一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2018-25,SP2018-5 (2018-06)

映像中の物体振動モードを利用した音源復元

布施　陽平†　安見　祐亮†　滝口　哲也†

† 神戸大学大学院システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †{fuse,yasumi}@me.cs.scitec.kobe-u.ac.jp, ††takigu@kobe-u.ac.jp

あらまし　音波は圧力の変動として周囲に伝播しており，物体に当たるとその表面に微小な振動を起こす．この振動を抽出することで振動の元となる音源を復元する研究が行われている．この技術は遠距離から音を収録できるという特性から，監視や安全保障の分野での応用が期待される．音による物体の振動は目には見えないほど速く微細であるが，物体を撮影したハイスピード映像の各フレームに complex steerable pyramid を用いることで，振動による物体の変化を各ピクセルの移動量として求めることができる．従来法では全ピクセルの移動量を足し合わせるため，音源とは関係のない変化に由来するノイズが混ざる可能性が考えられた．本研究では，周波数に対する物体の振動モードに着目して音を復元する手法を提案する．物体はその材質や形状，周波数などにより振動する部分が異なる場合がある．その振動は幾つかの基準振動の組み合わせで表現され，基準振動ごとに節や腹となる部分が存在する．周波数ごとにどの部分が振動しているのかを確認し，その応答をもとにフィルタをかけて音を復元する．どの部分が振動しているかは，その周波数の信号が最も大きい場合の各ピクセルの振幅応答から確認する．この応答と各ピクセルの信号の信頼度をフィルタとして用いる．実際に複数の物体の映像から音を復元し，手法の有効性を確かめた．
キーワード　音源復元，物体振動，ハイスピード映像，振動モード

Sound recovery using vibration mode of an object in video

Yohei FUSE†, Yusuke YASUMI†, and Tetsuya TAKIGUCHI†

† Graduate School of System Informatics, Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
E-mail: †{fuse,yasumi}@me.cs.scitec.kobe-u.ac.jp, ††takigu@kobe-u.ac.jp

Abstract　When a sound hits an object, it causes the surface of the object to vibrate. Some research has been carried out on the recovering of sounds by extracting the vibrations seen on video images. This research is expected to be applied in the field of surveillance and security because sounds can be recorded from relatively far away. The vibration of objects due to sound is so fast and minute that it is invisible. However, it is possible to observe such changes in objects by using the high-speed video as the movement of each pixel by using a complex steerable pyramid. In the conventional method, the movements of all pixels are added together to recover the sound. So it is possible that some noise source vibrations are mixed because there are some pixels that move independently of the sound source being focused upon. In this paper, we propose a sound recovery method focusing on the vibration modes of the object associated with the frequency. The vibrating parts of objects are different depending on the material, shape and frequency. The vibration is composed of some normal vibrations, and each has different loops and nodes. We confirm which part of the object is vibrating for each frequency of the sound, and recover the sound using a filter based on the response of the object. Which part is vibrating is confirmed from the amplitude response of each pixel when the signal of that frequency is the largest. This response and the reliability of the signal of each pixel are multiplied to each pixel as a filter. We recovered sounds from several objects in videos and ascertained the effectiveness of the method.
Key words　sound recovery, object vibration, high-speed video, vibration mode

1. はじめに

音波は圧力の変動として周囲に伝播するため，物体に音波が当たるとその表面には小さな動きが生じる．物体の特性により変形や移動など動きのパターンは異なるが，いずれの場合においても物体の特性を知るのに十分な情報がそのパターンには含まれている．例えば，映像から物体の振動特性を調査する方法が提案されており，棒状の金属の特定と布の硬さなどの特性の分類実験が行われている [1]．また，物体の振動特性を調査することにより，力が加えられたときの物体の動きを推定することができることも示された [2]．Justin G. Chen らは，風による建物の揺れを撮影した映像から建物の非破壊検査を行えることを示した [3]．本研究ではその情報から音源を復元することを考える．

