歩行動作のセンシングデータを入力とした足音合成法

吉田 赳[1,a)]　西村 竜一[1,b)]　入野 俊夫[1,c)]

概要：本研究では、人間の歩行動作のセンシングデータを入力し、足音の音響信号を合成するシステムを提案する。近年、ゲームコンテンツのグラフィックスにおいてフォトリアルな表現が可能になり、効果音においても多様性やリアルさが求められている。足音の自動合成手法について、既存研究では、リアリティの向上に必要な歩行動作の時系列の依存関係の考慮がなされていない。そこで、本研究では、センサデータの複数のフレームを連結して、変換モデルを学習することで、足音が発生する直前の歩行動作を考慮した足音合成手法を提案する。学習データとして、時間同期した歩行動作のセンシングデータと足音を収集した。収集した 20,971 秒分のデータを用いて足音波形の合成実験を行なった。その結果、DNN(ディープニューラルネットワーク) の出力を、抽象化し、次元数を減らした音響特徴量とした際に、元の波形と近い波形を合成できることを確認した。

Acoustic Signal Synthesis Method of Footsteps Based on Sensing Data of Walking Motion

Yoshida Isami[1,a)]　Nisimura Ryuichi[1,b)]　Irino Toshio[1,c)]

Abstract: This study proposes a method to synthesize acoustic signals of footsteps based on human walking motion sensing data. In recent years, photoreal expression has become possible in the graphics of game contents. We also require diversity and realism for sound effects. Regarding the automatic synthesizing method of footsteps, the time series dependence of walking motion necessary for improving reality is not considered in the previous study. Therefore, we propose a footstep synthesis method considering the state transition of the walking motion. It is realized by coupling multiple seinsing data frames as inputs in learning the DNNs (Deep Neural Networks). As for learning data, we collected sensing data of walking motion and acoustic signals of footsteps synchronized in time. In the experiments, we synthesized acousitc signals of the footsteps using data of 20,971 seconds we have collected. As a result, we have confirmed that acousitc waveforms close to the original can be synthesized when the abstracted acoustic features are adopted.

1. はじめに

本研究では、人間の歩行動作センシングデータから音響信号である足音を合成する手法を提案する。

近年のコンピュータグラフィックス技術の発展により、ゲームコンテンツ等においてフォトリアルな表現が利用されている。それに応じて、サウンドにおいても多様なキャラクターの動作や環境に対応した効果音を作成する必要がでてきた。しかし、多くの効果音は録音（フォーリーサウンド）やシンセサイザを用いてサウンドデザイナが手作業で作成しており、大きなコストを要している。

この問題を解決するために、近年プロシージャルオーディオと呼ばれる効果音自動生成の研究が進められている [1]。個体の衝突で生成される音のシミュレーション [2] や液体の音のシミュレーション [3] など、物理モデリングによる効果音合成の研究が存在する。特に足音は幅広いゲーム・映画コンテンツで利用されており、人間の歩行動作から足音を生成する物理モデル [4] と、そのインタフェースの設計方法 [5] の報告がある。しかし、その中でのモデルは足が接地した際の床からの反力（床反力）のみを考慮し

[1] 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University, 930, Sakaedani, Wakayama, Wakayama 640-8510, Japan
[a)] s165061@center.wakayama-u.ac.jp
[b)] nisimura@sys.wakayama-u.ac.jp
[c)] irino@sys.wakayama-u.ac.jp

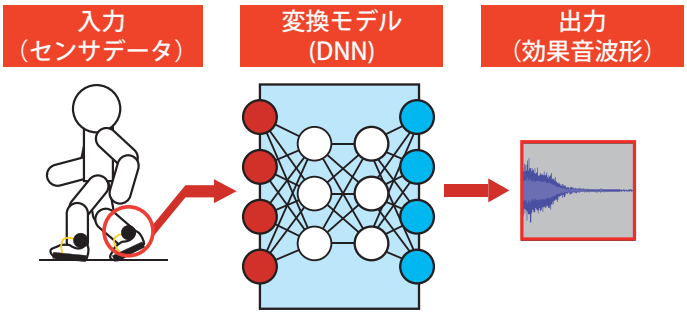

図 1　提案システムの概要図
Fig. 1　Outline of the proposal system

ており、足音が生成される前後の動作遷移は考慮していなかった。そこで本研究では、足の動作遷移を考慮した足音合成手法を提案する。具体的には、取得したセンサデータを複数フレーム結合し（本稿では、フレーム連結処理と呼ぶ)、システムの入力とする。さらに、少量のデータでも変換モデルを学習できるように、抽象化した低次元の音響特徴量を出力におけるパラメータとすることを検討した。

