CH-007

イベントカメラによる振動解析に基づいた高解像度光学的音響計測

白川 稜[1,a)]　曽我部 陽光[1,b)]　杉本 志織[1,c)]　松本 鮎美[1,d)]　北原 正樹[1,e)]

概要：光学的音響計測とはマイクロホンを用いた一般的な音響計測手法と異なり，光学デバイスを用いて音の振動に起因する光の変動を捉えることで音響を計測する技術である．レーザードップラー振動計や高速度カメラを用いた様々な手法が開発されている一方で，デバイス構成の複雑さや価格，時間分解能といった面で課題が残る．イベントカメラは被写体の輝度変化を検知するデバイスであり，高時間分解能，高ダイナミックレンジ，低消費電力といった特徴から多くの分野でその活用が期待されている．加えて，既存手法で用いられるデバイスと比較してイベントカメラのデバイス構成は非常に簡素かつ安価であることから，本稿では，イベントカメラを用いた被写体の振動解析及びそれらの情報を基にした高解像度な音響推定技術を提案する．

High-resolution Optical Sound Measurement based on Vibration Analysis with Event Camera

1. はじめに

音波とは空気の密度変化であり，音波が物体にぶつかることで対象の動きや形状変化を引き起こす．これらは振動として表現され，人間は鼓膜の振動を音として認識する．音の計測手法としてもっとも一般的であるマイクロフォンは，音波によるダイアフラムの振動を直接電気信号に変換することで計測を行う．近年のマイクロフォンは近くの音を非常に高精度に計測することが出来る一方で，音の減衰や遮蔽，指向性といった面から用途によっては所望音が計測できない場合がある．これらの解決策として期待されているのが光学的音響計測技術である．光学的音響計測とはマイクロフォンのように振動を直接電気信号に変換するのではなく，物体振動に起因する光の変動を解析することで音を計測する手法である．光は減衰率が低いことや半透明媒体に対する透過性，強い直進性を持つといった特徴があり，音波とは異なる伝達特性を有することからマイクロフォンには不向きな状況への応用可能性がある．

光学的音響計測の実装としてはレーザー反射光との干渉を利用したレーザードップラー振動計や高速度カメラを用いた手法 [1], [2] が開発されている一方，現状の実装手法の多くはデバイス構成の複雑さや価格，時間分解能といった面で課題が残る．したがって，本稿では高解像度な音響情報をより簡素かつ安価なデバイスを用いて計測することを目的とする．

イベントカメラ [3] は被写体の輝度変化を検知し各画素で非同期に出力する特殊カメラであり，高時間分解能，高ダイナミックレンジ，低消費電力といった特徴をもつ．加えてデータ取得の非同期性によりデータ効率が高い．デバイス構成についてはイメージセンサとレンズといった一般的なカメラと同様の構成であり，従来手法のデバイスと比較すると非常に簡素かつ安価である．そこで本稿ではイベントカメラの上述した利点に注目し，イベント情報を用いた被写体の振動解析及び音響推定手法を提案することで高解像度かつ簡便な光学的音響計測を検討する．

[1] 日本電信電話株式会社 NTT コンピュータ&データサイエンス研究所
〒 239–0847 神奈川県横須賀市光の丘１－１
[a)] ryo.shirakawa@ntt.com
[b)] yoko.sogabe@ntt.com
[c)] shiori.sugimoto@ntt.com
[d)] ayumi.matsumoto@ntt.com
[e)] masaki.kitahara@ntt.com

図 1: イベントカメラ (左:カメラ本体，右:イベント画像)

2. 関連研究

2.1 Visual Microphone

Visual Microphone[1] は高速度カメラを用いた光学的音響計測手法である．レーザードップラー振動計などの従来手法と異なり，被写体に対してレーザや照明等を当てる必要のない受動的な撮影を行う．撮影した高速度動画に対して complex steerable pyramiad[4] と呼ばれるガボールフィルタに似たウェーブレット変換を行い，各フレーム間の位相差を求めることで音の推定を行う．

利用する高速度カメラのサンプリング周波数は最大 20kHz を想定しており，一般的なマイクロフォンのサンプリング周波数が 44.1kHz であることを踏まえるとマイクロフォンと同等の音をとるには不十分である．ハイエンドのハイスピードカメラには 50kHz を超えるものもある一方で非常に高価であることから価格面において課題が残る．