音波による物体振動の抽出を利用した音源の復元は，直接音を聞き取ることができないような遠く離れた場所で発生した音の収録に利用できる．この特徴から，主に監視や安全保障などの分野における応用が期待される．従来，遠距離から音による物体の振動を取り出す方法としてレーザ光を用いた手法が検討されてきた．このような手法の多くでは物体からのレーザの反射光の位相を測定し，それにより物体の振動を記録した．レーザドップラ振動計 [4] では反射光のドップラシフトを計測し，物体表面の速さを求めることによって音を計測している．このようなレーザ光を用いた手法では，物体表面で光が適切に反射する必要があるほか，レーザと受信機の位置関係なども音の復元精度にかかわってくる．Zalevsky らは，反射光の斑点模様の変化をハイスピードカメラを用いて焦点をはずして記録して音を復元した [5]．

ハイスピード映像のみから音源を復元する手法は Abe Davis らによって考案された [6]．これにより，音波による物体の振動は，それ以前の音源復元の手法において用いられていた特殊なセンサやライトを用いずとも，映像を撮影するのに十分な光量さえあれば抽出することができるということが示された．また，通常のフレームレートの映像の場合においてもセンサの露光の仕方によっては音を復元することができるということも示された．この従来法では，映像の各ピクセルに対し求めた移動量を画像領域全体で足し合わせることで音源波形を復元した．従って，映像中における音源と関係がないような小さな変化もノイズとして足されてしまう可能性があった．[7] では，映像中に風などが原因の大きな変化が存在する場合において，位相差を取り出す方法を工夫することで物体の微小振動を取り出す手法を提案している．

本研究では，音源復元の精度の向上のために，振動モードに着目して物体振動の抽出方法を改善する．物体には材質や形状などに応じて振動しやすい部位やしにくい部位が存在する．そして，物体の振動の仕方には周波数ごとに特定のパターンが存在し，それらは振動モードと呼ばれる．この振動モードに着目することで，音源に由来する物体の振動を強調して抽出することができると考えた．実際には周波数ごとにどの部分が振動しているのかを確認し，その応答をもとにフィルタを作成し映像に適用したのち音源復元を行う．実際に複数の物体の映像から音を復元し，手法の有効性を確かめた．

2. ハイスピード映像を用いた音源復元

2.1 位相変化の抽出

簡単のために一次元の信号 $f(x)$ について考える．この信号をフーリエ級数展開すると信号は正弦波の集合として表される．角周波数を ω，振幅を A_ω とすると，以下のように表される．

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

ω ごとの複素正弦波についてみたとき，x_0 移動後の信号 $f(x - x_0)$ は，以下のように表される．

$$f(x - x_0) = \sum_{\omega=-\infty}^{\infty} A_\omega e^{i\omega(x - x_0)} \tag{2}$$

このとき，移動後の信号と元の信号の差は各複素正弦波における位相部分に表れる．従って信号を位相と振幅に分離させることで，信号の時間ごとの移動距離を位相部分から数値化し取り出すことができる．本研究では Complex Steerable Pyramid を用いて画像を位相と振幅に分解している．

2.2 Complex Steerable Pyramid

本研究は映像中の微小な変化を取り出す技術に基づいている．そのような技術の一つに complex steerable pyramid が挙げられる．Neal Wadhwa らは complex steerable pyramid を用いて画像を部分帯域ごとに分解し，振幅と位相に分離させることにより，位相変化の中に現れる物体の微小な振動をより強調させる映像処理に成功している [9]．また，Complex Steerable Pyramid を性能を維持したままコンパクトにした Riesz Pyramid により処理の高速化に成功し，実時間処理を可能にした [10]．Mohamed A. Elgharib らは，映像中の関心領域をあらかじめトラッキングすることで，映像の大きな動きの中から，関心領域内の微小な振動を強調した映像の作成に成功している [11]．

Complex Steerable Pyramid は，画像を縮尺 r，方向 θ ごとに複数の部分帯域に分割するフィルタバンクである．Fig. 1 はフィルタバンクを用いた画像分解の処理の流れを示している．本研究では Fig. 1 と同様に $r = 2$ のフィルタバンクを用いている．

入力画像に対し周波数領域における処理を行う．まず初めに高周波帯域を取り除き，残った帯域から低周波帯域を取り除く．中間の周波数帯域に対し方向フィルタをかけ，それぞれの方向に対応した部分帯域が逆変換されたものが複素画像として出力される．取り除かれた低周波帯域に対してサブサンプリングが行われ，折り返し雑音が起こらないようにさらに低周波の帯域に対して同様の処理が再帰的に繰り返される．Fig. 2 は処理を適用した画像の例である．

