Convolutional Neural Networksによる物体の微小振動からの音声復元*

☆布施陽平, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

音声は空間を媒体の圧力の変動として伝わっており，物体に衝突するとその表面に微小な振動を起こす．音声に由来する物体の振動は遠距離で発生した音声の取得に用いられ，監視や安全保障などの分野での応用が期待される．例えば，レーザー光を対象の物体に照射し，反射光のドップラー効果を利用して物体の振動を測定するレーザードップラー振動計 [1] を用いて音声を復元するレーザーマイクロフォンが提案されている．また，ハイスピードカメラで物体の振動を捉え，その映像に処理を加えることで音声を復元する手法 [2] も提案されている．この手法はレーザーマイクロフォンと比較すると，レーザー光やセンサなどの追加装置を必要とせず，対象の物体の反射性質などに関する制約も少ないといった利点がある．また，文献 [3] ではハイスピードカメラによる物体の振動映像からその物体の材質を推定する研究も行っており，物体振動の新たな分野への応用の可能性を示している．

映像から別のメディアへ変換する研究として，唇画像からテキスト情報を推定する Lip Reading [4] や，物体を棒で叩く映像などから音声を推定する Visually Indicated Sounds [5] などがある．これらはニューラルネットワークを用いた変換を行っており，メディア変換におけるニューラルネットワークの有用性を示したといえる．

本研究は，物体の微小振動を捉えた映像から特徴量を抽出する際に，Convolutional Neural Networks (CNN) による学習を行うことでより精度の高い音声復元を行うことを目的としている．

2 音声復元

2.1 従来手法

文献 [2] では，complex steerable pyramid [6] による特徴量抽出を行っている．これにより映像の各フレームはスケール r，オリエンテーション θ ごとにウェーブレット変換される．各ウェーブレットは画像中の座標 $\mathbf{x}$ に対し，振幅 $A(r,\theta,\mathbf{x})$ と位相 $\phi(r,\theta,\mathbf{x})$ を用いて

$$A(r,\theta,\mathbf{x})e^{i\phi(r,\theta,\mathbf{x})} \tag{1}$$

と表される．時刻 t のフレームの各ウェーブレットごとに初期時刻 t_0 のフレームとの位相差 $\phi_v(r,\theta,\mathbf{x},t)$ を求める．

$$\phi_v(r,\theta,\mathbf{x},t) = \phi(r,\theta,\mathbf{x},t) - \phi(r,\theta,\mathbf{x},t_0) \tag{2}$$

各ウェーブレットごとの移動量は振幅の 2 乗と位相差の積で表され，その総和を全体の移動量 $\Phi(r,\theta,t)$ とする．

$$\Phi(r,\theta,t) = \sum_{\mathbf{x}} A(r,\theta,\mathbf{x},t)^2 \phi(r,\theta,\mathbf{x},t) \tag{3}$$

求められた移動量を時間方向に並べたものを復元音声とする．また，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，Spectral Subtraction [7] や音声強調 [8] が行われている．

文献 [9] では，風による揺れやカメラ自身の振動などの外的要因による大きな変動が生じた場合に物体の微小振動を捉える手法が提案されている．この手法では，時刻 t のフレームに対し，直前の時刻 $t-1$ のフレームとの位相差をとる．

$$\phi_v(r,\theta,\mathbf{x},t) = \phi(r,\theta,\mathbf{x},t) - \phi(r,\theta,\mathbf{x},t-1) \tag{4}$$

時刻 t における全体の移動量を過去の移動量の総和とすることで，大きな変動の影響を減らしている．

$$\Phi(r,\theta,t) = \Phi(r,\theta,t) + \Phi(r,\theta,t-1) \tag{5}$$

2.2 提案手法

物体の微小振動から音声へ変換するニューラルネットワークモデルを Fig. 1 に示す．CNN を 2 層重ね，最後に全結合層を重ねた形になっている．入力は 32 × 32 pixel の大きさのグレースケール映像とし，音声の FFT パワースペクトルを教師信号とする．

