ハイスピード映像中の物体振動を利用した visual microphone の検討 *

☆安見祐亮, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

2014 年にハイスピード映像中の物体振動から音を復元することができるということが発表された [1]. これは，それまでに発表されてきた遠距離から物体の振動を取るための手法と異なり，レーザー光などを当てたりせずに物体の振動を取ることができる．物体を撮影するための光以外に必要なものがな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で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．本稿では，この技術について紹介するとともによりよい音をとるための振動の取り出し方を検討する．

2 音の復元について

2.1 振動抽出の仕組み

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].

$$f(\mathbf{x}) = \sum_{\omega=-\infty}^{\infty} A_w e^{iw\mathbf{x}} \tag{1}$$

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].

$$f(\mathbf{x}+\delta(t)) = \sum_{\omega=-\infty}^{\infty} A_w e^{iw(\mathbf{x}+\delta(t))} \tag{2}$$

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.

2.2 Visual Microphone について

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，位相差を取り出している．これにより，画像はスケール r，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．このとき，各局所ウェーブレットはコサイン，サインの両成分を保持しているので，局座標変換により以下のように表現される.

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

位相差を取り出すために，フレーム t の各局所ウェーブレットごとに参照フレーム t_0 における位相との差 $\phi_v(r,\theta,\mathbf{x},t)$ を出す.

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

振動の大きさは，各局所ウェーブレットごとに振幅の二乗と位相差の要素積の和 $\Phi(r,\theta,t)$ として表し，すべての局所ウェーブレットに対し最も大きくなるようにこれを足し合わせたものとなる.

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

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] やスピーチエンハンスメント [5] による処理が行われている.

2.3 振動抽出方法と音復元

振幅ごとの位相の変化をすべて足し合わせることによって音を取り出しているが，物体の移動量をより正確に表現することでよりよく音が取り出せる．振幅ごとに位相の変化が取り出されるが，振幅はテクスチャ強度により与えられるので振幅が小さい部分の位相の変化はノイズによるものが大きい.

高精度に物体の移動を取り出すために，各局所ウェーブレットにおける位相の変化だけを見てみることを考える．このとき，各局所ウェーブレットから得られた，振幅が一定以下の部分にマスクをかけた位相差の平均を出力とする．これにより，正弦波の移動量を直接位相差から取り出す.

また，よりはっきり変化を出すために単純なエッジの移動量の和を取り出した場合を考える．このとき，画像の x,y 方向の 1 ピクセルの差分画像 $I_x(\mathbf{x}), I_y(\mathbf{x})$ をとり，これらから求めた振幅 $A(\mathbf{x})$，および位相 $P(\mathbf{x})$ を音を取り出すのに用いる.

$$A(\mathbf{x}) = \sqrt{I_x(\mathbf{x})^2 + I_y(\mathbf{x})^2} \tag{6}$$

$$P(\mathbf{x}) = \tan^{-1}\left(\frac{I_y(\mathbf{x})}{I_x(\mathbf{x})}\right) \tag{7}$$

物体の大きさに対して振動による移動が小さい，移動により物体の位置が完全には変わらない場合には，この方法でも物体の移動量を数値化できる.

---

* Visual microphone using a vibrating object in high speed video. by Yûsuke Yasumi, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

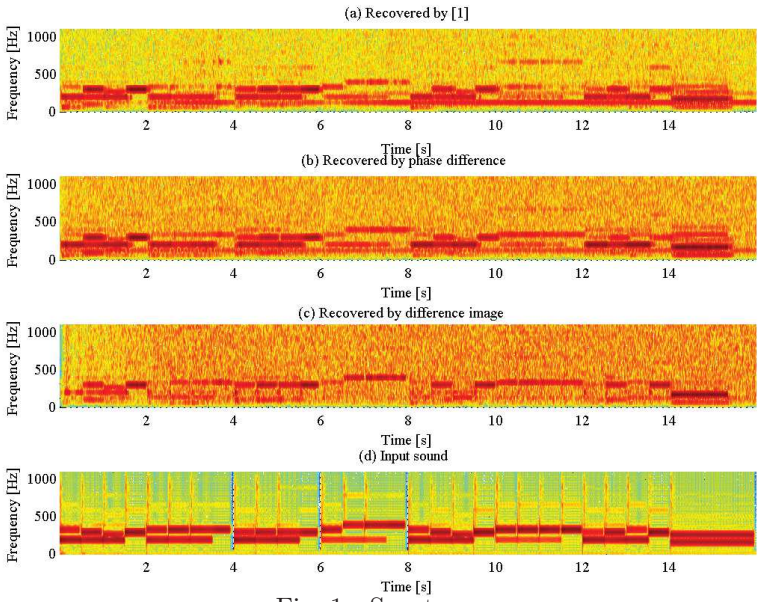

Fig. 1　Spectrogram

Table 1　SSNR of recoverd sound

	SSNR [dB]
Abe Davis *et al.*[1]	1.4851
Phase difference	1.5847
Difference image	-1.3465

3　実験

3.1　実験条件

使用した映像は，スーパーのレジ袋を拡大した映像で，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暗室にて撮影した．映像は 256x256 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 55 [Hz] 以下の信号のカットオフのみを行った．特徴量のスケールは 2 で，オリエンテーションは 2 である．

3.2　実験結果

Fig. 1 に入力音，各手法による復元音の短時間フーリエ変換による周波数スペクトラムを示す．また，これらの復元音の Segmental SNR (SSNR)[6] は Table 1 のようになった．SSNR の結果から，位相のみに着目した場合と文献 [1] の場合と比べてエッジの移動量を用いたほうがノイズが大きくなっていることがわかる．

4　おわりに

今回は，Visual Microphone の改善に関する検討とともに，振動の抽出方法に関する実験を行った．今後はより実際の環境に近づけたときに音をとる方法を考え，まずは移動する物体から振動を取り出すことについて考える．

参考文献

[1] Abe Davis *et al.*, The Visual Microphone: Passive Recovery of sound from Video, ACM Transactions on Graphics, 33(4), 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), 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, 113-120, 1979.

[5] Philipos 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, 857-869, 2005.

[6] John H. L. Hansen *et al.*, An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms, IN PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON SPEECH AND LANGUAGE PROCESSING, 2819-2822, 1998.