各縮尺 $r = 1, \ldots, n$，方向 $\theta = 1, \ldots, m$ の部分帯域より出力として返ってきた複素画像 $I_{r,\theta}$ はオイラーの公式から以下のように振幅 $A_{r,\theta}(\mathbf{x})$ と位相 $\phi_{r,\theta}(\mathbf{x})$ に分離することができる．$\mathbf{x}$

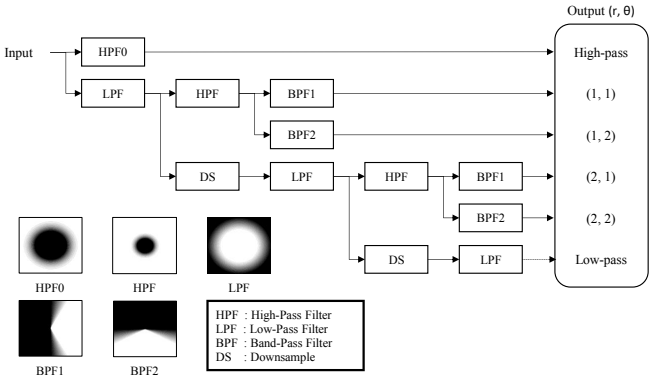

図 1　complex steerable pyramid による画像の分解の流れ $(r, \theta) = (2, 2)$

Fig. 1　Procedure of image decomposition using complex steerable pyramid $(r, \theta) = (2, 2)$

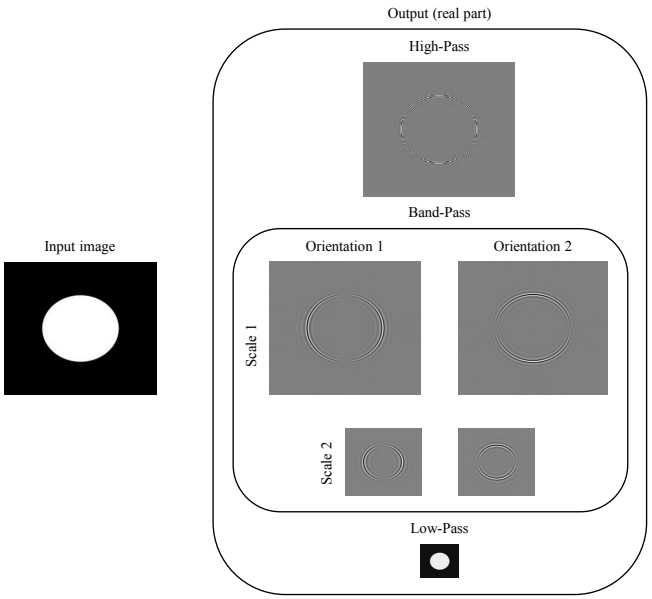

図 2　画像分解の例

Fig. 2　Example of image decomposition

は画像中の位置を表し，n，m は何番目の縮尺，方向かを表す．

$$A_{r,\theta}(\mathbf{x}) = \sqrt{Re(I_{r,\theta}(\mathbf{x}))^2 + Im(I_{r,\theta}(\mathbf{x}))^2} \tag{3}$$

$$\phi_{r,\theta}(\mathbf{x}) = \arctan \frac{Im(I_{r,\theta}(\mathbf{x}))}{Re(I_{r,\theta}(\mathbf{x}))} \tag{4}$$

2.3　従 来 手 法

Abe Davis らが提案した音源復元の従来手法について説明する．最初に，縮尺 r，方向 θ の部分帯域ごとに局所信号 $\Phi_{r,\theta}(t)$ を取り出す．まず，あるフレーム t における参照フレーム t_0 との位相差 $\phi^v_{r,\theta}(\mathbf{x}, t)$ を以下の式により求める．

$$\phi^v_{r,\theta}(\mathbf{x}, t) = \phi_{r,\theta}(\mathbf{x}, t) - \phi_{r,\theta}(\mathbf{x}, t_0) \tag{5}$$

位相差の取り出しの例を Fig. 3 に示す．この例では，白い円の 1 ピクセルのずれを取り出している．この位相差が物体の各ピクセルの移動距離を数値化したものとなる．次に各ピクセルごとに振幅 $A_{r,\theta}(\mathbf{x}, t)$ の 2 乗と位相差 $\phi^v_{r,\theta}(\mathbf{x}, t)$ を掛け合わせ，

