福井大学 学術研究院工学系部門 研究報告 第 71 巻 2023 年 3 月
Mem. Fac. Eng., Univ. Fukui, Vol. 71 (March 2023)

目　　次

【第 71 巻 11 月分】

【第 71 巻 3 月分】

Memoir of Faculty of Engineering, University of Fukui
Vol. 71, March 2023

CONTENTS

【Vol.71 November】

【Vol.71 March】

声質変換による合成音声話者の拡張

髙村 健也* 原田 楓* 小髙 知宏** 黒岩 丈介** 白井 治彦*** 諏訪 いずみ****

Increasing Variation of Synthetic Speech Speakers by Voice Conversion Technology

Kenya TAKAMURA*, Kaede HARADA*, Tomohiro ODAKA, Jousuke KUROIWA**, Haruhiko SHIRAI***, Izumi SUWA*****

(Received September 30, 2022)

In this paper, we used voice conversion techniques on synthetic speech to increase the variation of synthetic speech speakers. We performed voice conversion using 'CycleGAN-VC2' which is the non-parallel voice conversion model that does not impose significant restrictions on training data and evaluated output voices. The data set for the experiment was a synthetic voice with four different speakers (two female and two male) from the 'Google Cloud Text-to-Speech' service for source voices. For target voices, we prepared narration voices by the one male speaker.

As a result of the experiment, the voice conversion process was properly performed except for one type of female speaker, and it was possible to change the speaker. The problem with this method was that the conversion process loses the human-like voice quality found in source voices. The conclusion of our method is that it is useful for increasing variation of synthetic voice speakers if the transformation process can be made to sufficiently satisfy quality preservation.

***Key Words*:** Voice Conversion, Synthetic Speech, CycleGAN

1. 緒言

コンピュータによって音声を生成する音声合成技術は人工知能の根幹をなすディープニューラルネットワークモデルにより発展しており，最新の合成音声は人が話す音声に近い品質であることが示されている[1]. このような合成音声は, 電化製品を始めとした様々な物をインターネットに接続して活用する現代社会において，コンピュータからの動的な応答を人に伝える役割を持つ．その最たる例が，スマートフォンやスマート家電に搭載されている音声アシスタントであり，具体的には Google LLC による「Google アシスタント」や Amazon.com, Inc.による「Amazon Alexa」が挙げられる．このような合成音声の話者に着目すると，出力言語を日本語と設定した場合に選択可能な話者は女性 1 種類・男性 1 種類となっていることが多く，利用者はせいぜい 2 種類程度の話者を選択することに限られている現状がある．もしも合成音声話者のバリエーションが豊富であるならば，利用者はそれぞれの好みに合わせた話者を自由に選択し，コンピュータとのやり取りを自由に楽しむことができると考えられる．

そこで，本研究では合成音声に対して声質変換技術を用いることで，合成音声話者のバリエーションを拡張することを試みた. ここで声質変換技術とは，声の高さや抑揚などの音響的特徴を変換し，変換前の音声とは異なった声質を持つ音声を生成するものである[2].

* 大学院工学研究科知識社会基礎工学専攻
** 知能システム工学講座
*** 工学部技術部
**** 仁愛女子短期大学
* Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering
** Department of Human and Artificial Intelligent Systems
*** Technical Division
**** Jin-ai Women's College

2. 合成音声に対する声質変換の適用

2.1 声質変換の手法

声質変換を行う方法として，パラレル変換とノンパラレル変換の 2 種類がある．これらの方法は，声質変換モデルを構築するために必要な，変換の元とする音声データ（Source 音声）と変換によって再現したい声質の音声データ（Target 音声）の関係性によって分類される．パラレル変換手法では 2 種の音声データに対して，同一の文章を読み上げさせて発話内容を合わせた後に，音素レベルで発話タイミングを同期させることが必要となる．同期の精度を良い状態にするには手動での綿密な調整が不可欠であり，パラレル変換はデータセットの用意が困難である．一方でノンパラレル変換手法では，2 種の音声データに対して制約を課さないため複雑な前処理が不要となる．

合成音声話者の拡張を目的とする本研究においては，合成音声による Source 音声と再現目標である Target 音声の組み合わせは様々であり，複数種類の声質変換モデルを生成するための手法はより簡単な方が好ましい．そのため，本研究ではノンパラレル変換の手法を採用する．

2.2 使用する声質変換モデル

本研究ではノンパラレルな声質変換手法として，「CycleGAN-VC2」声質変換モデルを使用する[3]．このモデルは深層学習モデルである CycleGAN をベースとしている[4]．CycleGAN のアーキテクチャを図 1 に示す．CycleGAN は異なる 2 つのドメイン間のマッピングを行うモデルであり，4 種のニューラルネットワーク（Generator 2 種，Discriminator 2 種）から構成される．Generator は一方のドメインに属するデータをもう一方のドメインに属するデータに似せるようにマッピングを行う役割を持ち，Discriminator は Generator が生成したデータとドメインに属するデータとを分類する役割を持つ．CycleGAN では，これらのニューラルネットワークを適切に制御するために以下の 3 種類の損失を用いて学習を行う．

- Adversarial Loss
- Cycle Consistency Loss
- Identity Mapping Loss

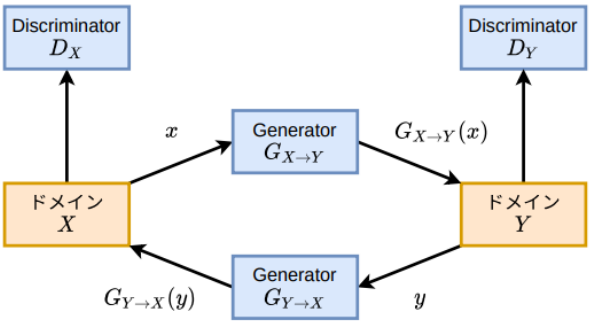

図 1　CycleGAN のアーキテクチャ

Adversarial Loss は，Generator と Discriminator が敵対するように機能させるための損失であり，Generator のマッピングにおけるデータの劣化を軽減する役割を持つ．Cycle Consistency Loss は，一方向の Generator の出力を逆方向の Generator の入力とし，最終的な出力を最初の入力と比較する損失であり，データの構造情報の保持を促進する役割を持つ．Identity Mapping Loss は，あるドメインに属するデータを，所属ドメインが変わらないように Generator でマッピングし，その出力を元の入力を比較する損失であり，声質変換タスクでは音声データに含まれる言語情報の保持を促進する役割を持つ[5]．

2.3 声質変換の流れ

学習済みの CycleGAN-VC2 モデルによる声質変換の流れを図 2 に示す．まず，音声合成分析器 WORLD（D4C edition）を使用して変換の対象とする音声から以下の 3 つの特徴量を抽出する[6] [7]．

- メルケプストラム（MCEP）
- 基本周波数
- 非周期性指標

MCEP は音の音色，基本周波数は音の高低，非周期性指標は有声音における雑音成分に対応する．

次に抽出した特徴量に対してパラメータ変換を行う．変換の対象とするのは MCEP と基本周波数であり，非周期性指標には何の処理も行わない．MCEP には，変換モデルを学習する際に使用するトレーニングデータから得られる統計情報を元にした線形的な変換処理と，Generator によるマッピング処理を組み合わせて適用する．基本周波数には，自然対数をとった後に同様の線形変換処理のみを適用する．ここで扱う線形変換は次のように表される．

$$f_{conerted} = \frac{f - \mu_s}{\sigma_s} * \sigma_t + \mu_t \quad (1)$$

ここで，fはある特徴量，μ_s, σ_sは Source 音声話者のトレーニングデータから得られる特徴量fの平均と標準偏差，μ_t, σ_tは Target 音声話者のトレーニングデータから得られる特徴量fの平均と標準偏差を意味している．

最後に，パラメータ変換処理を行った各特徴量を WORLD によって合成し，声質変換を適用した音声を出力する

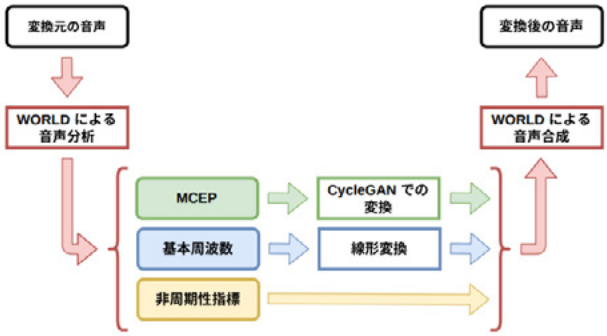

図 2　声質変換の流れ

3.　実験

3.1　音声データの取得と処理

実験で使用する音声データの種類は Target 音声と Source 音声の 2 種類であり， Source 音声に対して変換処理を行うことで，まるで Target 音声の話者が話しているような音声を生成する．

Target 音声を取得するために，書籍をナレーターや西友が朗読した音声データを配信するサービス「audiobook.jp」を利用した[8]．このような音声には目的とする音声以外にノイズや効果音がなどの音が含まれていないためである．Target 音声の話者としてプロの男性声優を設定し，この男性によるナレーション音声のうち，以下の 3 つの作品を購入した．

- 走れメロス（太宰治 著）
- 杜子春（芥川龍之介 著）
- 子供おばさん（山本文緒 著）

これらの音声は，拡張子が mp3，チャンネル数がステレオ，サンプリング周波数が 44.1[kHz]である．

取得した Target 音声に対して行う処理は以下の 3 つである．

- 有音区間の抽出
- 拡張子の変更
- チャンネル数，サンプリング周波数の変更

有音区間の抽出では，単語や文章の間に挟まれる無音の区間を排除する.「走れメロス」以外の音声では，音量の閾値を-24.0[dB]，無音部の最大持続時間を 0.80[s]とするなど，一定条件の設定下においてプログラム的に抽出した.「走れメロス」の音声では，作品の原文を参照しながら，文章の構造に従い手動で抽出を行った．拡張子の変更では，音声解析の際によく利用される wav 形式へと変更し，量子化ビット数を 16[bit]とする．チャンネル数，サンプリング周波数の変更では，後述する Source 音声のものと同期させるために，ステレオからモノラル，44.1[kHz]から 24.0[kHz]へと変更する．

Source 音声の取得では，本研究の目的を考慮して音声合成技術によって生成された音声データでなければならない．そのため，任意の自然言語によるテキストデータを入力することで，そのテキストの読み上げ音声を出力する音声合成サービス「Google Cloud Text-to-Speech」を利用した[9]．このサービスにおけるテキストを読み上げる話者は，高品質な日本語を話す女性 2 種，男性 2 種の 4 種類とした．

- ja-JP-Wavenet-A （女性）
- ja-JP-Wavenet-B （女性）
- ja-JP-Wavenet-C （男性）
- ja-JP-Wavenet-D （男性）

本論文では，これらの話者を「Source A」あるいは「Source 話者 A」といったように表記する．入力とするテキストについては，日本語のテキストとその読み上げ音声のセットである「JUST コーパス（ver1.1）」に含まれる「basic5000」と，Target 音声の処理において分割した「走れメロス」の音声それぞれと発話内容が一致するようにしたテキストである[10]．取得した Source 音声は，拡張子が wav，量子化ビット数が 16[bit]，チャンネル数がモノラル，サンプリング周波数が 24.0[kHz]である．

3.2　データセットの設定

取得した音声データを基にして，声質変換モデルを学習するために利用するトレーニングデータと，学習後の変換モデル性能を検証するために利用するテストデータを構築する．

まず Target 音声について,「杜子春」と「子供おばさん」をトレーニングデータ,「走れメロス」をテストデータとして利用する．Source 音声については,「basic5000」をトレーニングデータ,「走れメロス」をテストデータとして利用する．つまり，Target と Source 音声のテストデータにおいては，発話内容が同期している形式となっている．

次に，トレーニングデータの詳細な設定として，再生時間に着目してフィルタリングを行う．Target 音声のトレーニングデータは一定の条件下で有音区間を抽出したものであり，学習に適さないほど再生時間が短いものが存在しているためである．再生時間が 3.0[s]未満のデータをフィルタリングすることで，Target 音声の総ファイル数は 502 個となった．Source 音声のデータに対しても同様のフィルタリングを行い，残ったものから 502 個のファイルをランダムに選択した．トレーニングデータの詳細を表 1 に示す.

最後に，テストデータの詳細な設定として，再生時間とセリフ文に対するフィルタリングを行い，各話者間で発話内容が完全に同期するように除去を行う. 具体的には再生時間のフィルタリングは 1.5[s]未満とし，Target と Source を比較して読み方が大きく異なっているセリフ文を除く．これに加え，ある発話内容の音声が除去された際にはその発話内容を持つ音声を全体から除外する．これらの結果，各テストデータのファイル数は 266 個となった．テストデータの詳細を表 2 に示す.