2.2 Event-based Visual Microphone

Event-based Visual Microphone[5] はイベントカメラで取得した情報に Visual Microphone の音の推定手法を適用したものである．イベントカメラは各画素で非同期に情報を取得し同一時刻のフレームを持たないためフィルタ計算が不可能である．本技術では取得した全イベントに対して各イベント毎に近傍領域を利用した疑似フレームを定義することでフィルタ計算を実行する．

本技術は本稿と同様にイベントカメラに注目した音響計測手法である．イベントカメラの特徴である高時間分解能とデータ効率の高さを活用した手法であるが，全イベントデータに対する疑似フレームの定義とフィルタ計算が必要であることを考慮すると多くの計算量を要する．本稿では Visual Microphone の音の推定手法に依存せず，単純な周波数解析手法を利用することでより軽量な推定アプローチを検討する．

3. イベントカメラ

イベントカメラ [3](図 1) は生物の網膜を模倣した特殊なセンサにより，輝度の変化を非同期に検知しイベントとして出力する．各イベントは以下のような情報を持つ．

$$e = (x, y, p, t)$$

ここで $(x, y) \in \mathbb{Z}^2$ はイベントが発生した画素の座標，$p \in \{-1, 1\}$ は極性と呼ばれる対数輝度勾配であり，t はタイムスタンプである．

イベントカメラの特徴は高時間分解能，高ダイナミックレンジ，低消費電力という点であり，時間分解能については最大サンプリング周波数が 1MHz である．これは一般的なマイクロフォンのサンプリング周波数を大きく上回っており，性能面ではマイクロフォンと同等の音域を計測できる可能性がある．価格については 1 台数十万円であり，高速度カメラと比較すると十分安価である．

4. 提案手法

本章では音波を受けて振動する被写体から得られるイベントデータに対する周波数解析に基づく音の推定手法を提案する．

はじめに振動の周波数について考える．被写体が振動している場合，被写体表面のコントラストがある部分や被写体と背景とのエッジ部分の画素では周期的に輝度が変化する．したがって，各画素から出力されるイベント列を周波数解析にかけることで被写体がどのような周波数で振動しているかが推定されるものと期待できる．次に振動の強度について考える．振動の強度が強い，つまり振幅が大きいということは単位時間当たりの移動距離が長いということであり，単位時間あたりの輝度の変動が大きく，より多くのイベントが発生すると考えられる．したがって，単位時間当たりのイベント数を振動の強度に関係づけることができる．

上述した内容を踏まえて提案手法を説明する．なお，本稿では解析に十分なイベント数を確保するため計測範囲内の複数画素からイベントデータを取得する．提案手法を以下に示す．

まず，イベントデータ $e = (x, y, p, t)$ を計測範囲 X, Y 内の各画素から集め，イベントデータ集合 E を作る．以下ではイベントデータを $e = (p, t)$ と表記する．

$$E = \{e(p_1, t_1), e(p_2, t_2), \ldots, e(p_n, t_n)\}$$

次に，推定音響のサンプリング周波数 $\hat{f}$ とイベントカメラのサンプリング周波数 f_e に基づいてサンプリング単位 $\Delta t = \frac{f_e}{\hat{f}}$ を決定し，E についてサンプリング区間ごとのイベント数をカウントし，これを全区間並べたものを信号 $S(u)$ とする．

$$S(u) = \sum_{i=1}^{n} \mathbb{I}(t_i, u)$$

$$\text{where} \quad \mathbb{I}(t_i, u) = \begin{cases} 1 & \text{if } u\Delta t \leq t_i < (u+1)\Delta t \\ 0 & \text{otherwise} \end{cases}$$

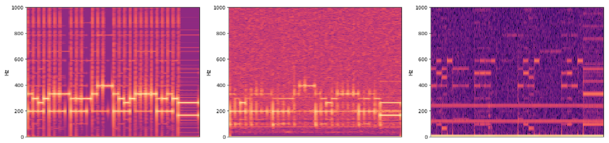


図 2: シミュレーション結果 (左:正解音源，中:Visual Microphone による推定，右:提案手法による推定)

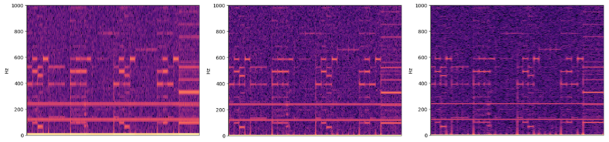


図 3: 推定音響のサンプリング周波数の影響 (左:44.1kHz，中:22.1kHz，右:11.0kHz)