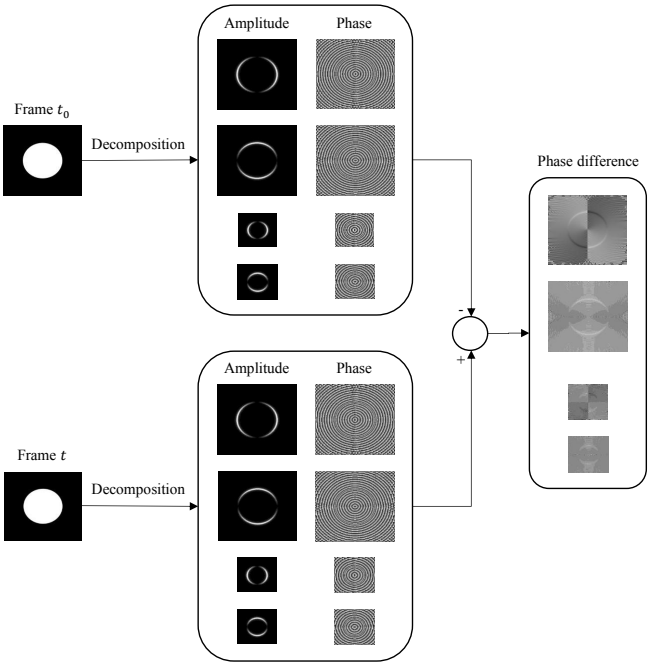

図 3　位相差抽出の流れ

Fig. 3　Procedure of extraction of phase difference

画像領域全体で総和をとったものをそのフレームにおける局所信号の出力とする．

$$\Phi_{r,\theta}(t) = \sum_{\mathbf{x}} A_{r,\theta}(\mathbf{x}, t)^2 \phi^v_{r,\theta}(\mathbf{x}, t) \tag{6}$$

この各局所信号をすべて足し合わせたときに最大になるように時間シフトさせながら足し合わせたものが最終的な復元音 $\hat{s}(t)$ となる．r_i，θ_i は，i 番目の部分帯域の縮尺，方向であり，Φ_i は i 番目の部分帯域による局所信号である．

$$t_i = \arg\max_{t_i} \Phi_0(r_0, \theta_0, t)^T \Phi_i(r_i, \theta_i, t - t_i) \tag{7}$$

$$\hat{s}(t) = \sum_i \Phi_i(r_i, \theta_i, t - t_i) \tag{8}$$

2.4　ノイズ処理

従来手法で得られる信号には低周波帯域において大きなノイズが発生してしまうので，得られた信号に対してカットオフ周波数をナイキスト周波数の 1/20 としたハイパスバタワースフィルタを適用する．また，各局所信号において特にノイズが強い場合には復元音として足し合わせる前の段階で各信号に適用する．最終的に得られた復元音に対して通常の音声信号に対するものと同じように明瞭度や正確性などの目的に応じて [12] [13] などのノイズ処理を適用する．

3.　振動モードを考慮した音源復元

3.1　振動モード

物体には振動しやすい場所とそうでない場所が存在し，その場所は周波数によって異なる．これは物体の振動が幾つかの基準振動から構成されており，基準振動ごとに振動する場所とそうでない場所が存在するためである．この基準振動が振動モー

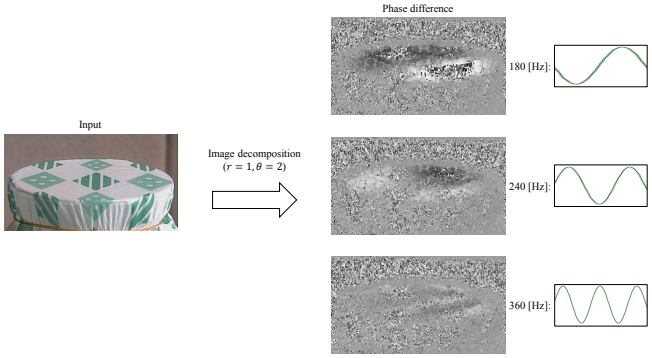

図 4　周波数ごとの位相差信号

Fig. 4　Phase difference at each frequency

ドと呼ばれる．Fig. 4 に例を示す．

Fig. 4 はコップにナイロンの蓋をしたものに特定の周波数の音を当てたときのある瞬間の様子である．$r = 1, \theta = 2$ の Complex steerable pyramid により求めた部分帯域の位相変化である．黒が $-\pi$，白が π の位相差を表し，それぞれ下方向と上方向の局所変動を表している．Fig. 4 から，周波数によって物体の振動する部分が異なることが確認できる．また，1 つの振動モードにおいても同じ瞬間に異なる方向に変化している部分が存在することが確認できる．このため，周波数ごとに異なる場所から音を復元することが音質の改善に繋がると考えられる．