表 1　トレーニングデータの詳細

話者	ファイル数	総再生時間 [s] (平均±標準偏差)
Target	502	2,850 (5.7±2.7)
Source A	502	3,107 (6.2±2.6)
Source B	502	2,892 (5.8±2.4)
Source C	502	2,765 (5.5±2.3)
Source D	502	2,805 (5.6±2.4)

表 2　テストデータの詳細

話者	ファイル数	総再生時間 [s] (平均±標準偏差)
Target	266	1,054 (4.0±2.6)
Source A	266	1,263 (4.7±2.6)
Source B	266	1,180 (4.4±2.4)
Source C	266	1,129 (4.2±2.3)
Source D	266	1,143 (4.3±2.4)

3.3　声質変換モデルの学習

はじめに，各話者のトレーニングデータから WORLD を用いて音響特徴量を抽出する．抽出する特徴量は基本周波数と MCEP の 2 つであり，音声の再生時間に対して 5.0[ms]を 1 つのフレームとして，それぞれの特徴量を抽出する. この時 MCEP は，512 次元のスペクトル包絡を取得して 36 次元の MCEP に変換することで抽出を行う.

次に，それぞれの特徴量から統計情報として平均と標準偏差を求めて保存する．基本周波数は自然対数をとった後にそのまま算出する．MCEP は平均と標準偏差を算出した後に，MCEP に対して，平均を 0，標準偏差を 1 とする正規化を行う．変換モデルの学習は，この処理によって得られた MCEP を用いて行う.

次に，それぞれの特徴量から統計情報として平均と標準偏差を求めて保存する．基本周波数は自然対数をとった後にそのまま算出する．MCEP は平均と標準偏差を算出した後に，MCEP に対して，平均を 0，標準偏差を 1 とする正規化を行う．変換モデルの学習は，この処理によって得られた MCEP を用いて行う.

モデルの学習では，1 人の Source 話者と Target 話者に対応するデータ間で行われるため，CycleGAN-VC2 における 2 つのデータドメイン(X, Y)の組み合わせは(Source A, Target)，(Source B, Target)，(Source C, Target)，(Source D, Target)の 4 つとなる．学習時に与える入力データは 36 次元の MCEP におけるランダムな個所の連続した 128 フレーム分のセグメント

表 3　実験環境

OS	Ubuntu 20.04.3 LTS
CPU	AMD Ryzen 7 1800X
GPU	NVIDIA Geforce 1080Ti （CUDA 11.3）
開発言語	Python 3.8.10
機械学習ライブラリ	PyTorch 1.10.0

とする．502 回のイテレーションを 1Epoch とし，学習回数を 20 から 200Epoch の範囲で行う．この際，Cycle Consistency Loss とトレードオフの関係にある Identity Mapping Loss を制御するために，学習の進行具合に応じて重み付けを定めるパラメータの値を 0 へ変更して複数種類のモデル学習し，それぞれのモデルの性能を評価する．

本実験における環境を表 3 に示す．モデルの学習には，機械学習ライブラリ PyTorch と並列計算を行うためのプラットフォームである CUDA を利用する．

3.4　声質変換モデルによる合成音声の評価

声質変換は 2.3 節で述べた流れに沿って行う．WORLD によって音響特徴量を抽出する手法は，モデルを学習する時と同様である．モデルによって変換される 36 次元の MCEP は，512 次元のスペクトル包絡に変換され，変換後の基本周波数と無変換の非周期性指標と共に合成される．

生成した音声を評価するために，本研究では以下の 2 つの点に着目した．

- 声質的評価
- 言語的評価

声質的評価では，変換モデルによって Target 話者の声質が適切に再現されているかどうかを評価する．声質の変換性能を客観的に評価する指標として，Mel-cepstral distortion (MCD)を使用した．MCD は次のように表される．

$$MCD[dB] = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{d=1}^{D-1}\{mc_d^X(t) - mc_d^Y(t)\}^2} \quad (2)$$

ここで，Dは MCEP の次元数，Tは MCEP のフレーム数であり，$mc_d^X(t)$と$mc_d^Y(t)$はt番目のフレーム，d次元の MCEP を表している．MCD は比較する MCEP の違いを算出するものであり，値が小さいほど違いの程度が小さくなることを意味する．MCD の算出時には，比較する MCEP に対して Dynamic Time Warping を適用してパラレルな状態とする．

言語的評価では，声質変換適用の前後において言語情報が保持できているかどうかを評価するために，音声データから読み上げられているテキスト情報を取得し，そのテキストを比較することで評価を行う．テキスト情報の取得には，Google 社が提供する「Google Cloud Speech-to-Text」を使用する[11]．取得した変換前と変換後のテキスト情報の比較は文字単位で行い，変換によって失われた言語情報を算出する．評価指標として失われた言語情報の割合[%]を，変換によって失われた語数をもとの文章全体の語数によって除算することで算出する．

各評価に使用するテストデータについて，声質的評価では全てのデータを使用する．一方で，言語的評価では Source 話者のテストデータから再生時間[s]が[1.5, 3.0), [3.0, 6.0), [6.0, 9.0)の各範囲にある音声データかを 5 つずつランダムに選択し，合計 15 個のデータを対象とした．この際に，全 Source 話者の間で 15 種類の各発話内容が同一となるようにする．

3.5　声質変換の処理時間計測

合成音声に対して声質変換を適用し話者を変更するという本手法において，元の合成音声を取得してから最終的な音声を再生するまでの間に処理の時間が必要となる．この時間は，声質変換を適用せずに音声を再生する場合を基準とすると，追加で掛かる時間である．この時間の程度を計測するために，学習モデルや統計情報をロードした状態の声質変換システムにおいて，変換する音声データファイルを読み込むタイミングで計測を開始し，変換を適用した音声データファイルを出力したタイミングで計測を停止する．

4.　実験結果

4.1　声質変換モデルの性能

声質変換を適用する前の各 Source 話者と Target 話者のデータ間で算出した声質的評価，および声質的評価と言語的評価の観点において最良である声質変換モデルの各評価値を表 4 に示す．表中の値は，平均値と標準偏差を意味する．声質的評価に着目すると，変換前，変換後ともに Source 話者 D の値が最小

表 4　声質変換モデルの評価値

話者	声質的評価		言語的評価
	変換前	変換後	
Source A	28.05±5.65	17.45±2.76	0.55±0.22
Source B	26.31±5.63	16.05±1.84	0.10±0.14
Source C	25.04±5.79	15.52±2.24	0.04±0.04
Source D	24.29±5.82	14.92±2.23	0.05±0.08

表 5　各 Source 話者の変換処理時間

話者	変換処理時間
Source A	0.28±0.01
Source B	0.28±0.01
Source C	0.30±0.01
Source D	0.30±0.01

であり，最も Target 音声を再現している．Source 話者 D の言語的評価も 0.05（5%）であり，変換による影響が非常に小さい．

声質的評価と言語的評価を総合した結果として，声質変換が適切に行われた話者は Source B，Source C，Source D である．一方で，変換後の音声をイヤホンで聞く主観的な評価として，元の合成音声が持つ自然さは少なからず失われており，ノイズが含まれていると感じられ，元の音声の品質を適切に保持できていないといえる．

4.2　声質変換の処理時間

異なる条件のもと学習を行った各モデルによる，声質変換の処理時間を計測し，各データの変換に要した処理時間をそのデータの再生時間で除算することで，1 秒間の音声を変換するのに必要な処理時間[s]の平均値および標準偏差を算出した．各 Source 話者の変換処理時間を表 5 に示す．結果として，学習条件の違いは処理時間に影響を与えず，各 Source 話者の変換モデルによる処理時間に大きな違いはない．

5.　考察

声質変換モデルによる声質的評価に着目すると，女性話者である Source A と Source B，男性話者である Source C と Source D の順で低くなっている．Target 話者の性別が男性であることを考慮すると，Source 話者と Target 話者同士の変換難易度は性別が同じ，あるいは声が似通っている場合に下がりやすいといえる．言語的評価にも着目した場合も，Source C，Source D の値が非常に小さく，ほかの Source 話者よりも声質変換が適切に行われているといえる．

声質変換の処理時間については，再生時間が 1 秒の音声を変換するために約 0.30[s]要する．つまり，対象とする音声が 5 秒間であれば 1.50[s]，10 秒であれば 3.0[s]の時間が必要となる．WaveNet モデルが 1 秒間に 20.0[s]の音声を出力する，すなわち 1 秒間の音声を生成するために 0.05[s]しか必要としないことを考慮すると，本手法の必要時間は少し長いといえる[1]．人とコンピュータが音声でやり取り可能な環境を提供するアプリケーションに本手法を組み込むことを考えると，コンピュータの応答を人に伝えるプロセスにおいて声質変換の時間がボトルネックになり，アプリケーションの応答時間が人に悪い印象を与える可能性は否めない．

声質変換の結果としては，変換自体は適切に行えているモデルを構築することができたが，元々の音声の品質が失われているものとなった．変換処理の後にノイズ除去などの品質向上を目的とした追加の処理を実行することを考えた場合，話者拡張に要する時間がさらに大きくなることは明白である．そのため，声質変換処理のみでこの問題を解決することが望ましく，モデルの改善や他の声質変換手法を検証することが必要となる．

6.　結言

本研究では，スマートフォン等に搭載されている音声アシスタントなどの音声合成を利用したサービスにおいて音声話者の種類が少ないことに着目し，話者のバリエーションを拡張するために声質変換技術を合成音声に適用して，その出力音声の評価を行った．変換元とする合成音声は Google Cloud Speech-to-Text サービスにおける 4 種類の話者（女性 2 種，男性 2 種）による音声とし，変換先とする音声は一般に販売されている 1 種の男性話者によるナレーション音声とした．結果として，1 種の女性話者を除いて，適切に話者性を変更することが可能であった．この結果から，対象とする音声同士の声質変換に対する相性の重要性を確認でき，和者数が 1 種類しか

存在しないような音声合成サービスにおいては本手法が適用できず，話者性を拡張できない可能性が考えられる．

本手法の問題点としては，最新の合成音声にみられる，まるで人が話しているような音声の品質が変換処理によって失われていることであった．そのため，変換処理によって品質の保持を十分に満たすことができれば，話者のバリエーションを拡張するのに有用であると判断できる．

参考文献

[1] Google Cloud 'Introducing Cloud Text-to-Speech powered by DeepMind WaveNet technology', <https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-text-to-speech-powered-by-deepmind-wavenet-technology>, (2022/01/10).

[2] 戸田智基．確率モデルに基づく声質変換技術日本音響学会誌，Vol.24，No.1，pp.34-39，2010．

[3] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6820-6824, 2019.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, pp. 2223–2232, 2017.

[5] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv:1711.11293, 2017.

[6] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE TRANSACTIONS on Information and Systems, Vol. E99-D, No. 7, pp.1877-1884, 2016.

[7] Masanori Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. Speech Communication, Vol. 84, pp.57-65, 2016.

[8] Audiobook.jp, <https://audiobook.jp/ >, (2020/07/13).

[9] Google Cloud Text-to-Speech, <https://cloud.google.com/speech-to-text>, (2021/04/21)

[10] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. arXiv preprint arXiv:1711.00354, 2017

[11] Google Cloud Speech-to-Text, <https://cloud.google.com/speech-to-text>, (2022/01/14)

Transformerモデルによる感情を基にした動画BGMの生成と評価

原田楓 * 小高知宏 ** 黒岩丈介 ** 諏訪いずみ *** 白井治彦 ****

Generation and Evaluation of Emotion-based Video BGM Using Transformer Model

Kaede HARADA* Tomohiro ODAKA Jousuke KUROIWA**
Izumi SUWA*** Haruhiko SHIRAI****

(Received September 30, 2022)

In this paper, we developed a system that automatically generates appropriate background music for a video using the emotions contained in the video as input. A dataset of piano music was created for each emotion and trained using Music Transformer models. Music was generated by inputting emotion words to the 4 models that had been trained. Two evaluation experiments were conducted on the generated songs. In the evaluation experiment of the music piece alone, we asked the subjects to listen to and evaluate 16 generated music. In the evaluation experiment of the videos as background music, we asked the subjects to watch 4 videos without sound, and determined the emotions of the videos. Then, 16 videos with background music were created by combining the 4 videos and 4 songs that contained each emotion. Then, we asked the subjects to watch the created videos and evaluate which background music was most suitable for them. For the experimental results, some of the models produced suitable music for the videos with expected emotions, but not all of the models were able to produce suitable music for the videos. In order to complete the system, we need to improve the training dataset and revise the experimental method.

Key words **:** music genelation, Transformer, BGM

1. 緒言

近年, 動画サイトの使用率, 使用人数は増加しており, 加えて動画投稿者の数も増えている. また, ショート動画を投稿する機能に力を入れたSNSアプリが普及し, 誰でも手軽に動画の制作や投稿ができる環境にあると言える. 加えて, 多種多様なジャンルの動画, 動画投稿者の増加によって日本を含め, 世界の様々な分野に影響を与えている. このことから, 動画というコンテンツの需要が増加傾向にあると考えられる.

動画制作の重要な要素の一つとして, 適切なBGMの選択が挙げられる. 動画の内容にマッチしたBGMを付与することで, 動画のクオリティ向上に繋がると考えられる. しかし, 個人利用, 商業利用が可能なBGMに絞ったとしても存在する楽曲数は膨大である. そのため, 個人が気に入るBGMの選択, 動画に合わせる編集などの作業には時間を要する.