最後に，$S(u)$ について短時間フーリエ変換をかけることで振動の周波数と強度を得る．

5. 実験

本章では [1] で公開されるデータを用いたシミュレーション実験並びに実際に撮影したデータを用いた実環境実験を通して提案手法を評価する．

5.1 シミュレーション実験

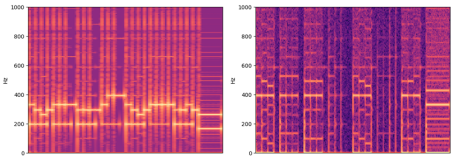
[1] で公開される高速度動画 (Chips2, 2200Hz, Mary had a little lamb MIDI) に対して Metavision SDK[*1] で用意されるフレーム-イベント変換を行うことでシミュレーションデータを作成する．短時間フーリエ変換の fft サイズは 4096，スライドサイズは 1024 に設定する．

はじめに，推定音響のサンプリング周波数を 44.1kH に設定した際の提案手法の結果と Visual Microphone の結果[*2]を比較する．正解音源とそれぞれの推定結果のスペクトログラムを図 2 に示す．図 2 を見ると提案手法の推定結果は正解音源よりも高周波側にシフトしている．これはイベントをサンプリング区間でカウントする際，振動方向を考慮していないことが原因だと考えられる．例えば単純に単振動を想定すると，被写体の移動方向が正の場合と負の場合とで，その振動の一周期内に二つのイベントが観測される．よって，一つの振動周期における輝度値の変化回数は，振動の回数の二倍となり，変化周期は半分となる．このことから本提案手法で推定される音響は実際の音響の倍の周波数で観測されることが予想され，実際の結果もそのとおりである．全ての振動に対して上記の振る舞いが発生するため，推定結果の全ての周波数を 1/2 に下げることで解決が可能である．また 120, 240Hz 辺りに定常信号が見られるが，これは電源周波数 60Hz の地域で発生する蛍光灯のフリッカーが原因だと考えられる．その他の信号については Visual Microphone の結果と比較しても明瞭であり，実際の音響を聞いても全ての高さの音が確認できた．

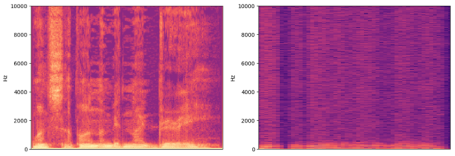
次に推定音響のサンプリング周波数を変動させた際の影響を検証する．推定音響のサンプリング周波数が低いということはイベント数をカウントする際のサンプリング区間が長いということであり，より頑健性の高い推定結果となることが予想される．図 3 に推定音響のサンプリング周波数を 44.1kHz，22.1kHz, 11.0kHz に設定した際の推定結果を示す．この結果から推定音響のサンプリング周波数を低く設定するほど推定結果の周波数の広がりが抑えられ，より明瞭な推定結果となることがわかる．ただし，サンプリング周波数を低くすることは計測可能な最大周波数の値を下げることにもつながるため，計測対象に合わせた適切な設定が必要である．本稿以下では 22.1kHz に設定する．

[*1] https://docs.prophesee.ai/stable/index.html
[*2] http://data.csail.mit.edu/vidmag/VisualMic/#results

図 4: 撮影環境



(a) Mary had a little lamb MIDI



(b) ”Once upon a midnight dreary”

図 5: 実環境実験の結果 (左:正解音源，右:推定音響)

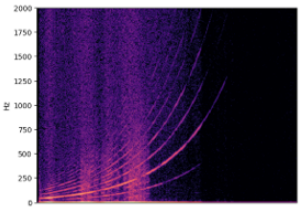


図 6: 計測可能範囲

5.2 実環境実験

次に実際にイベントカメラを用いた撮影及び音響推定の

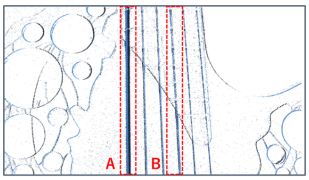


図 7: ギターのイベントデータ

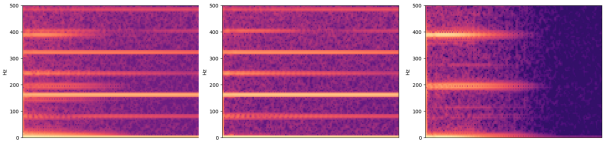


図 8: 領域毎の推定結果 (左:全体，中:領域 A，右:領域 B)