前処理として，映像の 256 フレームを 1 ブロックとしてシフト幅 128 で切り出す．その 1 ブロッ

*Recovering sound from small vibration of an object based on Convolutional Neural Networks by Yohei FUSE, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

ク $X_n = [X_{n,1}, X_{n,2}, ..., X_{n,256}]$ を 1 回の入力とする．学習時はブロックに対応する音声の 256 サンプル点の 129 次元 FFT パワースペクトル $Y_n = [y_{n,1}, y_{n,2}, ..., y_{n,129}]$ を用いる．各データに対し，入力は各ピクセルの時間次元の平均を 0 にし，時間次元の分散が画像領域内で最大のもので割るという正規化を行った．教師信号はスペクトログラム全体で平均 0，分散 1 の標準化を行った．

$$\tilde{X}_{ijt} = \frac{X_{ijt} - \bar{X}_{ij}}{\max_{i,j} \sigma_{X_{ij}}} \quad (6)$$

$$\tilde{Y}_{nf} = \frac{Y_{nf} - \bar{Y}}{\sigma_Y} \quad (7)$$

$\bar{X}_{ij}$，$\sigma_{X_{ij}}$ は X の各ピクセルの時間次元に対する平均，分散である．$\bar{Y}$，σ_Y は Y の平均，分散である．i，j は画像中の座標，t は時間，n はブロックの番号，f は周波数を表す．

畳み込み層の 1 層目の入力チャネルを映像フレームとすることで時間と空間のデータを同時に畳み込んでいる．注目するピクセルに対し，その周辺のピクセルの値も同時に処理することで，1 つのピクセルの変化のみを見るときよりも良い特徴量が得られることを期待している．さらに畳み込み層を重ね特徴量を圧縮することにより，ノイズにロバストになると予想される．2 層の畳み込み層を通した後，画像領域全体の平均をとり特徴量ベクトルとする．特徴量ベクトルを全結合層に通したものをモデルの出力とする．畳み込み層のチャネル数は第 1，2 層目についてそれぞれ 512, 128 とした．畳み込みフィルタの大きさは全て 3 × 3 とした．畳み込み層の活性化関数は 1 層目に絶対値，2 層目に Relu を用いた．全結合層の出力は線形活性とした．

今回のモデルでは映像中の振動そのものに注目する必要がある．正規化により入力は正負の値をとり，ネットワーク内部でも正負の値をとると考えられるが，特徴量の画像全体の平均をとる際に局所信号同士で打ち消し合ってしまう可能性が考えられる．そこで，ネットワーク内部で絶対値をとる処理を加えることで各局所信号の値を揃えられることを期待した．

映像データに対し CNN を用いて特徴量を抽出することは [10] など行動認識の分野でも行われており，今回のモデル構築の参考にした．

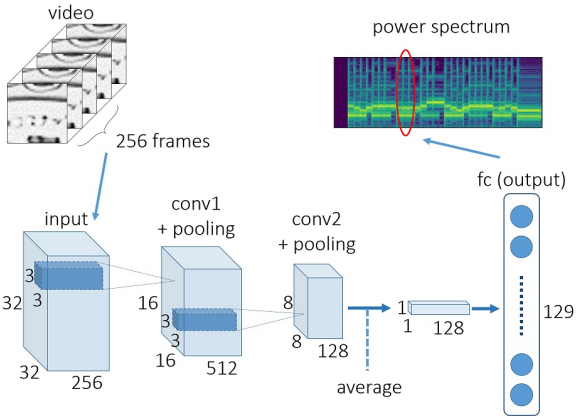

Fig. 1　物体の微小振動を音声のパワースペクトルに変換するニューラルネットワークモデル

3　実験

3.1　実験条件

cifar10 データセット中の 200 種類の画像からシミュレーション映像を作成し学習データセットとした (Fig.6)．cifar10 は，Alex Krizhevsky らにより整備されたラベル付き画像データセット [11] であり，一般物体認識タスクに用いられる．