本研究では, 動画に含まれる感情を入力として, 動画に適したBGMを自動生成するシステムの開発を目的とする. BGMを生成する手法には, 深層学習モデルであるTransformerを応用した音楽生成モデルMusic Trans-

*大学院工学研究科 知識社会基礎工学専攻
*Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering
**知能システム工学講座
**Department of Human and Artificial Intelligent Systems
***仁愛女子短期大学 生活科学学科
***Jin-ai Women's College
****工学部 技術部
****Technical Division

former を用いる. Music Transformer には楽曲の感情を基に作成したデータセットを学習させる. 学習させたモデルに感情を入力して生成された BGM を評価することで, 動画に適した BGM が生成されているか判定する.

本論文では, 2 章で現在の動画サイト等の動向, BGM 生成に関する研究について述べる. 3 章では Music Transformer を用いた楽曲の生成について述べる. 4 章で生成された楽曲の評価実験について説明し, 5 章で評価実験の結果を示す. 6 章で実験結果に対する考察を述べ, 7 章で本論文についてのまとめを述べる.

2. 動画サイトの現状と BGM の自動生成

近年では, 動画投稿を主軸とした SNS の普及によって, 動画サイトや SNS の利用者数は増加傾向にある. 加えて, 操作が簡易化された動画作成アプリの登場などにより, 動画の投稿者数も増加している. 動画投稿者の増加に伴い, 以前より多種多様なジャンルの動画が確立され, 投稿される動画の本数も増加している. また, 視覚と聴覚に情報を伝達できる動画というコンテンツの制作において, 適切な BGM の選択は重要な要素の一つである. このことから, BGM の選択にかかる時間の簡略化, 個人の気に入る BGM の生成という点に需要が存在すると考える. 2.1 節で現在の動画サイト等の動向, 2.2 節で動画と BGM の自動生成に関する過去の研究について述べる.

2.1 動画サイト等の動向

1.YouTube

「YouTube」は Google 社が運営する動画プラットフォームである. 2021 年には, 全世界で 23 億人以上 (ソーシャルメディアユーザーの 50 %以上) が月に 1 回利用している. 全世界ユーザーのうち, 日本のユーザー数は 6000 万にも上る. また, 投稿者として活動するために新しく開設されたチャンネルの数は 2019 年〜2020 年にかけて約 2 倍に増加している. クリエイターの増加により, 日本では音楽産業への貢献, 学習ツールとしての需要増加, 国を超えたカルチャーの発信などに影響を与えている [1][2] .

2.TikTok

「TikTok」は 2016 年にサービスを開始し, 2021 年には ios, Google Play を合わせると世界で最もダウンロードされたモバイル向けの動画プラットフォームである. 2022 年には, 月間アクティブユーザー数が 10 億人を超え, 世界第 6 位のソーシャルネットワークアプリとなった. 全世界ユーザーのうち, 約 50 %が 34 歳以下, 32.5 %は 10 歳〜19 歳の利用者であることから, 若年層を中心に成長している SNS アプリだと言える. TikTok の特徴として, エンゲージメント率 (視聴者の反応率) の高さが挙げられる. TikTok におけるマイクロインフルエンサーのエンゲージメント率は約 18 %, メガインフルエンサーのエンゲージメント率は約 5 %と他の動画プラットフォームの約 5 倍である. このことから, 動画メディアによるマーケティング効果が大きく期待できるアプリだと言える [3] .

これら以外にも, 様々な動画プラットフォームがアクティブユーザー数や動画投稿数の面で成長を見せており, 動画という情報メディアの需要が増加していることが伺える.

2.2 BGM の自動生成に関する研究

BGM の自動生成に関する研究は数多く存在する. 特に近年では, 自然言語処理の分野に登場した深層学習モデル Transformer を応用した音楽生成モデル Music Transformer が注目されている. 2018 年に Curtis Hawthorne らによって発表された研究では, ピアノ演奏の midi データで構成された大規模データセット「MAESTRO」を用いて, オリジナルのピアノ演奏を生成するモデルの開発とテストを行った [4] . この研究では, モデルの一部に Music Transformer を使用しており, 生成されたピアノ演奏楽曲はデータセットに含まれるピアノ演奏と同様の音楽的特性を持っているという結果が得られた. 加えて, 長期的な依存関係を含んだ約 1 分程度のピアノ演奏楽曲の生成が可能になった.

本研究の目的は, 感情を基にして動画に適した BGM を自動生成するシステムの開発である. そのため, 感情を基に収集した楽曲のデータセットを作成し, Music Transformer に学習させることで, ある感情の音楽的特徴を持った楽曲が生成できると考えた. この理由から, 本研究では楽曲の学習, 生成モデルに Music Transformer を使用することとした.

3. 楽曲の生成

本章では, Music Transformer を用いた感情語を入力とした楽曲の生成方法について述べる. まず, 感情語の取得方法や設定について説明する. 次に, 使用する Music Transformer モデルについて説明し, モデルに学習させるデータセットの作成方法を述べる. 最後に, Music Transformer を用いた楽曲生成の方法を説明する.

3.1 感情語の取得と設定

本研究では, 動画の内容から楽しい, 悲しいなどの感情語を取得し, 取得した感情語を Music Transformer に入力することで感情に適した BGM の生成を行う. 目標は, 自動で動画内容から感情語を取得できるようにすることだが, 今回は手動で動画に含まれる感情語を取得した. 対象者に動画を視てもらい, どのような感情を含んでいるか判定してもらうことで, 動画の持つ感情語を決定した. 得られた感情語を Music Transformer への入力に使用した. また, 感情には様々な種類の表現が存在するが, 今回は動画から取得する感情を楽しい, 穏やか, 悲しい, 恐ろしいの 4 種類に設定した.

3.2 使用モデルの説明

本研究では, データセットの学習, 楽曲の生成に Music Transformer を使用する. Music Transformer は 2018 年に Google LLC が発表した Transformer を用いた自動作曲 AI である [5] . 回帰型ニューラルネットワーク (RNN) や長短期記憶ネットワーク (LSTM) の登場により, 楽曲 (時系列データ) の過去情報を長期的に記憶することが可能になった. Music Transformer では, 情報の長期記憶性能が強化されている. これによって, 従来のモデルでは困難だった繰り返しフレーズ (楽曲中に複数回登場する似たメロディ) を持った楽曲の生成が可能になった. また, 音楽には文脈性が存在する. 音楽では, 前の時系列データ情報から, 次に使用できる音階 (スケール) や曲の調 (キー) などが決定される. しかし, 長い時系列になるほど文脈性の表現は難しくなる. Music Transformer の特徴として, 音楽の文脈性を従来のモデルより上手に表現できることも挙げられる.

3.3 データセットの作成方法

今回使用する Music Transformer モデルでは, 学習データ, 楽曲生成時の入力データにピアノ楽曲の midi データが必要となる. そのため, 以下の配布サイトでピアノ BGM を収集し, データ形式を変換することで, データセットの作成を行った.

- 音楽の卵 (https://ontama-m.com/)
- 甘茶の音楽工房 (https://amachamusic.chagasi.com/)
- ポケットサウンド (https://pocket-se.info/)
- 魔王魂 (https://maou.audio/)

はじめに, 各配布サイトにて, 4 種類の感情 (楽しい, 穏やか, 悲しい, 恐ろしい) をキーワードとして検索を行い, ピアノ BGM を wav データ形式でダウンロードした. そして, 私自身がダウンロードした BGM を聴き, 感情に沿った楽曲であるか判断を行った. 楽曲は各感情で 20 曲, 計 80 曲収集した. その後, オンラインファイルコンバーター AnyConv[6] を用いて, 収集した楽曲を wav データから midi データに変換した (図 1). これらの作業を行い, 各感情 (楽しい, 穏やか, 悲しい, 恐ろしい) のデータセットを作成した.

図 1 : データセットの作成工程

3.4 学習方法と生成方法

入力する感情語に適した楽曲を生成するためには, 各感情を持つ楽曲の特徴を学習したモデルが必要になる. そこで, 3.2 節で説明した各感情のピアノ楽曲データセットをそれぞれ Music Transformer に学習させる. 楽しいピアノ楽曲 20 曲分のデータセットを学習させた FUN モデル, 穏やかなピアノ楽曲 20 曲分のデータセットを学習させた RELAX モデル, 悲しいピアノ楽曲 20 曲分のデータセットを学習させた SAD モデル, 恐ろしいピアノ楽曲 20 曲分のデータセットを学習させた FEAR モデルの計 4 種類の楽曲生成モデルを作成する. 作成したモデルを用いて既存の BGM 数秒を始めとしたオリジナル楽曲を生成する.

作成した 4 種類のモデル (FUN モデル, RELAX モデル, SAD モデル, FEAR モデル) を用いて BGM の生成を行う. 学習させたモデルに感情語を入力すると, まずモデルは入力された感情語と関連性のある既存の BGM を選択する. そして, 選択された BGM の冒頭数秒を始めとして, 選択された BGM の持つ音楽的情報 (使用されている音, 音の長さなど) を基に続きとなるメロディーを生成する. 最後に, 生成された BGM を wav 形式で出力する (図 2).

今回は評価実験に使用するために, 4 種類のモデルでそれぞれ楽曲を 20 曲 (合計 80 曲) 生成を行った. 生成された BGM の長さは約 30 秒〜1 分 30 秒だった.

4. 楽曲の評価実験

本章では, 生成した楽曲が動画 BGM として適してるのか判定するために 2 種類の評価実験を行う. 4.1 節では生成楽曲の評価実験について, 4.2 節では生成楽曲を動画の BGM として使用した際の評価実験について説明する.

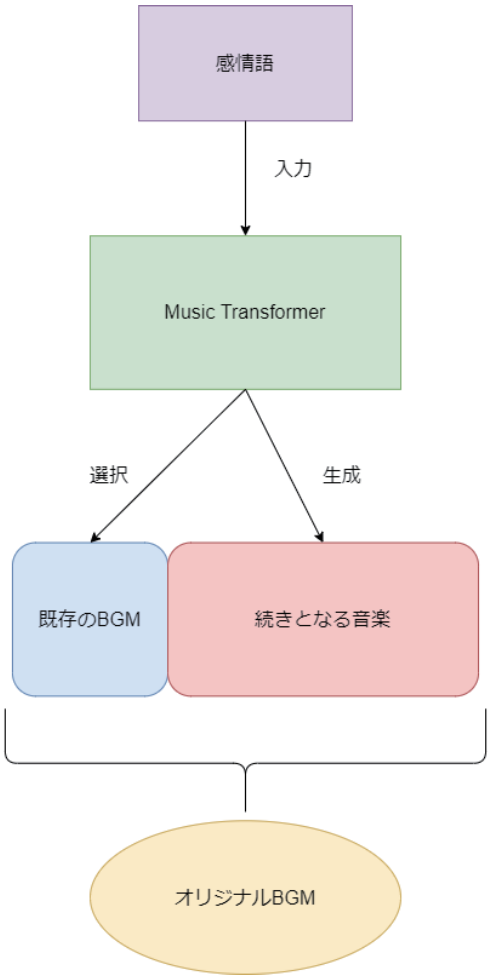

図 2 : BGM 生成の工程図

4.1 生成楽曲の評価実験

各モデルから生成された楽曲が期待する感情を含んでいるか確認するために評価実験を行った. まず, 各モデルで生成した楽曲 80 曲 (モデルごとに 20 曲) の中から, 私自身が対象者に聴いてもらう楽曲を 16 曲 (モデルごとに 4 曲) 選定した. その後, 対象者 4 名に選定した 16 曲を聴いてもらい, 楽曲が持つ感情を評価してもらった. また, 対象者 4 名は全員 20 代の男子大学生である.

楽曲の評価実験には, 対象者が楽曲の持つ感情を定量的に評価する指標が必要である. 今回は評価指標として, 曲に含まれる各感情 (楽しい, 穏やか, 悲しい, 恐ろしい) に 1〜4 の 4 段階の値を設定した. 設定した値が 4 に近いほど楽曲は対象の感情を含んでおり (強く感じる), 1 に近いほど対象の感情を含んでいない (あまり感じない) ものとした. この評価指標を基に対象者に楽曲を聴かせ, 各感情の値を決定してもらった. このようにして, 楽曲にどの感情が多く含まれているか, 各感情がどの程度含まれているのかを評価した.

4.2 動画 BGM としての評価実験

生成された楽曲が BGM として動画に適しているか確認するために, 動画と生成楽曲を組み合わせて評価実験を行った. こちらの実験も, 4.1 節の評価実験と同様の人物に対象者とした.

まず, 対象者 4 名に音が付いていない動画 4 本を視てもらい, 動画から感じた感情を評価してもらった. 今回の評価実験では, フリー素材配布サイト Pixabay[7] の動画を使用した. 動画の評価指標は 4.2 節で説明したのと同様に, 各感情 (楽しい, 穏やか, 悲しい, 恐ろしい) に 1〜4 の 4 段階の値を設定した. 対象者に値を決定してもらうことで, どの感情を多く含む動画か評価した.