3.2　周波数ごとの振動モードを考慮したフィルタ

入力映像の時刻 t のフレームの各ピクセルの変動の取り出しには従来手法と同様に complex steerable pyramid を用いる．取り出した位相変化 $\phi^v_{r,\theta}(\mathbf{x}, t)$ を時刻 t における座標 $\mathbf{x}$ での物体の移動量 $s_{r,\theta}(\mathbf{x}, t)$ とする．まず移動量の各周波数成分が最大となる区間 $n^{max}_{r,\theta}(\omega)$ を求めるため，各ピクセルの移動量を短時間フーリエ変換し，区間 n における移動量の周波数成分 $F_{r,\theta}(\mathbf{x}, \omega, n)$ を求める．

$$F_{r,\theta}(\mathbf{x}, \omega, n) = STFT[s_{r,\theta}(\mathbf{x}, t)] \tag{9}$$

$STFT[\cdot]$ は短時間フーリエ変換の計算を表す．このように求めた周波数成分と，各ピクセルの振幅の最小値 $A^{min}_{r,\theta}(\mathbf{x})$ をそれぞれ正規化し，それぞれを掛け合わせたものを部分帯域内のその成分の強さとして求める．この強さが一番大きくなる区間を $n^{max}_{r,\theta}(\omega)$ とする．

$$n^{max}_{r,\theta}(\omega) = \arg\max_{n} \sum_{\mathbf{x}} A^{min}_{r,\theta}(\mathbf{x}) F_{r,\theta}(\mathbf{x}, \omega, n) \tag{10}$$

これにより，各部分帯域の周波数ごとの物体の応答の最大値 $M_{r,\theta}(\mathbf{x}, \omega)$ が，$n^{max}_{r,\theta}(\omega)$ における振幅として求められる．

$$M_{r,\theta}(\mathbf{x}, \omega) = |F_{r,\theta}(\mathbf{x}, \omega, n^{max}_{r,\theta}(\omega))| \tag{11}$$

Fig. 5 に処理の流れを示す．

こうして得られた応答を周波数ごとの振動形態に応じたフィルタ，振幅の最小値をそれぞれのピクセルでの位相の信頼度に応じたフィルタとする．これらのフィルタをそれぞれ部分帯域に応じて正規化し，それぞれの区間における移動量に対し周波数領域において適用する．これを逆変換して得られた信号の実数部分を部分帯域内の全ピクセルに対して足し合わせたものを最終的な復元音とする．