本研究では、人間の歩行動作情報を入力とした変換モデルにより、足音を自動生成することを目的とする。将来的には、例えばモーションキャプチャしたデータから、効果音を自動生成できるようなシステムへの応用が考えられる。

2.　提案手法の概要

図 1 に提案手法の概要を示す。本節では、足音合成システムの入力（センシングデータ）と出力（効果音）及びその間を変換を担う変換モデルについて説明する。

2.1　変換モデルの入力と出力

変換モデルの入力となる、人間の歩行動作のセンシングデータは、9 軸センサと 6 つの圧力センサ両足分である（詳細は 2.1.1 節)。このデータをフレーム処理・連結した、サンプリング周波数 60Hz、長さ 1,000ms 分のデータ (1,800 次元) を入力とする。フレーム処理・連結に関しては、2.3 節にて説明する。

変換モデルの出力は、サンプリング周波数 8kHz、長さ 200ms 分の足音の音響信号を表現する音響特徴量とする。音響特徴量に関しては 4.1 節に記述する。出力された音響特徴量から合成した波形を、前フレームの波形とつなぎ合わせることで、連続した足音を出力する。

2.1.1　センサ

図 2 に、足にセンサを装着している様子を示す。足とスリッパの間に中敷きを敷き、その裏に圧力センサ 6 つを配置した。圧力センサの配置 (図 3) は、歩行動作の床反力を圧力センサで計測した既存研究 [6] を参考にした。くるぶし付近には Arduino を取り付け、圧力センサと 9 軸センサが接続されている。Arduino は Macbook と USB でシリアル通信接続しており、Macbook 上の Puredata で実装した

図 2　足に装着するセンサ
Fig. 2　Close up foot with sensors

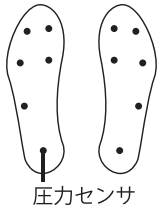

図 3　圧力センサの配置
Fig. 3　Location of pressure sensors

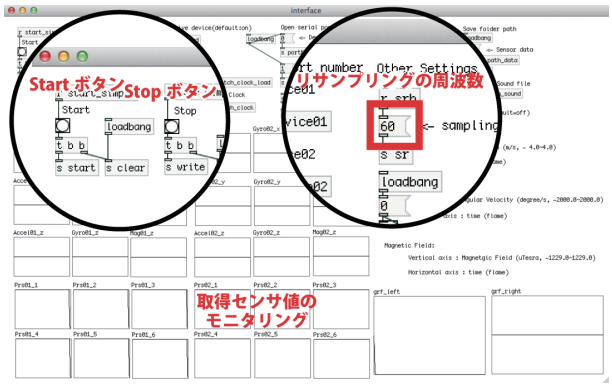

図 4　Puredata による学習データ収集ソフト
Fig. 4　Training data collection software by Puredata

データ収集ソフト（図 4）を用いて、すべてのセンサ出力値を両足分（計 30 軸）同時に記録する。センサに寄ってサンプリング周波数は異なるため、60Hz でリサンプリングする。なお、ここで用いた機材は以下の通りである。

- Arduino：Pololu 社 A-Star 32U4 Micro
- Macbook：Apple 社 Macbook Air (13-inch, Late 2010)
- 圧力センサ：Interlink Electronics 社 FSP400
- 9 軸センサ：InvenSense 社 MPU-9250
  - 3 軸加速度センサ
  - 3 軸ジャイロセンサ
  - 3 軸方位センサ

2.2　変換モデル

本研究では、センシングデータを足音に変換するアルゴリズムとして、ディープニューラルネットワーク（DNN）を採用した。DNN は近年、音声合成の分野で高い性能を示しており [7] [8]、最近では動画を入力とした効果音合成法 [9] に利用されている。本研究における DNN は、1,800 次元のセンシングデータを入力とし、効果音の音響特徴量を出力する変換器である。