その後, 4.1 節の評価実験で感情値が高かった楽曲を感情ごとに 1 曲ずつ (計 4 曲) を選定し, 評価してもらった動画 4 本に BGM として組み合わせた. つまり, 動画 4 本と楽曲 4 曲を組み合わせることで, BGM 付きの動画を合計 16 本分作成した (図 3).

そして, 作成した動画 16 本を対象者 4 名に視聴してもらい, 各動画に対してどの感情を多く含んだ楽曲が一番適していたか評価してもらった.

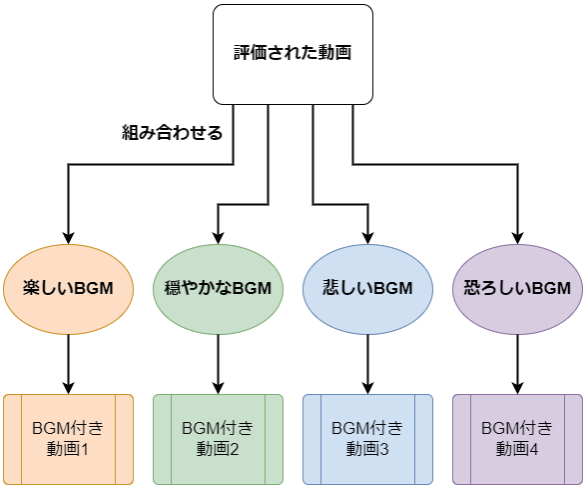

図 3 : ある評価された動画に対する BGM の組み合わせ

5. 評価結果

本章では, 4 章で説明した 2 種類の評価実験の結果を示す. 5.1 節では, 生成楽曲の評価実験結果, 5.2 節では, 生成楽曲を動画 BGM として使用した際の評価実験結果を示す.

5.1 生成楽曲の評価結果

各モデル (FUN モデル, RELAX モデル, SAD モデル, FEAR モデル) から生成した楽曲の評価平均を表 1〜4 に示す. また, 表の数値は楽曲の評価実験で対象者 4 名

が評価した各感情の数値 (1〜4) の平均を取ったものである.

RELAX モデルからは穏やかな楽曲が生成されることを期待しており, 期待通り生成された楽曲の多くが穏やかに感じると評価された (表 2). SAD モデルからは悲しい楽曲が生成されることを期待しており, 期待通り生成された楽曲の多くが悲しく感じると評価された (表 1). FEAR モデルからは恐ろしい楽曲が生成されることを期待しており, 期待通り生成された楽曲の多くが恐ろしく感じると評価された (表 4). これらの結果から, ほとんどのモデルからそれぞれに期待する感情を多く持っている楽曲が生成されていたと言える. しかし, FUN モデルからは楽しい楽曲が生成されることを期待していたが, 生成された楽曲の多くが穏やかに感じると評価された (表 1).

表 1 : FUN モデル生成楽曲の評価平均

曲	楽しい	穏やか	悲しい	恐ろしい
Acc1	2.75	3.5	1.25	1
Acc2	2.25	3.25	1.75	1
Loss1	3.5	2	1	1
Loss2	1.75	3.5	1.75	1

表 2 : RELAX モデル生成楽曲の評価平均

曲	楽しい	穏やか	悲しい	恐ろしい
Acc1	3.5	1.25	1	1.75
Acc2	1	3	2.5	1.25
Loss1	1.25	2.5	2.25	1.5
Loss2	2.5	3.25	1.75	1

表 3 : SAD モデル生成楽曲の評価平均

曲	楽しい	穏やか	悲しい	恐ろしい
Acc1	1.25	2.75	3.5	1
Acc2	1	1.5	3	1.25
Loss1	1	2.5	3.5	1
Loss2	1.25	3	2.25	1

表 4 : FEAR モデル生成楽曲の評価平均

曲	楽しい	穏やか	悲しい	恐ろしい
Acc1	1	1.25	2.25	3.25
Acc2	1	1.75	2.75	2.25
Loss1	1.75	2.75	1.5	1.75
Loss2	1	1	1.5	2.75

5.2 動画 BGM としての評価結果

対象者に映像のみで動画を評価してもらった結果、動画 1, 2 は明るい (楽しい, 穏やかな) 感情, 動画 3, 4 は暗い (悲しい, 恐ろしい) 感情を持っているという評価が得られた. 評価してもらった動画と合わせた際の各 BGM の適合率を表 5 に示す. また, 表の数値は対象者 4 名が各動画に対して, 最も適していると評価した BGM の割合を示している.

暗い感情を持った動画に対しては, 恐ろしい BGM が最も高い適合率が得られた. しかし, 悲しい BGM は適していないと評価された. 反対に, 明るい感情を持った動画に対しては, 悲しい BGM が最も適していると評価された. また, 全体的に楽しい BGM や穏やかな BGM は適合率が低い結果となった. これらの結果から, 動画に対して適切な BGM を付与できるように多くの改善が必要である.

表 5 : 動画に対する楽曲の平均適合率

	動画 1	動画 2	動画 3	動画 4
BGM1(楽しい)	25%	50%	0%	25%
BGM2(穏やか)	25%	0%	25%	25%
BGM3(悲しい)	50%	50%	0%	0%
BGM4(恐ろしい)	0%	0%	75%	50%

6. 考察

本章では Transformer モデルを用いて生成した楽曲の評価実験の考察を述べる. 6.1 節では生成楽曲単体の評価実験の考察, 6.2 節では動画 BGM として使用した際の評価実験の考察を述べる.

6.1 生成楽曲の評価実験考察

表 2, 3, 4 より, RELAX モデル, SAD モデル, FEAR モデルではそれぞれのモデルに期待する感情を含んだ楽曲が多く生成されているという結果が得られた. このことから, 作成した 4 種類のモデルのうち, 3 種類 (RELAX モデル, SAD モデル, FEAR モデル) は各感情を持つ楽曲生成に適したモデルであったと言える. それと同時にこれらのモデルに学習させたデータセット (楽曲) は, それぞれの感情 (穏やか, 悲しい, 恐ろしい) を持つ楽曲生成に適していたと言えるだろう.

しかし表 1 より, FUN モデルからは期待した感情と異なる感情を多く含んだ楽曲が生成された. また, RELAX モデル, SAD モデル, FEAR モデルからも少数だが期待する感情以外の感情評価が高い楽曲が生成された. これらの結果から各学習データセットの内容の見直しが必要であると考える. 今回は私自身の判断でそれぞれ

のデータセットに使用する楽曲を選定したが, 自分以外の人間にもデータセットに使用する楽曲を評価してもらうべきだと考える. 多くの人間に楽曲の持つ感情を評価してもらうことで学習に用いる楽曲の一般的な評価を得ることができ, より大勢の感情とマッチする学習データセットの作成に繋がるだろう. 学習データセットの改良を行うことで, 学習させたモデルから, より期待する感情を多く含んだ楽曲が生成されるようになると考える.

また, 生成した楽曲の中には, 終盤のメロディーが音楽として聴きづらいものがいくつか存在し, 楽曲の完成度に差が見られた. このような楽曲が生成された原因として, 学習する楽曲が少なかったことが考えられる. 今回は一つのモデルに学習させた楽曲数は 20 曲だったが, 学習させる楽曲数を増やすことで, より完成度の高いオリジナル BGM の生成が可能になると考える.

6.2 動画 BGM としての評価実験考察

表 5 より, 恐ろしい BGM が動画 3 では平均適合率 75 %, 動画 4 では 50 %と, 暗い感情を持った動画に対して最も高い適合率が得られた. また, 楽しい BGM は動画 2 において平均適合率 50 %という数値が得られた. これらの結果から, FEAR モデルからは動画に適した恐ろしい BGM が生成されていると言える. FUN モデルに関しても, 今回の実験では動画に適した楽しい BGM が生成できたと考える.

しかし, 悲しい BGM は動画 3, 4 ともに平均適合率 0 %と, 暗い感情を持った動画に対して適合率が低いという結果になった. 反対に動画 1, 2 では平均適合率 50 %と, 明るい感情を持った動画に対して適合率が高い結果が得られた. このような結果となった理由として, 楽曲が悲しい以外の感情を多く含んでいた可能性が考えられる. 楽曲の評価実験では悲しい感情を最も多く含んでいるという結果だったが, 悲しい以外の感情も多く含んでいたことによって, 動画と組み合わせた際の評価実験では対象者の評価が変化したと考えられる. また, 楽しい BGM, 穏やかな BGM は多くの動画に対して適合率が低い結果となった. この結果から, これらの楽曲には感情の音楽的特徴があまり含まれていなかった可能性がある. 音楽的特徴があまり含まれていないことにより, 対象者の印象に残りにくく, 他の楽曲の方が適していると判断されたと考えられる. ただし, これらの問題は 6.1 節で述べた学習データセットの改良によって解決できる可能性が高いだろう.

また, 今回は 20 代の男子大学生 4 名に対象者となってもらい, 評価実験を行った. しかし, 動画サイトや SNS では中学生や高校生, 30 歳以上の人々, 女性の方も動画を投稿している. そのため, 性別や年齢の異なる人々を対象者として評価実験を行うことで, より一般的な評価結果を得られると考える. 加えて, 今回は対象者の結果全体を平均して評価を行ったが, 対象者の個々の結果に注目して評価することも必要だと考えている.

7. 結言

本研究では, Transformer モデルを用いて 4 種類の感情語 (楽しい, 穏やか, 悲しい, 恐ろしい) から BGM を生成し, 生成された楽曲に対して評価実験を行うことで動画に適した BGM の生成ができているか判定した. 今回は動画に含まれる感情を基に生成を行うため, データセットを感情ごとの楽曲に分けて 4 種類作成した. 楽曲の学習, 生成には Music Transformer モデルを使用し, 4 種類のデータセットをそれぞれ学習させることで 4 種類のモデル (FUN モデル, RELAX モデル, SAD モデル, FEAR モデル) を作成した. 各モデルに感情語を入力し, 既存楽曲の冒頭数秒に続くメロディーを生成したものがオリジナル楽曲として出力された. 楽曲は各モデルで 20 曲 (合計 80 曲) 生成を行った.

生成された楽曲に対しては, 楽曲単体の評価実験と動画 BGM としての評価実験を行った. 楽曲単体の評価実験では, 対象者 4 名に生成した 80 曲の中から選定した 16 曲を聴いてもらい, 評価してもらった. 評価指標には曲に含まれる各感情に 1〜4 の 4 段階の値を設定した. 値が 4 に近いほど感情を含んでおり, 1 に近いほど感情を含んでいないものとし, 対象者には各感情の値を決定してもらうことで, 楽曲にどの感情が多く含まれているかを評価した. 評価結果より, 作成した 4 種類のモデルのうち, 3 種類 (RELAX モデル, SAD モデル, FEAR モデル) は各感情を持つ楽曲が生成された. しかし, FUN モデルからは期待した感情と異なる感情を多く含んだ楽曲が生成された.

動画 BGM としての評価実験では, 対象者 4 名に音の無い動画 4 本を視てもらい, 動画の持つ感情を決定した. その後, 評価してもらった動画 4 本と各感情を含んだ楽曲 4 曲を組み合わせて BGM 付き動画を 16 本作製した. そして, 作製した 16 本の動画を対象者に視聴してもらい, 動画に対してどの BGM が最も適しているか評価してもらった. 評価結果より, 恐ろしい BGM は暗い感情を持った動画に対して最も高い適合率が得られた. しかし, 悲しい BGM は暗い感情を持った動画に対して適合率が低く, 反対に明るい感情を持った動画に対して適合率が高い結果が得られた. また, 楽しい BGM, 穏やかな BGM は多くの動画に対して適合率が低い結果となった.

本研究では, 感情を基にして動画に適した BGM を自動生成するシステムの開発を研究目的とした. 評価実験の結果から, 作成した一部のモデルからは期待した感情を含み動画に適した楽曲が生成されたが, 全てのモデルからは動画に適した楽曲が生成できなかったので, 学習楽曲の再評価や楽曲数の増加など, 学習データセットの改良が必要であると考える. また, 評価実験の対象者の年齢層の拡大や対象者個々の結果に対して評価を行うことで, より一般的な評価を得ることが必要だと考える.

また, 今回は感情ごとにデータセットを作成し, 学習させた 4 種類のモデルから楽曲を生成した. しかし, 学習させていたデータセットを一つにまとめてモデルに学習させ, 生成した楽曲の評価を行うことで, どちらの形態が研究目的である動画に適した BGM を自動生成するシステムに適切か比較する必要がある.