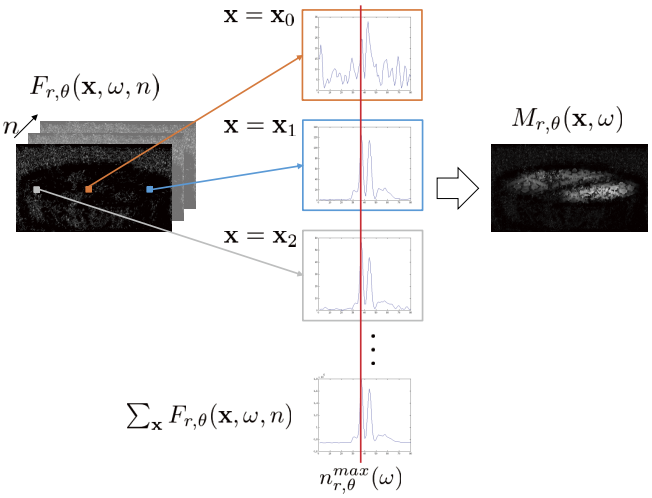

図 5　振動モードに基づくフィルタを求める処理の流れ

Fig. 5　Procedure for obtaining the vibration modes

4.　実　　験

4.1　実 験 条 件

カメラに対し垂直方向からスピーカを用いて物体に音を当てて撮影した．Fig. 6 に使用した映像のサンプルを示す．Object 1 と Object 2 は菓子の包装，Object 3 はポリ袋で，それぞれサイズの異なるプラスチックの物体である．音源には Fig. 7 のスペクトログラムが示すチャープ信号を用いた．物体とカメラは約 10 [cm]，物体とスピーカは約 30 [cm]　離れた位置で撮影された．映像は 256×256 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ワースフィルタによる 80 [Hz] 以下の信号のカットオフを行った．ただし，今回の手法では部分帯域ごとに復元を行っており，復元音は各部分帯域から復元した音のうち最もよいものとした．従来手法の特徴量の縮尺は 2 とした．

また，音源として音声を用いた場合の実験も行った．このとき，映像のフレームレートは 16000 [Hz] であり，物体 1 を撮影した．ノイズ除去についてはバタワースフィルタによる 80 [Hz] 以下の信号のカットオフの他に [13] のノイズ処理を適用した．Fig. 8 に元の音声のスペクトログラムを示す．

Object 1　　Object 2　　Object 3

図 6　使用した映像のサンプルフレーム

Fig. 6　Sample frames of videos

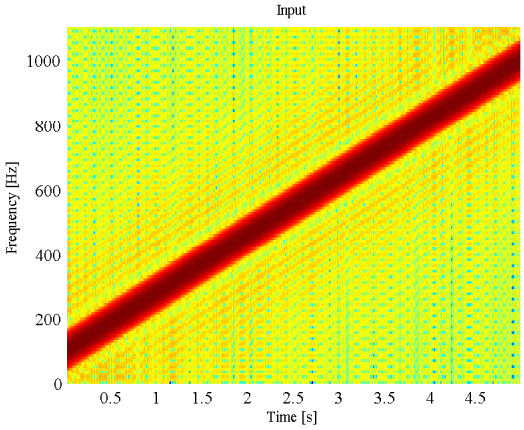

図 7　入力信号（チャープ信号）のスペクトログラム

Fig. 7　Spectrogram of input signal (chirp)

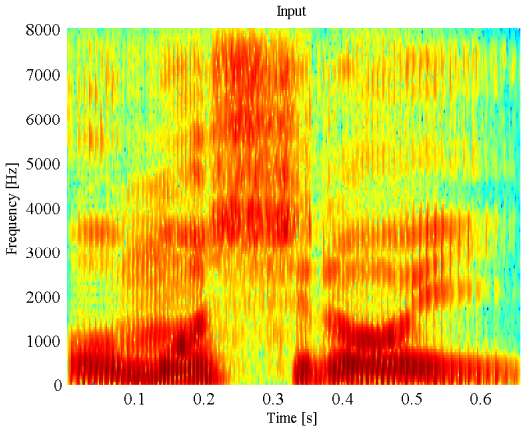

図 8　音声 /o m o sh i r o i/ のスペクトログラム

Fig. 8　Spectrogram of the utterance of /o m o sh i r o i/

表 1　チャープ信号に対する復元音の SSNR

Table 1　SSNR of recoverd sound for the chirp signal

SSNR [dB]	Object 1	Object 2	Object 3
Conventional method	0.6114	1.0259	0.8882
Proposed method	2.3520	2.1770	1.2008

4.2　実 験 結 果

それぞれの物体からの復元音のスペクトログラムを Figs. 9-11 に示す．次に従来手法，提案手法による復元音の Segmental SNR (SSNR) [14] を Table 1 に示す．SSNR は，信号を長さ N の M 個のフレームに分割し，それぞれの SNR を平均することで求められる．

$$SSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1}(y(n))^2}{\sum_{n=Nm}^{Nm+N-1}(s(n)-y(n))^2}. \quad (12)$$

$y(t)$ はもとの信号，$s(t)$ は処理を施した信号を表す．

また，音声に対する実験結果を示す．復元音のスペクトログラムを Fig. 12 に示す．従来手法，提案手法による復元音の SSNR と明瞭度の指標である short-time objective intelligibility (STOI) [15] を Table 2 に示す．

