

ハイスピード映像からの音源復元のための物体振動抽出手法の検討*

☆安見祐亮, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

本稿では、音響信号の影響により発生した（音源から離れた）物体の微細振動をハイスピードカメラで捉えることにより、映像から音響信号を復元する手法を検討する。従来、遠距離から物体の振動を取るための手法においてレーザー光などが用いられていた。近年、文献 [1] では、映像中の目に見えない物体の微細振動をハイスピードカメラで捉える手法が提案されている。この手法では、物体を撮影するための光以外に必要なものが無いので、センサなどの追加部品を必要とせず、また、カメラの物体への向きに関する大きな制約もない。監視や安全保障などの分野にて、応用が期待されるほか、物体の振動に対する特性の計測にも利用が可能である。筆者らはこれまで、外乱の影響下において映像中の物体振動から音を復元する研究をしてきた [8]。本稿では、振幅の大きな画像領域、位相差の大きい画像領域に注目した振動抽出手法を検討する。

2 音の復元について

2.1 従来研究

画像の輝度プロファイルを画像中の座標 $\mathbf{x}$ によって $f(\mathbf{x})$ と表現する。このとき、画像をフーリエ級数分解によって複素正弦波の集合で表すと次のようになる [2]。

$$f(\mathbf{x}) = \sum_{\omega=-\infty}^{\infty} A_{\omega} e^{i\omega\mathbf{x}} \quad (1)$$

このとき、フレーム t における画像中の点の時間移動が置換関数 $\delta(t)$ で表されるとすると、画像中の点の移動は位相における変化にのみ表れ移動後の画像は次のように表される [2]。

$$f(\mathbf{x} + \delta(t)) = \sum_{\omega=-\infty}^{\infty} A_{\omega} e^{i\omega(\mathbf{x} + \delta(t))} \quad (2)$$

よって物体の振動を取り出す際には、参照フレームとフレームとの位相差を取り出せばよい。

文献 [1] においては、Complex Steerable Pyramid [3] を特徴量として用いることにより、局所的な移動を位相変化として取り出している。この特徴量により、画像は周波数の大きさ（スケール r ）、周波数の方向（オリエンテーション θ ）ごとにウェーブレット

分解される。画像分解の例を Fig. 1 に示す。2 スケール 2 オリエンテーションで分解した。

このとき各ウェーブレットは実部虚部の両方を保持しており、極座標変換により振幅 $A(r, \theta, \mathbf{x})$ と位相 $\phi(r, \theta, \mathbf{x})$ に分解できる。各ウェーブレットは以下のように表現される。

$$A(r, \theta, \mathbf{x}) e^{i\phi(r, \theta, \mathbf{x})} \quad (3)$$

画像中の物体の移動量が小さいとき、各ピクセルにおける位相の差がそのウェーブレットにおけるピクセル毎の移動を表す。なので、ピクセル毎に時点 t における参照フレーム t_0 との位相差 $\phi_v(r, \theta, \mathbf{x}, t)$ を求める。

$$\phi_v(r, \theta, \mathbf{x}, t) = \phi(r, \theta, \mathbf{x}, t) - \phi(r, \theta, \mathbf{x}, t_0) \quad (4)$$

ウェーブレット毎に全ての位相差を足し合わせることで、各ウェーブレットの移動量 $\Phi(r, \theta, t)$ を求める。ただし、振幅の小さいピクセルはノイズが多いため各ピクセルの振幅の自乗を位相差に重み付けする。

$$\Phi(r, \theta, t) = \sum_{\mathbf{x}} A(r, \theta, \mathbf{x}, t)^2 \phi_v(r, \theta, \mathbf{x}, t) \quad (5)$$

最後に各ウェーブレットごとの移動量を最も移動量が大きくなるように足し合わせることで映像全体の移動量を求める。物体の移動が音によるものであるから、この映像全体の移動量が復元音を表す。また、この手法においてはノイズ処理として、バタワースフィルタによるナイキスト周波数の $1/20$ 以下の周波数のカットと、スペクトルサブトラクション [4] や音声強調 [5] による処理が行われている。

2.2 提案手法

本稿では、音質を改善するための物体振動の抽出手法として、特定の部分からのみ変動を取り出すことを考える。従来手法の強みは、物体の細かな変動を取り出せるようにするとともに、変動している部分を特定できるということである。細かな変動を取り出すだけであれば、画像の周波数領域で位相差を求めるだけでも取り出すことはできる。Fig. 2 に位相差の例を示す。入力フレームは参照フレームから半フレーム右に移動させたものである。位相差は白が π で、黒が $-\pi$ である。オリエンテーション 1 は周波数領域の右側

*Extracting object's vibration in a high-speed video for sound recovery. by Yūsuke Yasumi, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