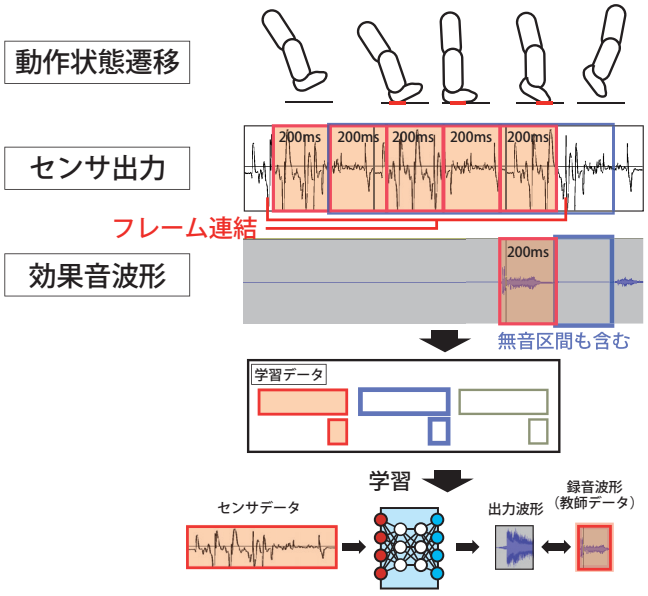

図 5　フレーム処理・連結処理フロー
Fig. 5　Software flow of frame processing and coupling multiple seinsing data

2.3　センシングデータのフレーム連結

本研究では、歩行動作の状態遷移を表現する方法として、複数フレームを結合したセンシングデータを学習データとする手法（フレーム連結処理）を提案する。

足音の音響信号を合成する際には、歩行動作の状態遷移を考慮すべきである。例えば、足音が実際に鳴るタイミングは地面に足が接着した瞬間であるが、その音に影響を与えるのは、着地直前の加速度等になる。しかし、足音はその時間長や連続する足音の時間間隔が変動するため、状態遷移を表現するには、HMM(Hidden Markov Model、出力確率の状態遷移を表現でき、入力信号の時間伸縮に柔軟に対応できる確率モデル) 等のアルゴリズムを組み込む必要があった。HMM を構築するには、学習データに対するラベリング作業が必要となるが、その作業は大きなコストを要する。そこで、ラベリングが不要となるように、フレーム連結処理（図 5）を以下のようにした。

(1) 時間同期されたセンシングデータ・足音波形を共にフレーム処理する。ここで、フレーム長とフレームシフトは同様に 200ms である。
(2) その後、現時刻のフレームに加え、過去 4 フレームを結合する。これは、足音生成に強く影響を与えているのは、足音の含まれるフレームよりも過去のフレームの動作であると考えたためである。
(3) 結果的に、足音波形は 200ms、センシングデータは 1,000ms 分が一対の学習データとなる。なお、足音の含まれない無音部分も同様にフレーム結合して学習データに含める事で、足音の間隔の変動を考慮することができる。

