構成の補間手法を検討した．抽出信号間には撮影開始のずれ等による遅延が存在す

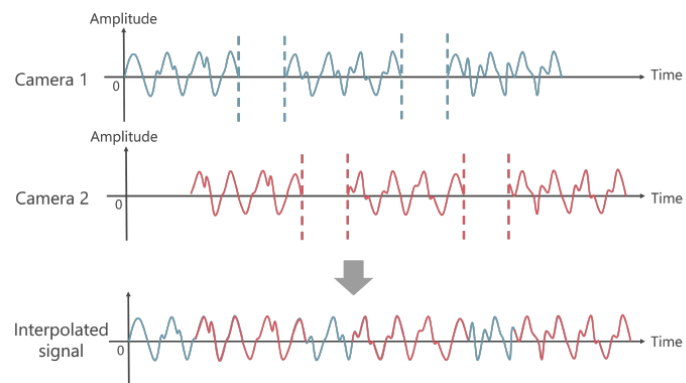


図 2 カメラアレイを用いた補間手法の概略

ると，信号間の正確な時間対応付けが不可欠となるが，本研究では，動的時間伸縮法（DTW）を拡張し，欠損値を含む時系列データにも適用可能な「DTW-AROW」を導入した．さらに，音響信号特有の周期構造による対応付けの誤りを防ぐため，ヒルベルト変換による信号の包絡線を用いて時間整列を行う手法を提案した．

ビジュアルマイクロホンの特性を模擬した信号を用いた評価実験では，単純な時系列置換と比較して，提案手法が PESQ（音質）および STOI（明瞭度）の両客観評価指標において有意な改善を示し，時間対応付けの有効性が実証された．しかしながら，実際のカメラから抽出された信号を用いた検証では，客観評価指標における有意な改善は認められなかった．この原因として，振動物体による高周波成分の物理的な吸収による波形の単純化や，解像度の限界による振幅の粗い量子化など，抽出段階での信号劣化が大きすぎるため，DTW による時間整列の優位性が発揮されなかったことが考えられる．

2.2 打音識別への応用

(1) 打音抽出の検証と識別実験

まず、CMOS カメラを用いて鉄琴の打音による物体振動を撮影し、音波抽出が可能か検証した。その結果、高周波の抽出は難しいものの、低周波の抽出は十分に可能であることが確認された。さらに、識別に使用可能な周波数帯域を調査したところ、基音が 2000 Hz までの打音に対しては、4000 Hz までの帯域が識別に有効であることが分かった。これを踏まえ、鉄琴の 2 つの音階 (F5, C6) を対象にサポートベクトルマシンを用いた識別を試みたところ、周波数ピーク位置の違いが明確であったこともあり、全てのサンプルで正しく識別ができた。これにより、CMOS カメラを用いた打音識別の基本的な可能性が示された。

より実用的なタスクを想定し、識別対象を鉄琴の高音域を含む 7 音階 (C7~B7) に拡張して識別実験を行った。その結果を図 3 に示す。全体の正解率は 94.2%と高い水準を維持した一方、図 4 に示すように、音階が高くなるにつれて撮影画像に表

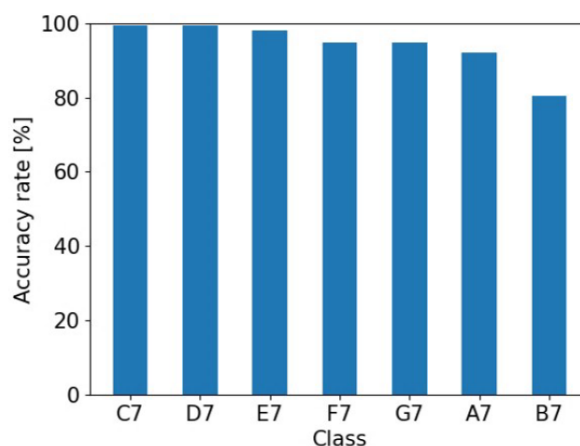


図 3 打音クラスごとの正解率

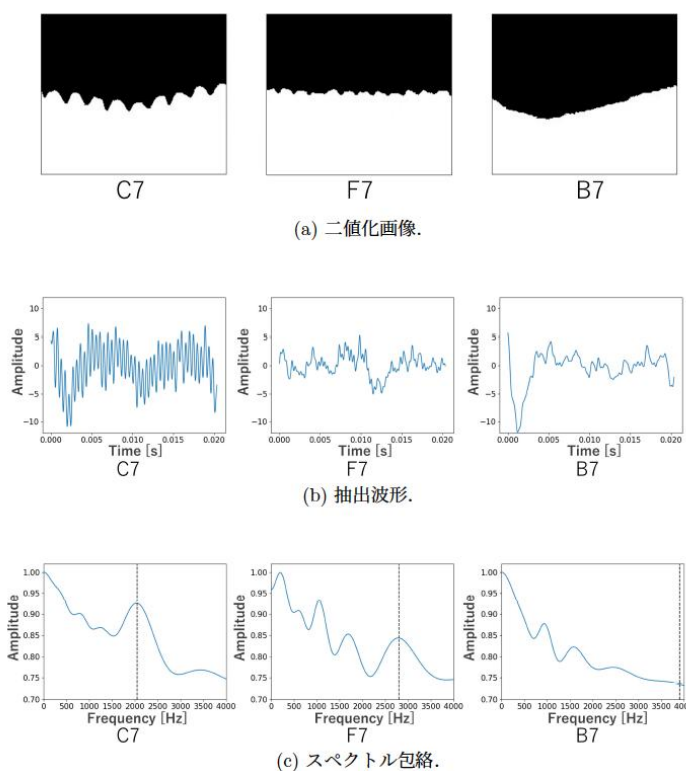


図 4 振動の微小化による情報欠落の可視化.