スペクトログラムおよび SSNR の値から，提案手法によりノイズが軽減されたことが確認できる．音声の場合でも音質を

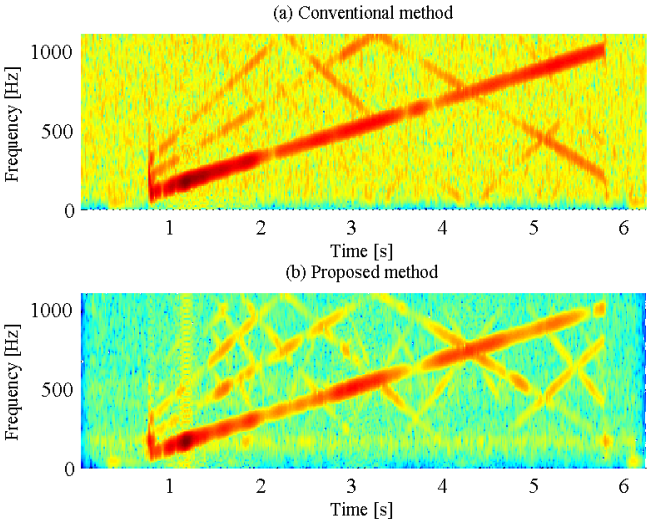

図 9　物体 1 からの復元音のスペクトログラム

Fig. 9　Spectrogram of sound recovered from object 1

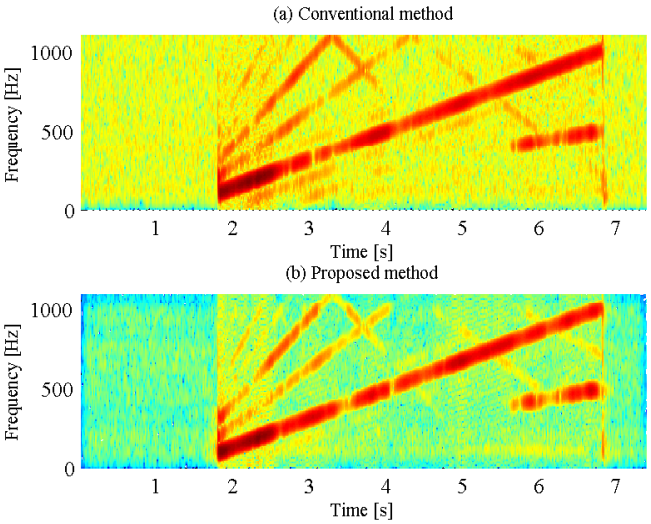

図 10　物体 2 からの復元音のスペクトログラム

Fig. 10　Spectrogram of sound recovered from object 2

表 2　復元音声 /o m o sh i r o i/ の評価

Table 2　Evaluation of recoverd sound for the utterance of /o m o sh i r o i/

	SSNR [dB]	STOI
Conventional method	-0.5806	0.5989
Proposed method	-0.4093	0.6227

改善できている．しかし，音が復元できていない部分が幾つか存在している．振動モードをうまく取り出せている周波数の音はうまく復元できているが，そうでない部分では復元することができなくなってしまっていると考えられる．物体の振動形態の取り出し方法の工夫によりさらに改善の余地があると考えられる．

提案手法では従来手法と同様に，実際には存在しない倍音を復元してしまっている．この現象は正弦波などの単純な音源の復元の際によく見られた．このことは物体の特性との関係も含めて調査する必要がある．