実験を行う．使用したイベントカメラは SilkyEvCam HD，被写体はポテトチップスの袋とし，スピーカーから出力する音によって振動を加える．撮影環境を図 4 に示す．正解音源には [1] で公開される音楽及び発話 {Mary had a little lamb MIDI, ”Once upon a midnight dreary”} のデータを利用し，結果を図 5 に示す．図 5 のスペクトログラム及び実際の推定音響を確認すると，音楽の音源については十分明瞭に推定できており全ての高さの音の聞き取りが可能である．一方で発話のデータについては，単語間の区切りは確認できるものの発話内容については推定音響がボケており聞き取りが不可能であった．音源のスペクトグラムから見てわかるように対象の発話は非常に低い周波数の音から構成されていることに加えてより複雑な周波数の広がりを持つ．推定音響では低周波の音は比較的よく推定できているが高周波の複雑な広がりは推定できておらず，この部分が発話内容の取得に重要であることが想定され提案手法では不十分な点である．

5.3 計測可能範囲

上述した提案手法の課題を明確にするため，提案手法の計測可能な周波数範囲を検証する．音源に 20Hz-20kHz のスイープ音を使い音響推定を行うことで上限と下限を調べる．被写体はポテトチップスの袋を使い，結果を図 6 に示す．図 6 及び上述した推定音響の周波数が倍になることを踏まえると計測上限は 600Hz 程度であり下限は 20Hz に近い音が取れることが分かる．高周波の音ほど被写体が振動しにくく，イベントとして観測できないため音響推定が困難となった．高周波の音を計測するためには，現撮影環境では不十分であり，より大きな音源や振動しやすい被写体，振動が計測しやすくなるような工夫が必要である．

5.4 音源分離

カメラは空間情報を扱えることから，光学的音響計測では指向性を考慮することができる．そこで本実験では被写体の領域毎の音響推定結果を用いて音源の分離を行う．図 7 に示すようにギターの 3 弦と 6 弦を同時に鳴らし，それぞれの弦に対応する領域について音響推定を行う．図 8 に全体及びそれぞれの領域での推定結果を示す．領域全体からの推定結果では 2 本の弦の周波数及び高調波・低調波が混ざっているが，領域 A からの推定結果は 6 弦について，領域 B からの推定結果は 3 弦についての周波数情報しかなく綺麗に分離ができていることがわかる．計測可能なイベント数に依存するものの，より多くの領域での分離も可能であり，この点がマイクロフォンに対する大きな利点である．

5.5 計算時間

提案手法は単純な信号処理のみからなるため非常に高速な計算が可能である．CPU:Intel(R) Core(TM) i7-1065G7 1.30GHz, memory:16GB, OS:Ubuntu20.04LTS のマシン上で [1] の (Mary had a little lamb MIDI) データを処理するのに要した計算時間は 1.2 秒であった．これは Visual Microphone の計算時間 (2-3 時間) と比較して非常に高速であることがわかり，リアルタイムでの推定を可能にし得る．

6. まとめ

本稿では，イベントカメラを用いた振動解析に基づく光学的音響計測手法を提案した．従来手法と比較して本手法は非常に高速に計算可能であり，良好な推定結果を示した．一方でイベント出力の数が著しく減少することから高周波の音の推定精度は不十分であり課題が残る．イベント出力数の増加のための撮影環境の調整並びにより高度な推定アルゴリズムの検討を今後の展望とする．

参考文献

[1] Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G., Durand, F. and Freeman, W. T.: The Visual Microphone: Passive Recovery of Sound from Video, *ACM Transactions on Graphics (Proc. SIGGRAPH)*, Vol. 33, No. 4, pp. 79:1–79:10 (2014).

[2] Nassi, B., Pirutin, Y., Shamir, A., Elovici, Y. and Zadov, B.: Lamphone: Real-Time Passive Sound Recovery from Light Bulb Vibrations.

[3] Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K. et al.: Event-based vision: A survey, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 44, No. 1, pp. 154–180 (2020).

[4] Portilla, J. and Simoncelli, E. P.: A parametric texture model based on joint statistics of complex wavelet coefficients, *International journal of computer vision*, Vol. 40, pp. 49–70 (2000).

[5] Niwa, R., Fushimi, T., Yamamoto, K. and Ochiai, Y.: Live Demonstration: Event-Based Visual Microphone, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4053–4054 (2023).