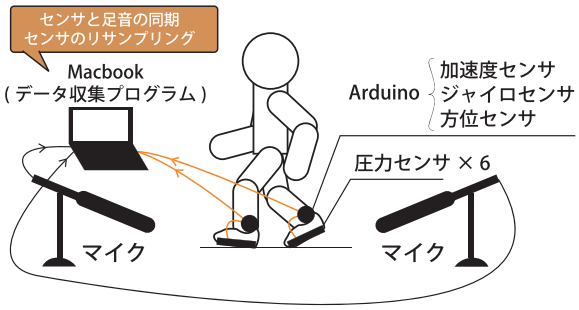

図 6　歩行データ収集システム概略図
Fig. 6　Outline of the walking data collection system

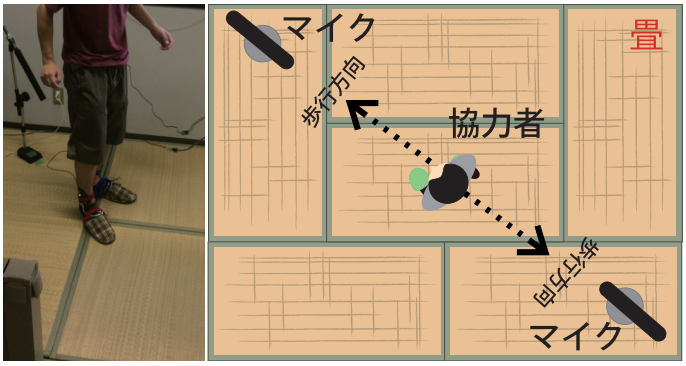

図 7　データ収集場所とマイクの配置
Fig. 7　Location of microphone and place of collecting data

3.　歩行に伴うセンシングデータと音響信号の収録

本研究では、変換モデルの学習データとして、歩行に伴うセンシングデータと、時間同期された音響信号（以下、歩行データ）を収集する必要があった。

3.1　収録システム

図 6 に歩行データ収録に用いたシステムの概略図を示す。人間の動作のセンシングデータの記録方法は、2.1 節で述べた。この本データ収集においては、センシングデータに加えて足音を録音した。図 7 のように配置したマイクを、オーディオインタフェースを通して Macbook と接続した。Puredata で実装したデータ収集ソフト（図 4）によってマイクの入力信号をサンプリング周波数 48kHz、量子化ビット数 16bit で録音した。ここで、図 4 に示したスタートボタンを押すことでセンシングデータと足音波形の記録を同時に開始しており、全センシングデータと録音波形はソフト上で時間的に同期する事を確認している。

3.2　収録の流れ

下記の流れで協力者の歩行データを収集した。なお、協力者には図 8 に示すように、10 種類の歩行を依頼した。

(1) 協力者の体重・身長を計測
(2) 協力者に 2.1 節で示したセンサを装着

1. 前向きに歩いた後、後ろ向きに歩く
2. 前向きに歩いた後、振り返って歩く
3. 旋回しながら歩く
4. 横向きに歩きながら左右に往復する
5. しゃがんだ体制で歩行する
6. 静かに（人に気づかれないように）歩行する
7. 前向きにジョギングした後、振り返ってジョギング
8. 旋回しながらジョギングする
9. 勢い良くジャンプして、両足で着地する
10. 小刻みにジャンプして、両足や片足で着地する

図 8　10 種類の歩行方法
Fig. 8　10 kinds of walking action

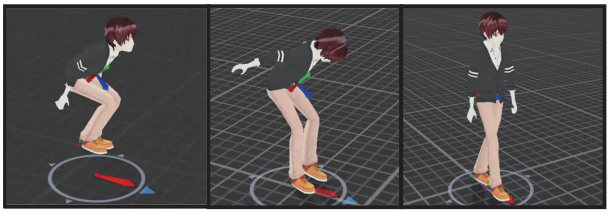

図 9　協力者に提示した動作指定動画の一例
Fig. 9　Examples of movement movie

(3) 協力者に歩行方法を動画 (図 9) で提示
この動画はゲームエンジンの Unity を用いて作成した。3D モデルは Unity Asset Store にて公開されている Game Asset Studio の Taichi Charactor Pack[10]、3D モーションは Carnegie-Mellon Univercity の Huge FBX Mocap Library[11] の一部を利用した。

(4) データ収集ソフトを起動・収集開始し、協力者は歩行を開始

(5) 歩行方法毎に設定した時間の歩行データを記録し、次の歩行方法に移る（(3) に戻る）。全ての歩行方法の収集を終えている場合は、データ収集終了

収集した歩行データの総時間は、一人あたり 1 時間である。図 7 で示したように、床素材には、足音が比較的大きくなり、かつ様々なコンテンツで使用される頻度の高い畳を選択した。なお、すべての協力者で同じスリッパを用いた。

4. 足音合成実験

本節では、提案法を用いた足音の音響信号の合成実験について説明する。

4.1 特徴量

今回、入力・出力層の特徴量を抽象化することで、変換モデルの汎化性能の確保を試みた。特に足音などの効果音は、様々な床素材・靴素材等の組み合わせがありうる。床素材の特徴を考慮していない現状では、各素材の組み合わせ毎に変換モデルを構築する必要がある。変換モデルのパラメータ数（DNN であればユニット間の結合重み）が増加するほど、（モデルが十分な汎化性能を持つために）必要な学習データ数も増加する。そこで本研究では、出力層の音響特徴量の次元数を減らすことで DNN モデルのパラメータ数を少なくする方針を採用した。本実験では、以下の 3 つの音響特徴量を比較した。