教師信号には，映像作成に用いた sin 波音源から生成したパワースペクトルを使用した．振動の周波数は十二平均律の A3 (220 Hz) から A5 (880Hz) の 25 種類，振動方向は上下方向と左右方向の 2 種類を用意した．クローズドテストおよび，ハイスピードカメラで撮影した菓子の包装の映像 (Fig. 7) を用いたテストを行った．得られた出力パワースペクトルから Griffin/Lim 法 [12] により波形を復元する．

実験に使用した映像と音声のサンプリング周波数はどちらも 2200 Hz である．

シミュレーション映像は，元となる画像の各ピクセルを音源の各サンプル点の振幅に合わせて移動させることで作成した．1 ピクセル未満の変動後の各ピクセルの値は周辺ピクセルの値の加重平均をとることで計算した．今回は最大振幅を 0.25 にランダムな誤差 (± 10%) を加えたものとしたが，入力には時間次元に対する正規化を行うため，映像間の振幅の差の影響はなくなる．

3.2　実験結果

Fig. 2 にクローズドテストの結果を示す．上段左が元の音源，右が従来手法による復元音声のスペクトル，中段，下段は提案手法により推定された結果である．image1，2 はそれぞれ異なる画

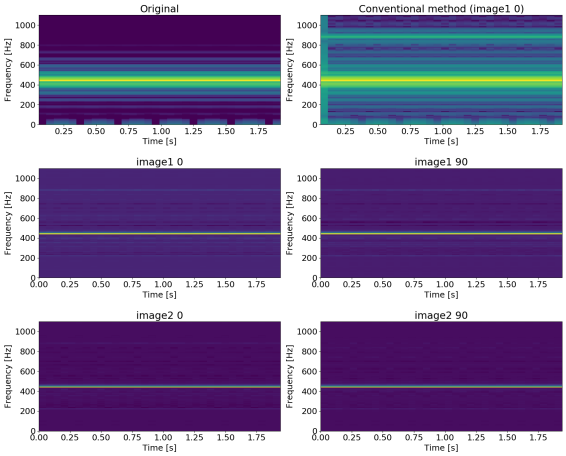

Fig. 2　sin 波 440 Hz の復元結果

像から作成した映像に対する出力である．0 および 90 はそれぞれ上下方向，左右方向の振動であることを表している．Fig. 2 に示した例以外でも異なる画像，異なる周波数に対して正解に近い出力が得られた．さらに多くの種類の画像からなる学習データセットを用いることで，未知の画像に対する応答はよくなることが予想される．

画像 200 種類，周波数 25 種類，振動方向 2 種類で合計 10,000 個の映像に対する出力と，それぞれの正解の音声のスペクトルを標準化したもののユークリッド距離を計算し，画像ごとに平均をとったものを Fig. 3 に示す．また，周波数ごとに平均をとったものを Fig. 4 に示す．画像ごとの距離を見ると，極端に大きな距離をとる出力が複数見られ, 現在のモデルでは音を復元しづらい画像が存在することが分かる．本研究では，未知の画像を含むあらゆる画像の振動映像に対して音を復元することを目的とするため，学習データセットを見直す必要がある．

Fig. 5 にハイスピード映像を用いたテストの結果を示す．上から，音源のスペクトル，visual microphone[2] から得られた音のスペクトル，提案手法の出力である．音源として，sin 波による楽曲を用いた．従来法はスケール 2，オリエンテーション 2 として復元した音声のスペクトルである．

提案手法ではかなりノイズが大きいが，わずかではあるが，未知の画像に対しても時間特徴量が捉えられていることが確認できた．シミュレーション映像を用いた学習により，実際にハイスピードカメラで撮影された映像の振動が捉えられたことから，シミュレーション映像の学習

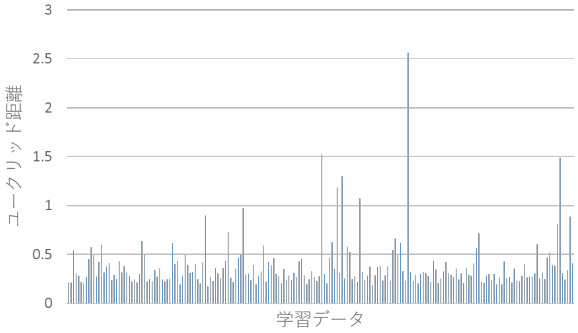

Fig. 3　音源のスペクトルと提案手法の出力とのユークリッド距離 (画像)