参考文献

[1] ”YouTube をめぐる 16 の統計データ”, https://www.infocubic.co.jp/blog/archives/15518/(2022/4/11)

[2] Oxford Economics, ”YouTube Impact Report：2021 年 日本における YouTube の経済的・社会的・文化的影響”, https://www.oxfordeconomics.com/resource/a-platform-for-japanese-opportunity-assessing-the-economic-societal-and-cultural-impact-of-youtube-in-japan-in-2021-jp/(2022/4/11)

[3] Business of Apps, ”TikTok Revenue and Usage Statistics (2022)”, https://www.businessofapps.com/data/tik-tok-statistics/(2022/4/11)

[4] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, … Douglas Eck, ”Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset”, In ICLR(2019)

[5] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, … Douglas Eck, ”Music Transformer: Generating Music with Long-Term Structure”, In ICLR(2019)

[6] ”無料のオンラインファイルコンバータ-AnyConv”, https://anyconv.com/ja/(2022/6/18)

[7] ”高品質なフリー動画素材-Pixabay”, https://pixabay.com/ja/(2022/6/20)

AR 技術を用いたオープンキャンパス支援システムにおけるコンテンツ作成支援システムの開発

岸本 雄氏*　黒岩 丈介**　小高 知宏**　諏訪 いずみ***　白井 治彦****

Development of Content Creation Support System for Open Campus Support System Using AR Technology

Yuji KISHIMOTO*,　Jousuke KUROIWA**, Tomohiro ODAKA**
Izumi SUWA*** and Haruhiko SHIRAI****

(Received September 30, 2022)

In this paper, we investigate an open campus support system, where provide three functions, a campus tour guidance function, a laboratory tour guidance function and a laboratory introduction function. However, it has taken too much cost to develop laboratory introduction contents in realizing the laboratory introduction function. Therefore, the purpose of this paper is to construct a content creation support system with user friend operating functions, which reduces efforts in developing the contents. We have constructed the support system and performed evaluation experiment operability assessment. We obtain positive feedback regarding with the reduction of content creation time and the operability.

Key words :*Support system, AR*

1. はじめに

これまで，オープンキャンパスなどで研究室を訪れた学生に対し，研究室紹介を教授や研究室の学生が口頭説明によって行われてきた．しかし，教員や研究室の学生が不在の場合，直接情報を得ることが出来ないため，紙媒体による情報発信を行ってきた．紙媒体による情報発信では，提供する情報が限定されてしまうことが問題となっていた．

*大学院工学研究科 知識社会基礎工学専攻
*Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering
**知能システム工学講座
**Department of Human and Artificial Intelligent Systems
***仁愛女子短期大学 生活科学学科
***Jin-ai Women's College
****工学部 技術部
****Technical Division

そこで，観光案内や教育などの分野で活用されているAR技術に着目した．ARとは「Augmented Reality」の略称で，主に拡張現実と訳されている．ARの定義としてはAzuma氏が提案したもので，現実と仮想の組み合わせであること，実時間で動作する応答性を備えていること，三次元的に整合性が取れていることとされている[1]．また，特定の出力機器に限定するものではなく，視覚メディアに限定するものでもなく，技術的に実現は難しいが，音声，触覚，嗅覚，味覚に関するARもこの範疇に含まれている．

観光分野では情報提供システムの先行研究として，画像認識型AR技術を用いて観光情報を提供するシステムの研究が行われている[2]．教育の分野ではAR技術を活用する事例として，ARを用いてタブレットPCで教科書の特定のページを認識すると，動画教材が再生されるシステムの構築を行う研究も行われている[3]．

我々はこれまでに，観光情報を提供する研究から，

AR 技術を大学構内の施設や訪れた研究室の情報を表示する機能に応用することが出来，屋外や屋内の案内をスマートフォンで行えるようにすることが出来るのではないかと考えた [4]．また，教育分野の研究を特定のマーカーを研究内容を紹介する動画を表示するという機能にして実装することで，研究室紹介などの口頭で行っていた情報提供が出来るようになり，負担軽減につながると考えた，それに加えて，人は AR 上の キャラクターも人間だと感じてしまうという先行研究があり，教員や学生の代わりに研究の内容を紹介する 3D モデルキャラクターを動画とともに配置することで，より効率よく情報を得ることが出来る機能を有したシステムの検討を行った [5]．さらに，AR 技術を利用したオープンキャンパス支援システムの開発を行っており，各研究室ごとに異なるマーカーを配置し，読み取ったマーカーに対応した研究室の研究内容を動画を用いて紹介するという機能を開発していた．しかし，先行研究の問題点として，研究紹介動画の作成やのキャラクターの動作作成，キャラクターの動作を行う時間の指定などがコンテンツ作成者への負担が大きいという点があった．

そこで本研究では，研究室紹介機能において使用する研究紹介動画の作成を支援するシステムの構築を行う．合成音声の作成や動画編集などの研究紹介動画を作成する際に負担となっていた作業を可能な限り簡単に作成することが出来るようにする必要がある．また，3D モデルのキャラクターに研究の内容等を紹介させる際，動作を指示するスクリプトの作成を手入力で行っており，これもまた負担となっていた．そのため，動作を指示するスクリプトの作成を簡易化するシステムを構築することが必要であると考えた．以上より，本研究の目的は，研究紹介動画の作成に必要なコンテンツの作成を支援する研究室紹介動画作成支援システムを構築することである．

2. AR 機能を有するオープンキャンパス支援システム

2.1 システムの設計

現状のオープンキャンパスのやり方として，屋外・屋内の建物案内は来場者に紙媒体の地図を配り，学問説明や研究室紹介でも同様に紙媒体の資料を配布した上で教授や研究室の学生による口頭説明を行っていた．しかし，教員や研究室の学生が不在の場合，直接情報を得ることが出来ないため，紙媒体のみの情報発信となり，提供する情報が限定されてしまうことが問題となっていた．

そこで先行研究では，屋外では地図の表示や GPS を利用した目的の建物や施設までのルート案内や目印を表示，屋内では研究室などの目的の部屋への案内や情報を表示し，研究室紹介では教授や学生が行ってきた研究内容の紹介を代わりに行うようなシステムの開発を検討していた．研究室紹介機能では，AR マーカーを読み取ることで，訪れた研究室を判定し，3D モデルのキャラクターである「説明者アバター」に研究室や研究の内容を説明させていた．研究内容の説明の際には，学会発表等で用いた Power Point スライドへ合成音声をつけた 「スライド動画」 という動画を作成した．説明者アバターにスライド動画へポインティングさせながら紹介することで，研究紹介機能として実装した．

2.2 システムの問題点

研究紹介システムの設計，構築を行ってきた．本システムの実装をするにあたって，先行研究では 4 つの研究紹介動画を作成したが，音声合成ソフトを用いて合成音声を 69 個作成するのに 1 週間，動画編集ソフトを用いてスライド動画を作成するのに 4 週間かかっていた．そのため，本システムを利用する研究室が増えるにつれて，コンテンツの制作者への負担が大きくなっていくことが想定される．そこで，この問題を解決する方法を考えていく．スライド動画を動画編集ソフトで作成していたが，PowerPoint のスライドを使用する場合，そもそも PowerPoint に動画を出力する機能があるため，これを利用すれば複雑な編集は不要であると考えられる．この機能では，設定したアニメーションを実行した上で動画として出力するため，スライドに作成した合成音声をアニメーションとして埋め込むことで，合成音声とスライド動画を動画編集ソフトで纏める作業も短縮することが出来る．次に，合成音声の作成についてである．PowerPoint をはじめとする Microsoft office には文字を読み上げる機能があり，PowerPoint のノートに書かれている内容も読み上げることが出来る．合成音声の作成には学会等の発表で利用した読み原稿をもとに作成しており，発表で用いる PowerPoint ファイルはノートに読み原稿を書き込んで，練習を行っていることが想定される．ノートに書き込まれている内容を音声読み上げ機能を利用し，何らかの方法で音声をスライドに埋め込むことが出来れば，動画編集ソフトを殆ど用いずにスライド動画を作成することが出来ると考えた．また，説明者アバターの動作を指示するスクリプトを，手入力で行っていることもまた，コンテンツの作成者の負担になることが分かってい

る．そこで，直感的に操作でき，かつ簡単に動作指示を行うスクリプトの作成が出来るようなシステムを構築することで問題を解決できると考えた．

3. AR 動画作成支援システムの設計

研究室紹介システムにおいて，スライド動画やアバターの動作作成などのコンテンツの作成に多大な労力が必要となり，製作者への負担が大きいという問題点を抱えていた．本システムでは，AR 技術をもちいた研究室紹介機能によって利用されるスライド動画の作成とシナリオがいるの作成を支援し，コンテンツ作成者の負担を軽減するシステムの構築を目指す．

3.1 シナリオファイル作成機能

先行研究では，研究紹介動画を作成する際に，スライド動画のどの位置にポインティングするかという動作を指定するシナリオファイルをスライド動画を見ながら手入力で行っており，負担となっていた．この機能は，その負担を軽減するためのものである．スライド動画をアプリケーション上で再生し，スライド動画のポインティングしたい位置をクリックすることによって，座標を取得し，説明者アバターに動作を行わせる．クリックした際に，現在のスライド動画の再生時間，実行した動作の種類，クリックした座標をシナリオファイルに格納し，出力する．これによって，手入力することなく，動画を見ながら直感的な操作でシナリオファイルを作成することが出来る．

3.2 スライド動画作成支援機能

研究を紹介するスライド動画の作成には合成音声の作成，動画の編集など製作者への負担が大きいという問題があった．スライド動画は，PowerPoint ファイルのスライドを使用しており，合成音声はノート等に書かれている読み原稿を外部の合成音声ソフトに入力することで作成していた．この機能は，その負担を軽減するためのものである．PowerPoint の機能は合成音声によってノートの内容を読み上げる機能が存在しており，合成音声をスライドに埋め込むというアドインを作成することで，合成音声を作成する風単の軽減を行った．また，PowerPoint の機能には動画として出力する機能も存在しており，アニメーションを実行した上で，スライドショー形式で動画を出力するものである．埋め込まれた合成音声はアニメーションとして扱われているため，スライドに合わせて研究を紹介する動画が作成できる．

3.3 システムの構成

本システムでは，大きく分けて説明者アバターの動作指示ファイルであるシナリオファイル作成を行う機能，スライド動画作成支援機能の 2 つに分けられる．シナリオファイル作成機能は動画表示機能，動画接触判定による配列格納機能，説明者アバターの動作生成機能，再生時間巻き戻し機能，シナリオファイル出力機能によって構成されており，スライド動画作成支援機能はアドイン追加機能，音声埋め込み機能，動画出力機能によって構成されている．

3.4 システムの設計

シナリオファイルの作成機能とスライド動画作成支援機能を構成する各機能についての設計を以下に述べる．

3.5 システムの概要

3.5.1 シナリオファイル作成機能

シナリオファイル作成機能の構成を図 1 に示す．研究の内容を紹介する際に，スライド動画に合わせて説明者アバターにポインティング等の動作をさせる必要がある．先行研究では，手入力で動作のタイミングを指定しており，負担となっていた．シナリオファイル作成機能では，動画表示機能によって表示されたスライド動画を閲覧し，動画接触判定によるシナリオファイル作成機能によって直感的に動作の指定をすることが出来る．指示を出す際にアバターの動作生成機能によって，研究室紹介システムでどのように説明者

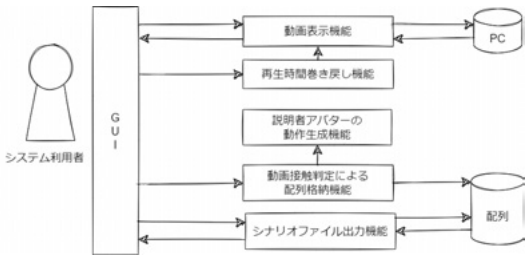

図 1 シナリオファイル作成支援機能の構成

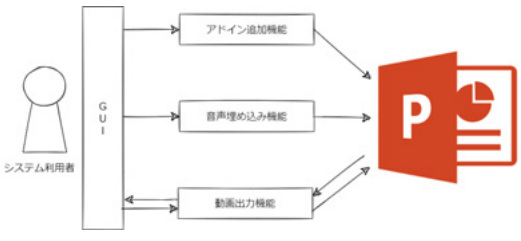

図 2 スライド動画作成支援機能の構成

アバターが動作するかを視覚的に把握することが出来る．動画全体で作成した後に，CSV 形式のファイルとしてシナリオファイルを出力することが出来る．

3.5.2 スライド動画作成支援機能

スライド動画作成支援機能の構成を図 3 に示す．スライド動画は PowerPoint ファイルのスライドを使用しており，合成音声はノートなどに書かれている読み原稿を外部の合成音声作成ソフトに入力することで作成していた．この機能は，その負担を軽減するためのものである．PowerPoint の機能には合成音声によってノートの内容を読み上げる機能が存在している．この機能を利用して音声ファイルを作成し，スライドに埋め込むという音声埋め込み機能を有したアドインを作成することで，スライド動画全体の合成音声を作成することが出来る．アドイン追加機能によって，複雑な設定をすることなくアドインを追加することが可能となっており，実行後は 「アドイン」 という項目を選択後，追加されたアドインのアイコンをクリックするだけで音声の埋め込みが実行される．PowerPoint の機能には作成したスライドを動画として出力する機能も存在しており，前述した機能と併用することで簡単にスライド動画を作成することが出来る．