- LPC(Linear Prediction Coefficients, 線形予測分析) 係数
- PARCOR(Partial auto-correlation, 編自己相関) 係数
- LSP(Line Spectral Pairs, 線スペクトル対) 係数

これらの音響特徴量は、音声分野で次元数を減らした状態でスペクトル包絡を表現するためのものである。本研究は効果音が対象ではあるが、スペクトルの包絡形状を再現できれば音響信号の合成は可能であると考えた。これらの特徴量は数学的には等価であるが、一般的に LPC 係数は量子化等による誤差によってフィルタの安定性が保証されない [12]。対して、PARCOR 係数、LSP 係数は誤差に対して比較的耐性があることが知られている。DNN モデルにおいて、教師データと全く同じ値を出力することは不可能であるため、学習する特徴量は誤差に対して耐性のあるものが望ましい。よって、PARCOR 係数や LSP 係数の利用が望ましいと考えた。

また、比較対象として、高次元特徴量であるフーリエスペクトルを出力層とした場合も検討した。離散フーリエ変換の際のフレーム長は 1,600 サンプル点 (サンプリング周波数 48kHz での録音を 8kHz にダウンサウンプリング、窓長 200ms) である。ここで入力信号が実数の場合、フーリエスペクトルは正と負で対称 (実部は偶関数、虚部は奇関数) となるため、正の周波数の情報のみで学習した。よって、出力層は 801 次元となる。なお、DNN モデルにおいて複素数をそのまま学習することは困難なため、フーリエスペクトルの実部と虚部をそれぞれ別の DNN として構築した。

4.2 変換モデルの学習手順と実験条件

収集した歩行データを用いて、変換モデルを学習する手順を以下に示す。

(1) 歩行データをフレーム処理し、センシングデータのみをフレーム連結

(2) Auto-Encoder を用いて、各層を教師なし学習によって Pre-training

(3) Backpropagation による教師あり学習である Fine-tuning
Pre-training した各層を接続し、収集した波形の音響特徴量を用いた教師あり学習を行なった。本研究では、DNN の出力と教師データの誤差を元に結合重みを更新する Backpropagation を用いた。

本実験における学習データの条件を表 1、DNN の学習パラメータを表 2 に示す。各隠れ層のユニット数とミニバッチサイズは、事前のグリッドサーチによって最も教師デー