れる振動が微小になり、それに伴ってスペクトル情報が失われてしまうために、特徴抽出が困難になることで識別精度が徐々に低下する傾向が明確に確認された。この結果から、識別精度を根本的に向上させるには、微小な振動となる高音の抽出精度の改善が不可欠であることが示唆された。

(2) 識別精度向上のための対策検証

高解像度・高倍率カメラの導入といった機材コストを上げずに高音域の識別精度を向上させるため、抽出信号に対するスペクトルサブトラクションと、撮影画像の超解像化という 2 つのアプローチを検証した。

第 1 のアプローチであるスペクトルサブトラクションは、抽出信号の無音区間から背景雑音を推定して減算する手法である。適用させたところ、低音域のスペクトルは強調されたものの、高音域では振動自体が微小すぎるために雑音とともに目的音のピークをも消滅させる場合があり、本来の目的は達成できないことが明らかとなった。第 2 のアプローチとして、撮影動画像を Bicubic 法などの補間型超解像手法によって縦横 2 倍に超解像化し、そこから音波を抽出する実験を行った。しかしながら、原画像から抽出した波形と超解像化画像から抽出した波形に有意な違いは見られず、識別精度の向上は認められなかった。これは、周囲の画素値から計算する単純な補間型手法では、撮影時に欠落した微小な振動情報を新たに復元することはできないためであることを示唆している。同時に、大量の振動情報を学習した深層学習ベースの超解像手法の検討の必要性が促されている。

2.3 振動分布の可視化と振動源検知

(1) 単一カメラを用いた複数箇所の振幅比較による振動源検知

まず、単一の CMOS カメラを用いて振動物体表面の複数箇所を撮影し、各箇所から抽出された信号の振幅差を比較する手法を検証した。アルミニウム板および鉄琴を対象とし、表面に設定した計 9 箇所の観測点において、打撃による微小変位を画像から抽出した。得られた各箇所の振幅の最大値・最小値の差を算出し、ヒートマップとして振動分布を可視化した結果、振幅が最大となる箇所と実際の振動源となる打撃位置が概ね一致することを確認した。特に鉄琴を用いた検証では、実際の振動源と検知された位置がほぼ正確に一致し、本手法の有効性が示唆するものとなった。

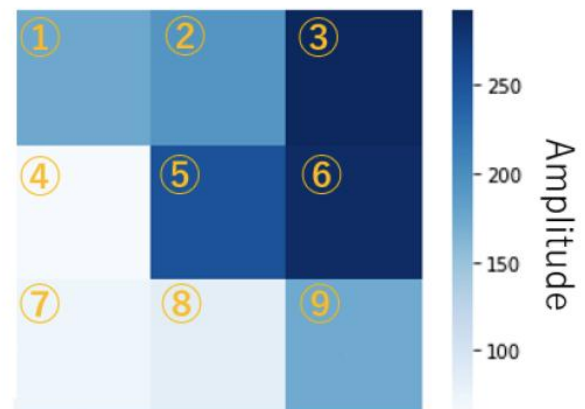


図 5 ③の位置でアルミニウム板を打撃した場合のヒートマップによる振動源の可視化。

一方で、物体を固定している箇所、すなわち振動の節となる付近では振幅が減衰し最小となるなど、物体の振動モードが振幅分布に影響を与えることも明らかとなり、固定条件や対象物

の材質に応じた配慮が必要であることが示唆された。

(2) 複数カメラを用いた異なる視点からの振動源方向推定

次に、空間的な振動源の方向推定へと拡張するため、図 6 に示すように 2 台の CMOS カメラを用いた非同期分散型の観測システムを構築した。振動源として振動物体に吸着可能なスピーカを用い、その設置位置を振動物体の右・中央・左と変化させながら、異なる角度から 2 台のカメラで同時撮影を行った。両カメラから抽出された信号の時間平均の振幅値およびスペクトログラムを比較した結果、振動源に近い側

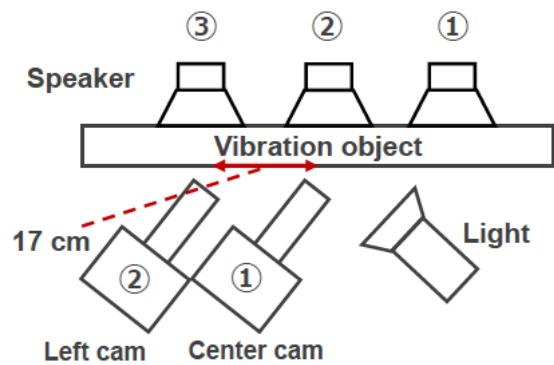


図 6 2 台の CMOS カメラによる振動源方向推定実験の機材配置図

のカメラから抽出された信号の方が振幅値が大きく、スペクトログラムにおいても目的周波数のパワーがより明瞭に強く表れることが確認された。2 台のカメラ配置を入れ替えた追加実験でも同様の傾向が確認され、機材の性能差ではなく、振動源とカメラとの空間的な位置関係に起因する特徴であることが証明された。これにより、複数カメラから得られる振幅値と周波数パワーの差異を比較することで、振動源の方向を概ね推定できることが示唆された。