3.6 システムの実装

各機能において実装した機能を下記に示す．

3.6.1 シナリオファイル作成支援機能

- 動画表示機能
  Unity に動画を取り込むことなく，ファイルパスを用いて動画を再生する機能である．また，ファイルダイアログからファイルを読み込む機能を実装しているため，ファイルパスを手入力しなくとも PC 内の動画ファイルをマウスで選択するだけで動画のファイルパスを入力することも可能である．この機能によって，シナリオファイル作成の際に，細かい設定をすることなく Unity 内での動画の表示され方を確認することが出来る．

- 動画接触判定による配列格納機能
  動画表示機能によって表示された動画をクリックすることで，クリックした位置の座標を取得する．クリックした座標に応じて動画の上段，中段，下段に分けられ，更に動画の右半分をクリックした際にはポインターを表示し，動画の再生時間とともに動作を指示するシナリオファイルへと出力するための配列に格納される．この機能によって，スライド動画のポインティングしてほしい位置を動画を見ながら直感的に指定することが出来る．

- 説明者アバターの動作生成機能
  動画に接触判定をつける機能を利用してポインティングしたい位置を指定した際，ポインティングした座標に対応した動作を行う機能である．また，シナリオファイルの情報を読み込み，再生している動画の時間とシナリオファイルに格納されている時間が一致した際に，特定の動作を行うことが出来る．

- 再生時間巻き戻し機能
  動画を視聴しながら作成する都合上，動作をしてほしい個所を見逃してシナリオファイルを作成してしまう可能性がある．そこで本機能では，現在の再生時間から 5 秒巻き戻す機能を実装した．

- シナリオファイル出力機能
  格納した配列からシナリオファイルを出力する機能である．再生時間巻き戻し機能を使用して作成されたシナリオファイルは，格納した動画の時間が昇順になっておらず，順番が不規則となっている．この状態のままシナリオファイルの読み込みを行うと再生時間が一致しなくなり，動作を行わなくなってしまう．そのため，この機能にはシナリオファイルを出力する際に，格納されたスライド動画の時間を昇順にソートし，かつ動作の種類も連動してソートする機能も含まれている．

3.6.2 スライド動画作成支援機能

- アドイン追加機能
  PowerPoint ファイルにアドインを追加する機能である．PowerPoint にアドインを追加する際には,「ファイル」→「オプション」→「アドイン」 を選択し,「管理」→「PowerPoint アドイン」を選択する．その後,「アドイン」 のダイアログボックスから 「新規追加」 を選択.「新しい PowerPoint アドインの追加」 から追加するアドインを参照するという手順が必要であるが，本機能では，その手順を省略することが出来る．

- 音声埋め込み機能
  研究の内容を紹介するスライド動画の作成において，使用される PowerPoint ファイルのノートの内容を音声ファイルとして自動生成し，スライドに埋め込む機能である．PowerPoint の機能の 1 つである動画を出力する機能を利用することによって，音声を手動で作成することもなく，最小限の動画編集のみでスライド動画を作成することが出来る．
- 動画出力機能
  PowerPoint スライドを動画として出力する機能である．スライドに設定されたアニメーションを実行した上で，動画として出力する．PowerPoint に備わっている機能である．

4. 評価実験

実装した研究紹介 AR 動画作成支援システムを用いたスライド動画の制作時間，及びアンケート評価実験を行った．本章では，実験目的と方法，結果について述べる．

4.1 スライド動画制作時間の比較による評価実験

4.1.1 実験の目的と方法

今回実装した AR 動画作成支援システムでは，先行研究の研究紹介システムのコンテンツ作成の負担が大きいという問題点を解決することを目的として作成したシステムである．そのため，目的の達成度合いの確認を目的とした評価実験を行った．実験方法としては，男子大学生 4 名 (平均年齢 22.75 歳) に本システムを利用させ，先行研究における研究紹介動画の作成手法と本システムを利用した際の紹介動画制作時間を比較することで評価を行う．

4.1.2 実験結果

先行研究の手法では 7 分 44 秒の動画に対して，シナリオファイルの作成に約 4 時間かかっていた．また，スライド動画の作成には 9 日間かかっていた．本システム利用時にはシナリオファイルの作成，スライド動画の作成それぞれ約 30 分で行うことが出来た．

4.2 アンケート評価実験

4.2.1 実験の目的と方法

次に AR 動画作成支援システムの改善点を知ることを目的とした評価実験を行った．実験方法としては，男子大学生 4 名 (平均年齢 22.75 歳) を対象に以下の手順で行った．

1. システムの使い方を口頭で説明する．
2. 被験者にシステムを利用させ，スライド動画作成支援機能を活用し，スライド動画の作成を行ってもらう．
3. シナリオファイル作成機能を利用し，シナリオファイルの作成を行ってもらう．
4. システム利用後，アンケートに答えてもらう．

アンケートは 2 つの機能を利用した上で，現時点でのシステムの有用性を測る 5 つの設問を用意した．以下に設問の内容を示す．

1. 本システムの操作方法は分かりやすいと思うか．
2. 研究を紹介するにあたって，説明者アバターの動きは適切であったと思うか．
3. 思った通りにポインターは動いたと思うか．
4. 本システムは研究紹介動画を作成する手法として成り立っていると思うか．
5. 本システムで気になった点や改善点が何かあれば教えてください．

上記の設問 1 から 5 について，とても思う，思う，どちらともいえない，あまり思わない，思わない，の 5 段階リッカート尺度で評価してもらい，5 番目の設問については自由記述とした．

4.2.2 実験結果

システムの操作方法は分かりやすいかという，直感的な操作を行うことが出来るかを把握する設問に対して，図 3 のようにとても思う，思うという評価をそれぞれ 2 人が回答していた．説明者アバターの動きは適切であったかという設問に対しては，図 4 のようにとても思う，思う，どちらともいえない，あまり思わないの 4 つの回答に分かれた．思った通りにポインターは動いたかという設問に対しては，図 5 のように全員からどちらともいえないという評価を得た．研究紹介動画を作成する手法として成り立っているかという設問に対しては，図 6 のように全員から思うという評価を得た．5 つ目の設問については，以下のような回答を得た．

- システムの操作をした際に，ユーザー側に実行した旨を伝える機能が欲しい．
- 説明者アバターの動作を実行するまでのレスポンスが長い．
- 説明者アバターの動作がポインティングしたい位置とずれている点が気になる．
- ポインターを表示するだけでなく，左右移動や円運動などの動きが欲しい．

4.3 評価実験の考察

4.3.1 制作時間の比較における評価実験についての考察

実験結果から，研究紹介 AR 動画の制作時間を大幅に短縮することが出来た．スライド動画作成支援機能に実装した PowerPoint ファイルにアドインを追加する機能を用いて，追加した合成音声作成のアドインを実行するだけで合成音声を作成することが出来るというものであった．これによって，これまで使用していた音声合成ソフトを用いるという手間を省くことが出来たという点が時間短縮につながったと考えられる．また，30 分から更に短縮することは，合成音声の誤読の確認・修正に多少時間がかかるため，難しいと考えた．良い結果を得ることが出来たが，一部の PowerPoint ファイルにおいてアドインを実行することが出来ないという不具合が見つかっている．また，現在のアドインでは PowerPoint ファイル内全てのスライドに対して音声を作成し，埋め込みを行っている．スライドの枚数によっては実行完了までに時間がかかってしまうという問題点が考えられる．今後の課題としては，アドインを実行できない不具合の原因究明と対応，現在表示しているスライドにのみ音声を埋め込むアドインを新規で作成することである．

4.3.2 アンケート評価実験についての考察

AR 動画作成支援システムの操作方法について分かりやすいと思うかという設問については，とても思う，思うという肯定的な評価を得ることが出来た．この結果からシステムを利用する上でのユーザーインターフェースには問題が無いという事が考えられる．その反面，自由記述の設問では，システムを操作した際にユーザー側へ実行した旨を伝える機能が欲しいという改善点の指摘があった．現在のシステムには操作の手順を案内するような機能が搭載されてお

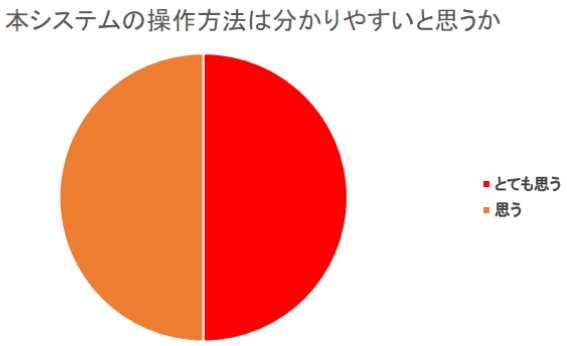

図 3 質問 1 の結果

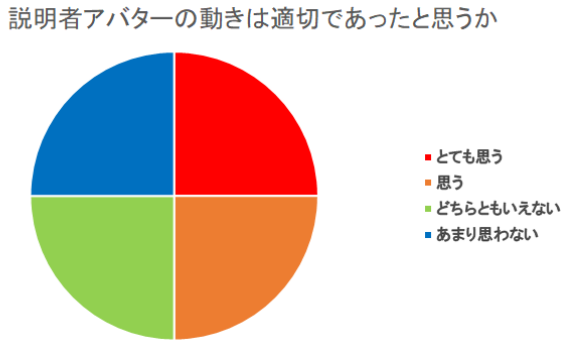

図 4 質問 2 の結果

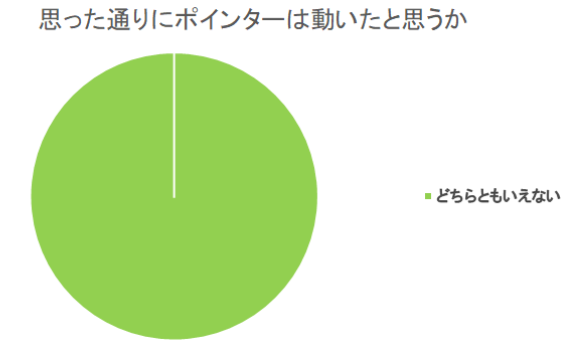

図 5 質問 3 の結果

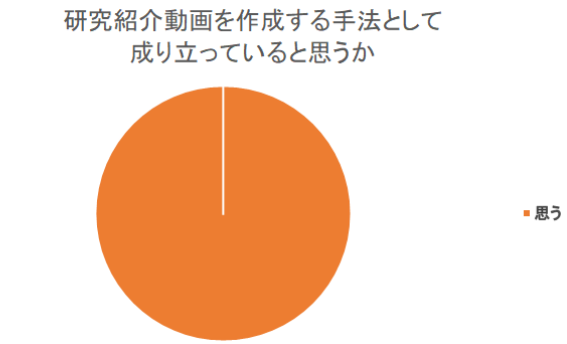

図 6 質問 4 の結果

らず，シナリオファイル作成支援機能においては，シナリオファイル作成機能を正しく実行できているかどうか視覚的に分かりにくいという点が考えられる．そのため，実行する機能を実行した際には，メッセージウインドウ等で現在実行した内容を文字や音声で案内する機能の追加が必要であると考えた．

説明者アバターの動きは適切であったかという設問に対しては，とても思う，思う，どちらともいえない，あまり思わないのそれぞれに 1 人ずつ回答していた．自由記述における回答では，説明者アバターの動作を実行するまでのレスポンスが長い，説明者アバターの動作がポインティングしたい位置とずれているという点が挙げられていた．実行するまでのレスポンスが長いことから，ポインティングのタイミングがずれてしまうという問題点だけでなく，慣れるまでユーザーの思い通りに動かないというストレスを与えてしまうという問題点がある．また，説明者アバターの動作とポインティングしたい位置のずれについては，現在のシステムではスライドの上段，中段，下段へ腕を動かすという 3 つしか動作が無いことによるものであると考える．今後の課題としては，説明者アバターが動作を行うまでの時間を短縮，動作の種類に上段と中段，中段と下段の中間の位置を指す動作を追加することである．

思った通りにポインターは動いたかという設問に対しては，全員からどちらともいえないという回答を得た．これは自由記述におけるポインターに動きが欲しいという点だけでなく，ポインターを表示する際の仕様も関係していると考えられる．3.2.1 項でも述べたように，動画接触判定による配列格納機能の 1 つであるポインターは，動画表示機能によって表示された動画の右半分をクリックした時にのみ表示されるというものである．この仕様はポインター自体がスライド動画の注目して欲しい箇所を見せるためである．そのため，説明者アバターの動作だけでポインティングを行うことが出来る左半分はポインターの表示は不要であると考えたためであった．しかし，ユーザー視点ではその仕様が分かりにくく，説明者アバターの動作のレスポンスが遅いということもあり，正しく実行できているか分からず，何度も画面をクリックするといった行動をしている被験者が 2,3 名ほどいた．そのため，動画の左半分をクリックした際にも，ポインターを表示するように変更する必要があると考えた．また，ポインターに左右の移動や円運動などの動きが欲しいという点では，ポインターが動くことによってスライド動画よりもポインターの方へ注目しやすくなってしまうという考えから，ポインターの動作については静止状態のみで良いと考えた．