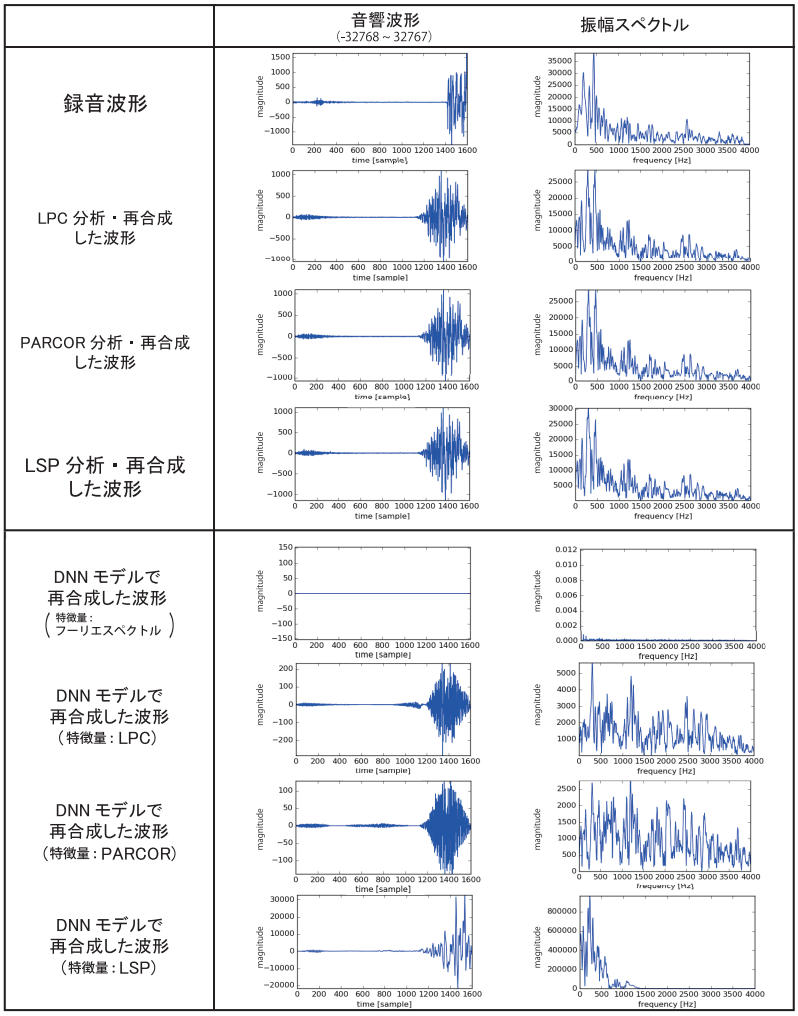

図 **10**　録音波形と出力波形

Fig. 10　Recorded signals and represented waveform

表 **1**　学習データに関する条件

Table 1　Conditions for training data

音の長さ	0.2s
音のサンプリング周波数	8kHz
センサデータの長さ	1.0s
センサデータのサンプリング周波数	60Hz
窓関数	ハミング窓
学習データ数	104,856

表 **2**　DNN の学習パラメータ

Table 2　Parameters of training DNN

DNN の形式		stacked Auto-Encoder
活性化関数		ReLU
Dropout		50%
学習方式		ミニバッチ
バッチサイズ		2,590
エポック数 (Pre-training)		1
エポック数 (Fine-tuning)		7,500
次元数	入力層	1,800
	隠れ層 1	460
	隠れ層 2	375
	出力層 (フーリエスペクトル)	801
	出力層 (LPC / PARCOR / LSP)	420

タとの平均二乗誤差が小さかった組み合わせの条件を求めて採用した。グリッドサーチは、隠れ層のユニット数については 5〜1,000 を 5 おきに、ミニバッチサイズは 10〜3,000 を 10 おきに試行した。

モデルの学習には Python 向けの深層学習ライブラリである Chainer[13] を利用した。LPC 係数、PARCOR 係数、LSP 係数の抽出には SPTK(Speech Signal Processing Toolkit)[14] を使用した。

4.3　合成結果

本実験では、学習データと同一のセンシングデータをテストデータとして用いる、クローズド条件下で音響信号の合成を行なった。図 10 に、録音波形と合成波形の具体例を示す。音響特徴量にフーリエスペクトルを利用した場合は、全ての波形がほぼ無音に近い出力となった。対して、低次元特徴量である LPC、PARCOR、LSP 係数を利用した場合は、録音波形に近い波形を合成できることを確認した。特に LSP 係数は、低音が強い元波形のスペクトルの傾向を比較的良く再現している。

Pre-training、Fine-tunung の学習過程を確認するため、DNN 内の重みの変化を可視化した（図 4.3)。第二隠れ層

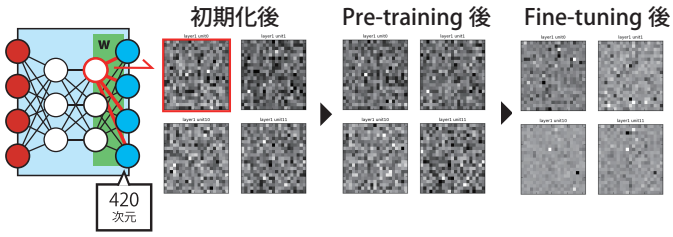

図 11　学習による DNN 重みパラメータの変化
Fig. 11　Trangition of weighting parameters in training DNN

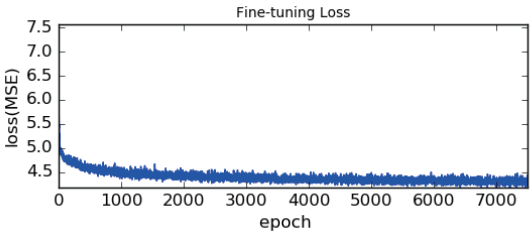

図 12　学習回数に対する平均二乗誤差の遷移
Fig. 12　Trangision of mean squared error for the training number of times

と出力層の接続部分の重みを、濃淡画像（色が濃いほど重みが大きい）で表している。学習する度に、初期化時のランダムな値から重みの大きい部分と小さい部分が分かれることがわかる。また、図 12 に示すとおり、学習回数が増える度に平均二乗誤差が小さくなっており、モデルの学習が進んでいることを確認できた。