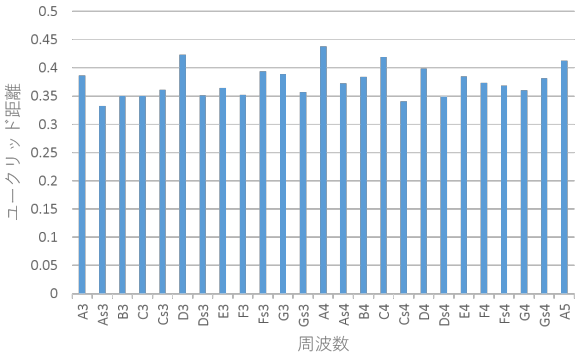

Fig. 4　音源のスペクトルと提案手法の出力とのユークリッド距離 (周波数)

データとしての有用性が示された．

学習データには入っていない低周波帯の振動に対してはほとんど反応がない．この問題は学習データの振動周波数の種類を増やすことで解決すると考えられる．

4　おわりに

今回は，物体の振動を CNN により学習し音声を復元する手法を提案した．シミュレーション映像を用いた学習により実際のハイスピードカメラの映像から音声を復元することができたが，現状ではまだノイズが大きい．しかし，画像の違いによらず振動の特徴量が得られることが確認できたため，パラメータ調整次第でより良い出力が得られるようになると考えられる．また，今回は事前学習をせずに実験を行ったため，重みの初期値に関する検討も必要である．

今後は，CNN の総数やフィルタ数の検討，LSTM (Long Short Term Memory) など別のモデルを用いた学習や，学習データセットのサイズや種類の検討などを行い，音声復元の精度を高

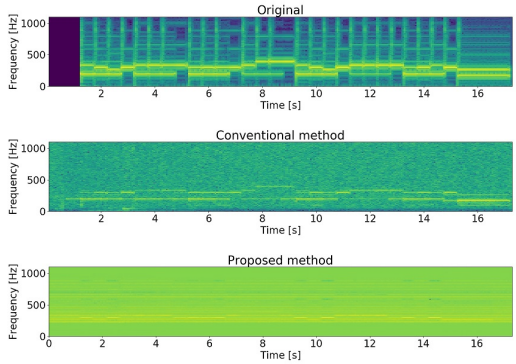

Fig. 5　ハイスピード映像からの音の復元結果

める手法を考える．

参考文献

[1] Rothberg, S., Baker, J., and Halliwell, N. A. “Laser vibrometry: pseudo-vibrations,” Journal of Sound and Vibration 135 (3), 516522, 1989.

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

[3] Abe Davis *et al.*, “Visual Vibrometry: Estimating Material Properties from Small Motions in Video,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (4), 732-745, 2017.

[4] J. S. Chung *et al.*, “Lip Reading Sentences in the Wild,” IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[5] Andrew Owens *et al.*, “Visually Indicated Sounds,” Computer Vision and Pattern Recognition (CVPR), 2016.

[6] 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.

[7] 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.

[8] 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, 13 (5), 857-869, 2005.

[9] Yusuke Yasumi *et al.*, “Visual Sound Recovery Using Momentary Phase Variations ”, The 23rd International Workshop on Frontiers of Computer Vision, P2-6, 2017.

[10] Simonyan, K. and Zisserman, A. “Two-stream convolutional networks for action recognition in videos,” In Advances in Neural Information Processing Systems, 2014a.

[11] A. Krizhevsky. “Learning multiple layers of features from tiny images,” Master ’ s thesis, Department of Computer Science, University of Toronto, 2009.

[12] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” IEEE Transactions on Acoustics, Speech, and Signal Processing, 32 (2), 236-243, 1984.

Fig. 6　学習データセットの例

Fig. 7　ハイスピードカメラで撮影した菓子のプラスチック包装の一部 (テストデータ)