研究紹介動画を作成する手法として成り立っていると思うかという設問に対しては，被験者全員から思うという評価を得ることが出来た．上述したように，各設問における改善点はあるものの，AR 動画作成支援システムは研究紹介機能において利用されるコンテンツの作成支援システムとして成り立っており，オープンキャンパス支援システムを作成する上での問題点の 1 つを解決することが出来たと考える．

4.4 システム実装についての考察

AR 紹介動画作成支援システムは，先行研究のおいて検討・開発を行っていた AR 技術を用いたオープンキャンパス支援システムの研究紹介機能におけるコンテンツの作成を行う負担を軽減することを目的として実装した．本システムを実装していく中で，コンテンツの負担軽減だけでなく先行研究のシステム自体にも改善点が見つかったため， 先行研究において有用であるとされた機能を踏襲した新たなシステムを構築する必要があると考えた．先行研究では，説明者アバターとスライド動画を動画編集ソフトを用いて研究紹介動画として作成していた．これは，スマートフォンの性能によっては 3D モデルの描写が上手く行えず，本来の挙動と違う動作をすることを懸念してのことであったが，近年のスマートフォンの性能の向上から考慮する必要が無いと考えた．今後の課題としては，先行研究において開発を行っていたシステムの改修である．

5. おわりに

本研究では，研究紹介機能において利用されるコンテンツ作成の負担軽減を目的としたシステムの開発を行った．AR 動画作成支援システムでは，直感的にシナリオファイルを作成することが出来るシナリオファイル作成支援機能と PowerPoint ファイルを利用したスライド動画作成支援機能の 2 つの機能を組み合わせることによって実装した．また，AR 紹介動画作成支援システムに対して，研究紹介動画の制作時間の比較及びアンケートによる評価実験を行った．得られた結果は，以下である．

- 約 9 日かかっていた研究紹介動画の作成時間を 1 時間に短縮することが出来た．
- 直感的な操作は可能であるが，各機能の挙動に改善点がある．
- システム全体の評価としては支援出来ていると

いえる．

制作時間の短縮，評価の観点から支援システムとして十分に成り立っているといえる．今後の課題は，以下である．

- シナリオファイル作成支援機能における，動画をクリックした際に説明者アバターが動作を実行するまでのレスポンスの改善.
- 説明者アバターの動作の種類を追加することによるポインティングの精度向上.
- ポインター表示範囲の修正.
- 先行研究で開発を行っていたオープンキャンパス支援システムの改修.

参考文献

[1] Ronald.T.Azuma:A Survey of Augmented Reality,Presence:Teleoperators and Virtual Environments,4,355-385,(1997).

[2] 深田秀実, 船木達也, 兒玉松男, 宮下直也, 大津晶: 画像認識型 AR 技術を用いた観光情報提供システムの提案, 情報処理学会研究報告書,2011-IS-115 13,1-8.

[3] 初谷拓郎, 岡村拓哉, 伊與田光宏. 拡張現実を用いた授業支援教材における教育効果の検証. 情報科学技術フォーラム,12,457-458(2013).

[4] 池本武史, 黒岩丈介, 小高知宏, 白井治彦, 諏訪いずみ :AR 技術を用いた研究室紹介システムの実現，電気・情報関係学会北陸支部連合大会，F1-2-2(2020).

[5] Social interaction in augmented reality,<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0216290>(2022/9/27).

サポートベクターマシンによるスワイプデータからの個人認証手法の確立

阿部 僚馬*　黒岩 丈介**　小高 知宏**　諏訪 いずみ***　白井 治彦****

Support Vector Machines for Swipe Data Establishment of Personal Authentication Method

Ryoma ABE*,　Jousuke KUROIWA**, Tomohiro ODAKA**
Izumi SUWA*** and Haruhiko SHIRAI****

(Received September 30, 2022)

In recent years, smartphones have not only functioned as cell phones, but have also a variety of other functions, such as electronic payment and Internet banking. Then, they contain a lot of personal information. Android OS smartphones, which have a high market share worldwide, still use pattern lock authentication, but it is poor secure. Therefore, the purpose of this study is to investigate a more secure lock by exploiting the potential features of pattern locks during swiping with support vector machine. In computer experiment, we trained a support vector machine with personal features obtained from five subjects and evaluated the false reject rate and false accept rate. From the results, we have succeeded to provide a robust and practical authentication method based on the swiping actions with support vector machine.

Key words :*pattern lock authentication*, *a support vector machine*

1. はじめに

スマートフォンの普及は進み，老若男女関係なく，1人1台持つことが当たり前の時代となってきた．スマートフォンの高性能化により，電話番号や電話帳のデータだけでなく，銀行の口座番号，ID，パスワード，GPSなどによる位置情報が記憶されている．そのため，スマートフォンの紛失は個人情報の漏洩となりかねない．これを未然に防ぐために，パスコードや生体認証を用いた個人認証方法を用いてロックをかける．

モバイル端末のOSは現在，Google社が提供するAndroid OSとApple社が提供するiOSの2つが99％を占めている．その中でもAndroid OSのシェア率が最も高い．Android OSの端末にはパスワード，PIN，パターンロック，生体認証を用いたロック方法が搭載されている．パターンロック認証は3 x 3の目印点をスワイプする指の軌跡情報から本人を認証する手法であり，簡便ではあるが，認証動作を覗き見されることにより第3者でも容易に不正侵入することができてしまう．[1]

また，実験例として，ディスプレイに付着した油脂を撮影し，解読した結果，68%の精度でパターンの解読が可能であった報告もある．[2] さらに，覗き見による認証率は高く，1回の覗き見で解読可能な確率が64.2%，システムがオートロックされる5回までに解読可能な確率は95%という実験結果も挙げられる．[3]

*大学院工学研究科 知識社会基礎工学専攻
*Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering
**知能システム工学講座
**Department of Human and Artificial Intelligent Systems
***仁愛女子短期大学 生活科学学科
***Jin-ai Women's College
****工学部 技術部
****Technical Division

そのため，様々な個人情報を守る上でも，より強固かつ容易な操作で行える認証手法が求められている．パターンロック認証には，座標データ以外に，スワイプ操作の速度，ディスプレイに対する指の接触面積または圧力など情報も存在する．これらの情報は意識的にスワイプする座標データとは異なり，人が無意識的に発するデータと言える．本研究ではこのようなデータを無意識的特徴と呼ぶ．この無意識的特徴は身体的特徴に近く，模倣が困難であるため，覗き見に対する耐性があると考えられる． そこで先行研究では，スワイプ動作の指の位置情報だけでなく，時間情報や，指の移動速度及び指の接触面積などの情報を個人特徴として注目し，より強固なパターンロック認証手法を確立することを可能とした. [4]

まず，軌跡情報取得システムからユーザーのスワイプ動作の時系列データを取得する．取得したデータから x 軸方向及び y 軸方向の速度，指の接触面積及び時間データを算出し，個人特徴の抽出をした．その際，ノイズを除去するために単純移動平均法を施した．

個人認証実験では被験者 6 名に対して計測を行い，各被験者の特徴量を抽出して独立認証を行う．そして，被験者分類には標準化ユークリッド距離を用いた認証が一番精度が良く，FRR が 18.6% ，FAR が 1.6% だった. [5] しかし，スマートフォンに実用化することを考えたときに，FAR を 0% とすることが理想である．そのため，標準化ユークリッド距離を用いた手法では，実用化には不十分である．

そこで，本研究では分類精度の高いサポートベクターマシンを用いて，認証精度のさらなる向上を目指す.

2. サポートベクターマシン

サポートベクターマシンはニューロンのモデルとして最も単純な線形しきい素子を用いて，2 クラスののパターン識別器を構成する手法である.

2 クラスのパターン分類問題を考えるとき, 特徴ベクトルを $\boldsymbol{x}^T = (x_0, x_1, \cdots, x_d)^T$，パラメータを $\boldsymbol{w}^T = (w_0, w_1, \cdots, w_d)^T$, しきい値を h とすると，線形識別関数は

$$y = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} - h) \tag{1}$$

と表現され，2 値の出力値を計算する．関数 sign(u) は $u > 1$ のとき 1 をとり，$u \leq 0$ のとき -1 をとる符号関数である．訓練サンプル集合が線形分離可能なら

$$t_i(\boldsymbol{w}^T\boldsymbol{x} - h) \geq 1, (i = 1, ..., N) \tag{2}$$

を満たすようなパラメータが存在する．これは 2 枚の超平面で訓練サンプルが分離されていることを表している．このとき，識別平面と超平面のマージンの大きさは $\frac{1}{||\boldsymbol{w}||}$ となる．サポートベクターマシンでは訓練サンプルをなるべく余裕を持って分けるような平面が求められる．そのため，マージンを最大とするパラメータ $\boldsymbol{w}$ と h を求める問題は，制約条件

$$t_i(\boldsymbol{w}^T\boldsymbol{x} - h) \geq 1, (i = 1, ..., N) \tag{3}$$

のもとで，目的関数

$$L(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||^2 \tag{4}$$

を最小とするパラメータを求める問題と等価になる.

2.1 ソフトマージン

実際，パターン認識の分野において線形分離可能な場合は稀である．そのため，非線形のものを分離するには，多少の誤りを許容する必要がある．これはソフトマージンと呼ばれる．ソフトマージンでは，誤ったサンプルと境界平面の距離をパラメータ $\xi_i(\geq 0)$ を用いて $\frac{\xi_i}{||\boldsymbol{w}||}$ と表すと，その和は

$$\sum_{i=1}^{N} \frac{\xi_i}{||\boldsymbol{w}||} \tag{5}$$

なるべく小さいことが望ましい．これからの条件から，最適な識別面を求める問題は，制約条件

$$\xi_i(\geq 0), t_i(\boldsymbol{w}^T\boldsymbol{x} - h) \geq 1 - \xi_i, (i = 1, ..., N) \tag{6}$$

の下で，目的関数

$$L(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}||\boldsymbol{w}||^2 \tag{7}$$

を最小とするパラメータを求める問題に帰着する．パラメータ γ は第 1 項のマージンの大きさと第 2 項のはみ出しの程度とのバランスを決める定数である.

2.2 カーネルトリック

ソフトマージン法を用いることで，非線形の場合でも素子のパラメータを求めることができるようになった．しかし, これによって全ての非線形の複雑な問題に対して，良い性能の識別器を構成できるとは限らない．複雑な非線形の問題に対応する方法として，特徴ベクトルを非線形変換して，その空間で線形の識別を行う方法をカーネルトリックと呼ぶ.

今，元の特徴ベクトル $\boldsymbol{x}$ を非線形の写像 $\phi(\boldsymbol{x})$ によって変換し，その空間で線形識別を行うことを考え

てみる．例えば，写像 ϕ として，入力特徴を 2 次の多項式に変換する写像を用いるとすると，写像した先で線形識別を行うことは，元の空間で 2 次の識別関数を構成することに対応する．一般に，こうした非線形の写像によって変換した特徴空間の次元は非常に大きくなりがちであるが，サポートベクターマシンの場合には，目的関数や識別関数が入力パターンの内積のみに依存した形になっており，内積が計算できれば最適な識別関数を構成することが可能である．もし，非線形に写像した空間での二つの要素 $\phi(\boldsymbol{x}_1)^T$ と $\phi(\boldsymbol{x}_2)$ の内積が

$$\phi(\boldsymbol{x}_1)^T\phi(\boldsymbol{x}_2) = \boldsymbol{K}(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{8}$$

のように，入力特徴 $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ のみから計算できるなら，非線形写像によって変換された特徴空間での特徴 $\phi(\boldsymbol{x}_1)$ や $\phi(\boldsymbol{x}_2)$ を計算する代わりに，$\boldsymbol{K}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ から最適な非線形写像を構成できる．

3. スワイプデータとその特徴

3.1 実験に用いる個人データ

本実験には先行研究で用いられたデータを用いる．右利きの男性 5 名の被験者（ user A, user B, user C, user D, user E ）に動作が安定するまで十分なスワイプ操作を行った後にデータの取得を行い，被験者全員が同様の室内環境，体調，姿勢でデータ取得を行った．実験経路を図 1 に示す．青が目印点を表し，赤色は経路を表す．経路は ABCDACBDCABD の順であり，「Z」→「∞」→「四角」の順にスワイプする．1 人あたり，1 つの経路に対して 20 回分データの取得を行う．