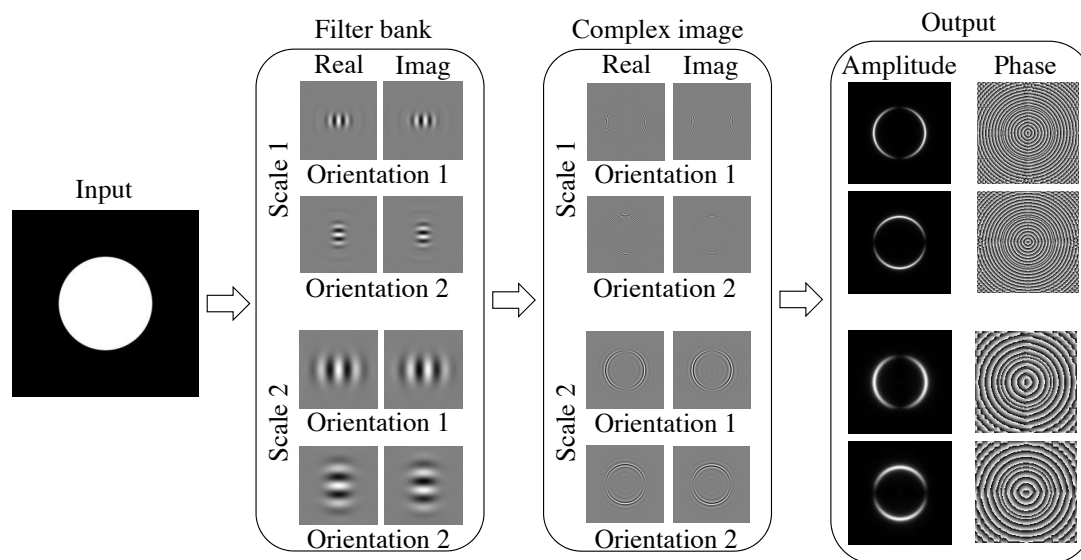


Fig. 1 Image decomposition

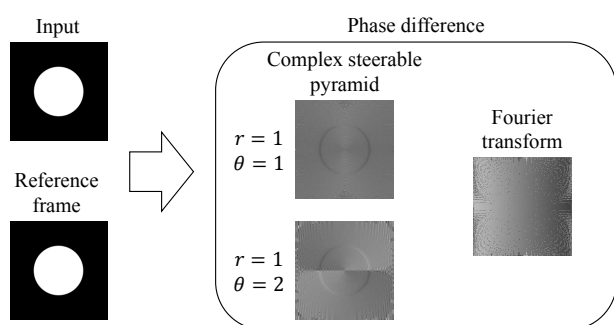


Fig. 2 Phase difference

の一部を逆変換したものであり、図からウェーブレットにおける変動の方向が周波数領域の変動の方向と同じであることがわかる。Fig. 2 から、画像の周波数領域からでも移動の方向はわかるが、従来手法の場合は実際に移動している部分を知ることができる。

実際の物体は様々な方向に振動するため、同じウェーブレット内でも全てのピクセルが同じ方向に変化することは少ない。そのため、全てのピクセルでの移動量を足し合わせる際に移動量の減衰が起こっていると考えられる。この減衰の影響を小さくするために、振動している部分の中でもできる限り同じ方向に変化している部分を特定したい。文献 [7] では、画像をいくつかのパッチに分割してそれぞれのパッチ毎に音を復元することで、画像全体で復元するよりもよい

結果が得られたと報告されている。画像を小さく分割したほうがよい結果が出ているということからも、画像中の復元に不要な部分を取り除く必要があるということがわかる。よって、様々な方法で音の復元に必要な部分を取り出してみる。

まず最初の方法として、振幅による重み付けではなく、ある一定値以上の振幅になる部分のみから復元を行ってみる。振幅の小さな部分は位相差にノイズが大きくなるため取り除く必要がある。しかし、振幅の平均値で分類する場合、埃や撮影の際の光学ノイズの影響を受けることがある。また、各時点での振幅の最大値で分類を行う場合にも同じく埃や光学ノイズの影響を受けることがある。そのため映像全体での振幅の最小値で分類を行う。分類結果に従い取り出された部分からのみ位相差を取り出し、重み付け無しで足し合わせたものを移動量とし音を復元する

次に位相差の変動量の大きな箇所から取り出すという方法である。物体には変動しやすい部分とそうでない部分がある。そのため、映像中において同程度の振幅値であっても位相変化が大きい部分とそうでない部分が存在する。物体は音の周波数によっても振動する部分に変化する。これは物体が振動する際に周波数によって異なる形に変形するためである。Fig. 3 に周波数による物体の振動形態の違いの例を示す。この物体はコップにスーパーのレジ袋の切れ端で蓋をし、輪ゴムでとめたものである。各周波数で振動しているときのある時点での物体の位相差を表し