5. おわりに

本研究は、人間の動作情報から足音の音響信号を合成する手法を検討した。変換モデルの次元圧縮のために線形予測系の低次元特徴量を用い、結果的に元波形に似た足音波形を合成できる可能性を示した。以下の事項が今後の課題として挙げられる。

- オープンテスト結果の改善
  本稿ではクローズドテストの結果を示したが、オープンテストでは、ほぼ全ての波形が無音に近い出力となった。この問題に対して学習データの拡充が必要である。また、入力信号に対しても抽象化が必要である。

さらに、提案システムの実用化では、以下の課題が挙げられる。

- 床素材を考慮した入力データの検討
  今回、床素材は畳に限定したが、多様な床素材における足音を一つのモデルで表現できれば、より有用なシステムが実現できる。例えば床の画像など、床素材を表現できる特徴量を検討し、それを入力データに含むことでこのようなシステムは実現できる可能性がある。
- リアルタイム合成
  今日は事前に収録したデータから効果音を合成することを前提とした。効果音作成の負担軽減という目的では十分であるが、ゲーム中の様々な状況変化に応じたリアルタイム効果音合成の需要は高いと考えられる。例えば現時刻の足音の波形より前フレームのセンサ情報のみを入力とすることで、未来の足音を予測するような検討が必要である。

謝辞　本研究は JSPS 科研費 JP15K01069, JP16H01734, JP15H02726 の助成を受けたものです。

参考文献

[1] 社団法人コンピュータエンタテインメント協会, CESA ゲーム開発技術ロードマップ 2015 年度版サウンド部門 (http://cedec.cesa.or.jp/2016/documents/roadmap/SND_2015.pdf) (2016.12.22 アクセス確認).
[2] K. van den Doel, P.G. Kry and D.K. Pai, FoleyAutomatic: Physically-based Sound Effects for Interactive Simulation and Animation, in Computer Graphics (ACM SIGGRAPH 01 Conference Proceedings), 2001.
[3] K. van den Doel, PHYSICALLY-BASED MODELS FOR LIQUID SOUNDS, Proc. of ICAD 04-Tenth Meeting of the International Conference on Auditory Display, 2004.
[4] A.J. Farnell, Marching onwards: procedural synthetic foot- steps for video games and animation, Proc. of the Pure Data Convention, 2007.
[5] Turchet, L., Serafin, S., Dimitrov, S., Nordahl, R, Physically Based Sound Synthesis and Control of Footsteps Sounds, Proc. of the 13th International Conference on Digital Audio Effects (DAFx-10), vol. 1, pp.161-168, 2010.
[6] 松田拓也・甲斐義弘・井上喜雄・谷岡哲也, 足底圧計測装置を用いたインテリジェント歩行支援機の制御, 第 45 回システム制御情報学会研究発表講演会講演論文集, pp.221-222, 2000.
[7] 高木信二, 山岸順一, Deep Neural Network に基づく音響特徴抽出・音響モデルを用いた統計的音声合成システムの構築, 情報処理学会研究報告, 2015-SLP-105(2), 1-6, 2015.
[8] H. Zen, A. Senior, and M. Schuster, STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS, Proceedings of ICASSP, pp.7962-7966, 2013.
[9] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, William T. Freeman, Visually Indicated Sounds, Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2405-2413, 2016.
[10] Game Asset Studio, Taichi Charactor Pack (https://www.assetstore.unity3d.com/jp/#!/content/15667) (2016.12.25 アクセス確認).
[11] Carnegie-Mellon Univercity, Huge FBX Mocap Library (https://www.assetstore.unity3d.com/jp/#!/content/19991) (2016.12.25 アクセス確認).
[12] 菅村昇, トピックス 14 線形予測分析 LPC の発明→ PARCOR → LSP へ (〈特集〉-音響学における 20 世紀の成果と 21 世紀に残された課題-), 日本音響学会誌 vol.57, no.1, pp.66, 2000.
[13] Preferred Research, Chainer (http://chainer.org/) (2016.12.25 アクセス確認).
[14] Satoshi Imai, Takao Kobayashi, Speech Signal Processing Toolkit (SPTK) (http://sp-tk.sourceforge.net/) (2016.12.25 アクセス確認).