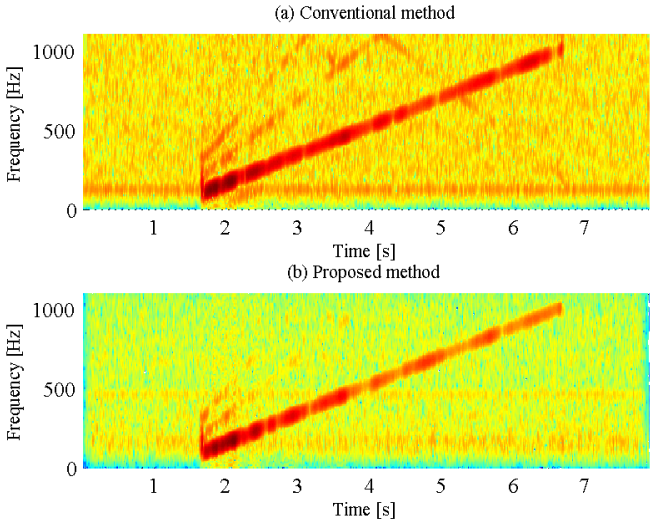

図 11　物体 3 からの復元音のスペクトログラム

Fig. 11　Spectrogram of sound recovered from object 3

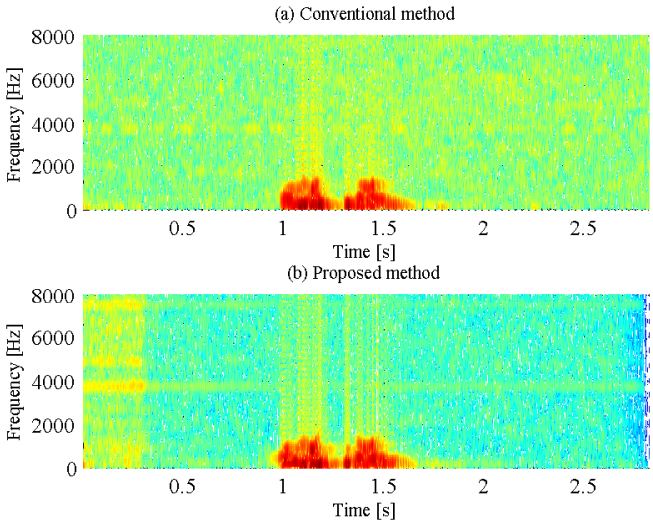

図 12　復元音声のスペクトログラム /o m o sh i r o i/

Fig. 12　Spectrogram of recovered utterance of /o m o sh i r o i/

5. おわりに

本研究では，物体の微小な振動をハイスピード映像から抽出する際，物体の振動モードに着目し，特に振動している部分を重みづけして抽出する手法を提案した．提案手法と従来手法による音源復元を行い比較することで音質が向上することが確認できた．今後は振動モードの特定方法の改善について考える．

文　献

[1] Abe Davis *et al.*, “Visual Vibrometry: Estimating Material Properties from Small Motion in Video,” IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),2015.

[2] Abe Davis *et al.*, “Image-space modal bases for plausible manipulation of objects in video,” ACM Transactions on Graphics,34(6),239:1-239:7,2015.

[3] Justin G. Chen *et al.*, “Video camera-based vibration measurement for Condition Assessment of Civil Infrastructure,” NDT-CE International Symposium Non-Destructive Testing in Civil Engineering,15-17,2015.

[4] Rothberg, S., Baker, J., and Halliwell, N. A., “Laser vibrometry: pseudo-vibrations,” Journal of Sound and Vibration, 135 (3), 516-522, 1989.

[5] Zeev Zalevsky *et al.*, “Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern,” Optics Express, 17(24), 21566-21580, 2009.

[6] Abe Davis *et al.*, “The Visual Microphone: Passive Recovery of Sound from Video,” ACM Transactions on Graphics, 33 (4), 79:1-79:10, 2014.

[7] Yusuke Yasumi *et al.*, “Visual Sound Recovery Using Momentary Phase Variations,” The 23rd International Workshop on Frontiers of Computer Vision, 2-6, 2017.

[8] Javier Portilla *et al.*, “A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients,” International Journal of Computer Vision, 40 (1), 49-71, 2000.

[9] Neal Wadhwa *et al.*, “Phase-Based Video Motion Processing,” ACM Transactions on Graphics, 32(4), 80:1-80:10, 2013.

[10] Neal Wadhwa*et al.*, “Riesz Pyramids for Fast Phase-Based Video Magnification,” IEEE International Conference on Computational Photography (ICCP), 1-10, 2014.

[11] Elgharib, M.A., Hefeeda, M., Durand, F., and Freeman, W.T., “Video Magnification in Presence of Large Motions,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[12] Steven F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Transactions on Acoustics, Speech and Signal Processing, 27(2), 113-120, 1979.

[13] Lu, Yang and Loizou, P. C., “A geometric approach to spectral subtraction,” Speech Communication, 50(6), 453-466, 2008.

[14] John H. L. Hansen and Bryan L. Pellom, “An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms,” Proceedings of the International Conference on Speech and Language Processing, 2819-2822, 1998.

[15] Taal, Cees H., Hendriks, Richard C., Heusdens, R., and Jensen, J., “An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech,” IEEE Transactions on Audio, Speech and Language Processing, 19(7), 2125-2136, 2011.