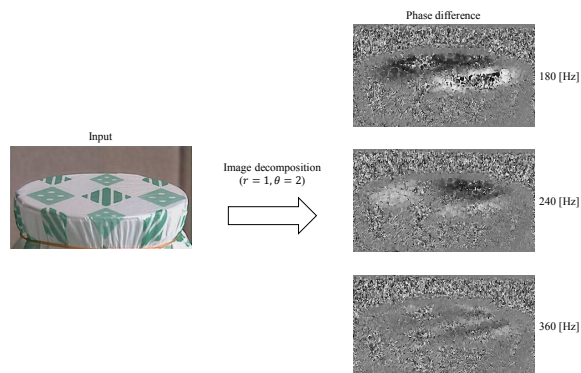


Fig. 3 Phase difference at each frequency

ている。画像はスケール 2 オリエンテーション 2 で分解され、Fig. 3 に示されているのはスケール 1 オリエンテーション 2 のウェーブレットにおける位相差である。周波数によって異なる振動形態をとっていることがわかる。これらを考慮し、今回は事前に映像全体の位相変化を調査してあらかじめ取り出す部分を決めておくことにした。Fig. 3 からわかるように、物体は様々な方向に振動するが振動の周波数に応じて周囲と同じ方向に振動する部分がいくつか生じる。振動の周波数毎に同じ方向に振動する部分の大きさは異なり、高周波数であるほど小さくなる。そして、高周波数においても周囲が同じ方向に移動している部分は低周波においても同様に同じ方向に振動していることが多い。そのため振動が高周波であるときに周囲が同じ方向に振動している部分を音を取り出す部分とした。こうして決められた部分から取り出された位相差を重み付け無しで足し合わせたものを移動量とし音を復元する。

最初の方法では振動の変化形態が考慮されていないため、異なる方向の移動による減衰が起こると考えられる。しかし、振幅の最小値の基準の調整によりこの減衰がなくなることも考えられるため、実験を行うことにした。

3 実験

3.1 実験条件

使用した映像は、Fig. 2 で用いた物体を撮影した映像で、カメラに対し垂直方向からスピーカを用いて音を当てて撮影した。これらの位置関係は Fig. 4 に示されている。映像は 400×224 ピクセルで、フレームレートが 2200 [Hz] である。また、ノイズ除去についてはバターワースフィルタによる 55 [Hz] 以下の信号のカットオフを行った。音には 100 [Hz] から 1000 [Hz] に五秒間で線形増加する音源を用いた。今

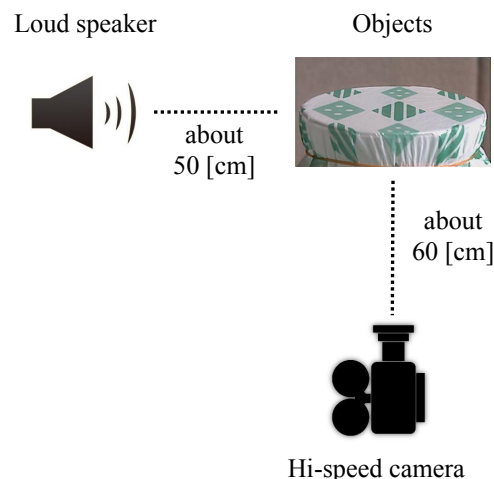


Fig. 4 Experimental setup

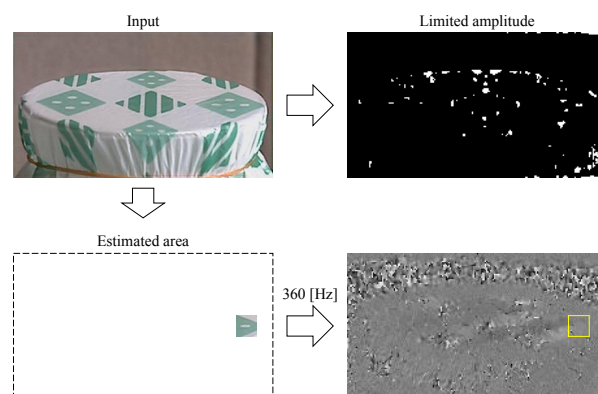


Fig. 5 Selected area

回の実験においては物体の変動が縦方向に大きいことがあらかじめわかっていたため、各ウェーブレットから求めた移動量を足し合わせるのではなく最もよい結果のウェーブレットの結果を用いる。特徴量はスケールが 4、オリエンテーションが 2 である。また、提案手法ではどちらの方法でも取り出すピクセルの数は同じになるようにしており、どちらも 32×32 ピクセル分（スケール 1 において、スケール 2 ではその四分の一）だけ取り出している。Fig. 5 に実際にそれぞれの方法で音を取り出した領域を示す。振幅により分類する手法では白が 1 黒が 0 の値のフィルタとして用いられた。位相差の変動量の大きな部分としては、黄色の線の枠内の部分を 360 [Hz] の位相変化を基に取り出す部分として決めた。

3.2 実験結果

従来手法、提案手法による復元音のスペクトログラムを Fig. 6 に示す。従来手法、提案手法による復元音の波形を Fig. 7 に示す。実験結果から、提案手法

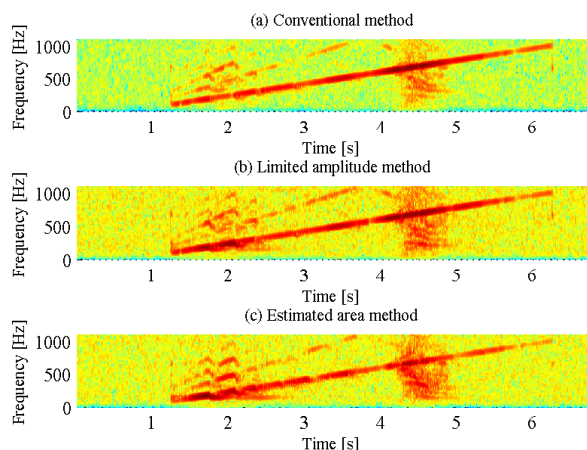


Fig. 6 Spectrogram of recovered sound

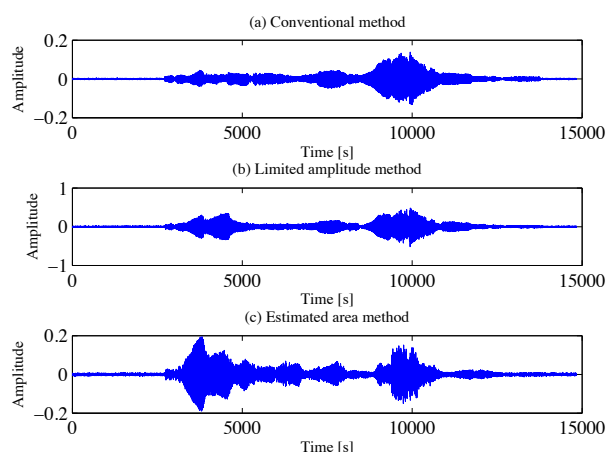


Fig. 7 Waveform of recovered sound

のどちらにおいても全体としてはノイズが少し増えていることがわかる。しかし、180 [Hz] 付近において従来手法ではうまく復元できていなかった場所の復元ができています。特に二番目の提案手法については低周波域では強い信号を復元することができています。

3.3 考察

提案手法において従来手法よりも全体としてノイズが多くなったのは、最初の提案手法ではピクセル数の不足、二番目の提案手法ではノイズ領域の混入が考えられる。最初の提案手法ではピクセル数の条件を二番目の提案手法と合わせるために、最もうまく音を復元できると考えられるピクセル数の最大値まで取り出せていなかったと考えられる。二番目の手法では位相変化を基に取り出す部分を選ぶ際に選択する部分が大きくなりノイズの多い部分も含まれてしまったと考えられる。また、サイズが大きくなってしまったことで高周波では打ち消しあう部分が含まれ

てしまったということも考えられる。特定の部分から取り出す方法では、ある特定の周波数ではうまく取り出すことができていても全体としてうまく取り出すことはできていない。他の物体でも同様であるのか確かめる必要がある。

4 おわりに

今回は、音質を改善できる物体振動の抽出手法として特定の部分から取り出し復元する方法を提案し、実験によりその効果を確認した。今後は、今回の手法の問題点を改善できるように音の周波数毎にうまく取り出すことを考える。

参考文献

- [1] Abe Davis *et al.*, “The Visual Microphone: Passive Recovery of sound from Video,” *ACM Transactions on Graphics*, 33(4), pp. 79:1-79:10, 2014.
- [2] Neal Wadhwa *et al.*, “Phase-Based Video Motion Processing,” *ACM Transactions on Graphics*, 32(4), 2013.
- [3] Javier Portilla *et al.*, “A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients,” *International Journal of Computer Vision* 40(1), pp. 49-71, 2000.
- [4] Steven F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.
- [5] Philippos C. Loizou, “Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum,” *Speech and Audio Processing, IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857-869, 2005.
- [6] John H. L. Hansen *et al.*, “An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms,” In *Proc. of the International Conf. on Speech and Language Processing*, pp. 2819-2822, 1998.
- [7] Juhyun Ahn *et al.*, “Patch-based visual microphone for improving quality of sound,” *23rd International Conference on Pattern Recognition*, pp. 3927-3932, 2016.
- [8] Yusuke Yasumi *et al.*, “Visual Sound Recovery Using Momentary Phase Variations,” *The 23rd International Workshop on Frontiers of Computer Vision*, pp. 2-6, 2017.