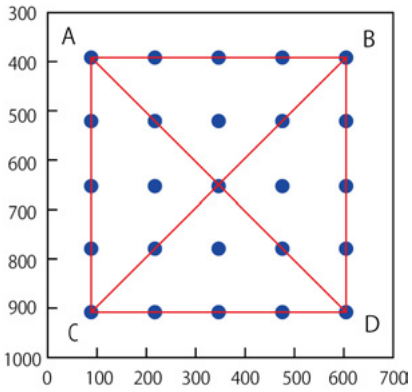

図 1 実験経路

3.2 ノイズ除去

取得した 20 回分のデータから x 軸速度，y 軸速度，接触面積のデータに分ける．取り出した x 軸速度，y 軸速度にはノイズが載っているため，平滑化を行うことでノイズを除去する．ノイズの除去には移動平均法を用いるが，移動平均法の計算には主に 3 種類存在し，単純移動平均法 (SMA)，加重移動平均法 (WMA)，指数移動平均法 (EMA) が挙げられる．先行研究より単純移動平均を用いた平滑化がほかの移動平均法よりも優れていることが分かっているため，今回の実験では単純移動平均法を用いた．ここで，単純移動平均法について説明する．は時系列データにおいてある時刻 k ステップ目を中心とし，その前後 n 個のデータの平均値を求め，その結果を k ステップ目の結果とする．すなわち $2n+1$ 個のデータの平均値が単純移動平均での結果となる．ここで単純移動平均法を以下の式で定義する．k ステップ目のデータを $A(k)$, k ステップ目の演算結果を $S(k)$ とする．ただし，データの先頭 n 個，末尾 n 個のデータに対してこの処理は行わない．この式を数回繰り返すことによってノイズを除去し，平滑化された波形を得る．

$$S(k) = \frac{1}{2n+1}\sum_{i=-n}^{n} A(k+i) \tag{9}$$

user A の x 軸速度のデータの一部を図 2 に示す．青色はスワイプデータから取得した元データを指し，橙色は青色の元データに対して平滑化した結果を表す．

3.3 特徴抽出

次に平滑化を行った x 軸速度，y 軸速度，接触面積から特徴量を抽出する．今回の実験では x 軸速度，y 軸速度に対しては極値，指の接触面積の変化に対し

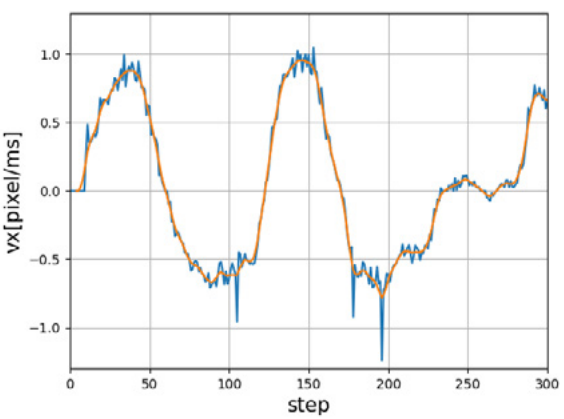

図 2 user A の平滑化後の x 方向速度

ては目印点通過時のデータを特徴量とする．速度の例として，AB 間のスワイプを考える．現在 A の座標（100，400），B の座標（600，400）である．A から B にスワイプする時，x 軸速度は正の方向に加速し，y 軸速度は変化しないため 0 であることが分かる．また，スワイプ開始点 A と目印点 B でともに速度は 0 になるため，速度の極大値が AB 間に存在していることが分かる．本実験ではこのような極値を速度の特徴点として抽出する．続いて接触面積の特徴量について説明する．図 1 から 1 回のスワイプで通過する目印点は 12 であることが分かる．また，x 軸速度，y 軸速度のデータと接触面積のデータの性質は異なり極値を取り出すことが困難なことが分かっている．そこで本実験では目印点を通過した瞬間のデータ値を取得する．

図 3 に user A の x 軸速度，図 4 に user A の 接触面積のデータを示す．どちらも赤点が抽出する特徴点の位置を表す．

4. サポートベクターマシンを用いた個人認証実験

4.1 実験方法

- 独立認証

  平滑化された x 軸速度，y 軸速度のデータからは特徴量を 7 点抽出し，接触面積のデータからは特徴量を 12 点抽出する．そして抽出した特徴量を用いて被験者ごとに x 軸速度，y 軸速度，接触面積の分類器を作成して分類を行う．

- 多数決認証

  そして，x 軸速度，y 軸速度，接触面積の独立認証の結果に対して多数決を取る．以下に多数決認証の定義式を示す．

表 1 既知 user 認証実験の user の組み合わせ

分類器	学習データ	テストデータ
A	A，B，C	A，B，C
B	B，C，D	B，C，D
C	C，D，E	C，D，E
D	D，E，A	D，E，A
E	E，A，B	E，A，B

表 2 未知 user 認証実験の user の組み合わせ

分類器	テストデータ
A	D，E
B	E，A
C	A，B
D	B，C
E	C，D

$$ANDi(S_i, X_i, Y_i) = \begin{cases} 1 & (Vote(S_i, X_i, Y_i) \geq 0.5) \\ 0 & (Vote(S_i, X_i, Y_i) < 0.5) \end{cases} \tag{10}$$

$ANDi$ は i 回目のスワイプの多数決認証の結果である．

多数決認証は，1 つの特徴で異常な結果を返した場合でも，正常な結果を返すことができ，独立認証よりも高精度な認証精度が得られることが期待できる．

今回の実験には 2 種類のテストデータを用いて，多数決認証実験を行っていく．既知 user 認証実験では，学習データとテストデータが同じ user である．これ

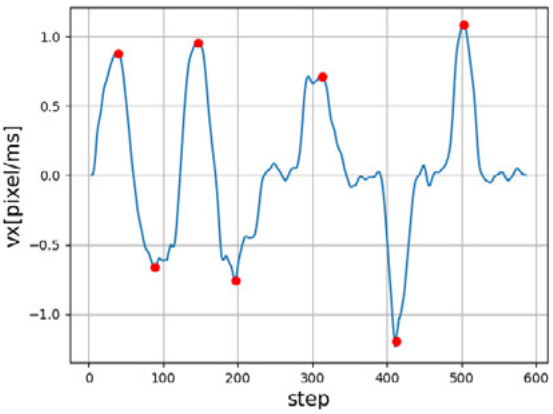

図 3 user A の x 軸速度の特徴点

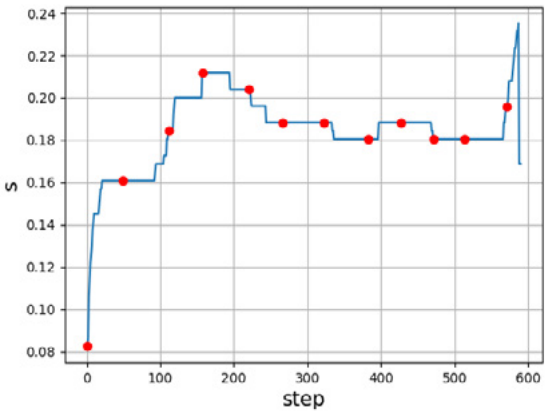

図 4 user A の接触面積の特徴点

表 3 既知 user 認証の結果

2*	x 軸速度		y 軸速度		接触面積		多数決認証	
	FRR (%)	FAR (%)	FRR (%)	FAR (%)	FRR (%)	FAR (%)	FRR (%)	FAR (%)
user A	10	0	20	0	0	0	0	0
user B	60	0	40	5	0	5	30	0
user C	20	0	10	0	0	0	10	0
user D	10	5	0	5	10	0	0	0
user E	0	15	10	0	0	0	0	0
平均	20	4	16	2	2	1	8	0

表 4 未知 user 認証の結果

2*	x 軸速度	y 軸速度	接触面積	多数決認証
	FAR (%)	FAR (%)	FAR (%)	FAR (%)
user A	0	0	3	0
user B	0	0	0	0
user C	0	0	15	0
user D	20	20	50	28
user E	0	0	3	0
平均	4	4	14.2	5.6

は，先行研究と同じ実験方法である．そして，未知 user 認証実験では，学習データに用いていない user をテストデータに用いる．表 1 の user A の分類器を例に説明する．

1. A の奇数データを出力 1 で A と学習させる．
2. B, C の奇数データを出力 0 で A でないと学習させる．
3. A, B, C の偶数かつ窓数 n = 5 のデータをテストデータとして検証する．

4.2 既知 user 認証実験

学習データに, user 3 人の奇数番目を用い，テストデータには，学習データに用いた user の偶数番目のデータを用いて評価を行う．学習データとテストデータの user の組み合わせを表 1 に示す．表 1 の分類器の欄に記載されている人は，その人の分類器を作成して分類していくことを表している．

4.3 未知 user 認証実験

既知 user 認証実験と同様，学習データに，user 3 人の奇数番目を用い，テストデータには学習に用いていない user のデータを用いて評価を行う．学習データとテストデータの user の組み合わせを表 2 に示す.

4.4 評価方法

認証精度は, 本人拒否率 FRR (False Reject Rate) と他人拒否率 FAR (False Accept Rate) で評価を行う．

$$\mathrm{FRR} = \frac{\text{本人拒否回数}}{\text{試行回数}}, \mathrm{FAR} = \frac{\text{他人受け入れ回数}}{\text{試行回数}} \tag{11}$$

4.5 パラメータの決定

SVM のカーネルには rbf カーネルを用いる．ハイパーパラメータの C は 1 と固定し，γ は学習データの FRR と FAR が収束したときの値を用いて検証を行った．

$$\boldsymbol{K}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp \frac{||\boldsymbol{x}_1 - \boldsymbol{x}_2||^2}{2\sigma^2} \tag{12}$$

4.6 実験結果

既知 user 認証実験の結果を表 3，未知 user 認証実験の結果を表 4 に示す.

表 3 を見ると，x 軸速度，y 軸速度に比べて，接触面積は本人拒否率と，他人受け入れ率の精度が高かった．x 軸速度，y 軸速度の本人拒否率はユーザー平均を取ると 20% と 16% と高いが，多数決認証をすることで，平均値が 8% と低く抑えられていることが分かる．また，各特徴量の他人受け入れ率が数%あっても多数決を取ることで 0% になった．そして，先行研究の標準化ユークリッド距離よりも高い精度を実現することができた．

表 4 を見ると，接触面積の他人受け入れ率が x 軸速度，y 軸速度と比較して，高いことが分かる．user ごとに見ると user B は各特徴量の他人受け入れ率が 0%となった．そして，多数決認証の結果では，user D 以外は 0%となり，平均を取ると，5.6%となった．

5. 考察

既知 user 認証実験の多数決認証の FAR が 0 %，未知 user 認証実験の FAR が 5.6%であったことから，学習データと未知のデータに似た特徴量を持つ user が存在すると考えられる．スマートフォンに実装するこ

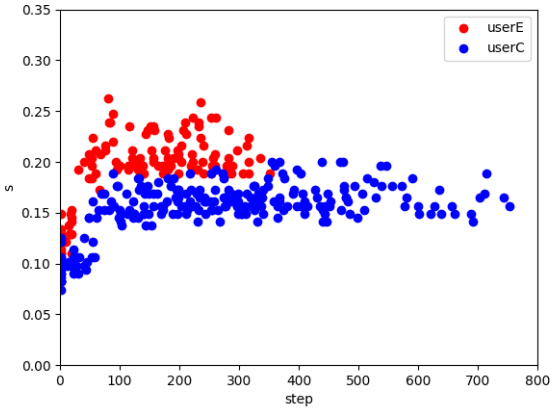

図 5 user C と user E の接触面積

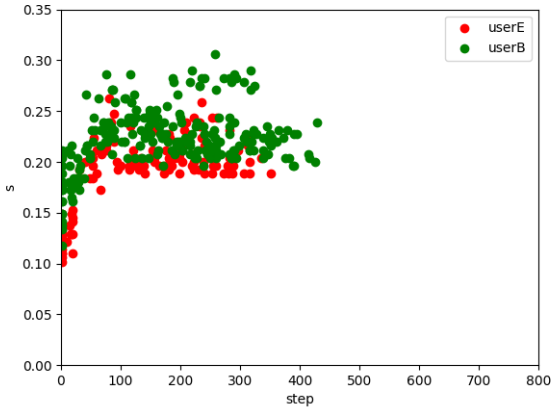

図 6 user B と user E の接触面積

とを考えたときに第 3 者による侵入を防ぐことのほうが重要であるため，本人拒否率よりも他人受け入れ率を低く抑えることのほうが重要である．今回の実験で user D 他人受け入れ率が 0%でなかった理由を明らかにしていく必要がある．

user D の分類器では，user B と user C を未知データとして用いた．多数決認証の結果では，user C より user B の方が誤認識している割合が多く，なかでも接触面積を用いた分類では，user B のすべてのデータを user D 誤認識していた．そこで，user B と user D の接触面積の特徴量の比較を行った．user C と user D の接触面積のデータを図 3 に示し，user B と user D を図 4 に示す．図 3 と図 4 を比較すると，user C と user D は点の重なっている部分が少なく，特徴量の違いが現れていることが分かる．一方で user B と user D は user C と比較して点が重なっている部分が多く，似た特徴量を持っていることが分かる．よって今回の接触面積の分類で user D の FAR が 50% と高かったのは，user B と user D の特徴量が似ていたからだと考えることができる．

6. まとめ

今回，先行研究の標準化ユークリッド距離を用いた手法よりも高い精度を出すために，サポートベクターマシン用いた実験を行った．既知 user 認証実験では，FRR が 8%，FAR が 0%となり，先行研究より高い精度を出すことができた．また，未知 user 認証実験では，FAR が 5.6%であった．未知 user 認証実験の FAR が 0%に抑えられなかった原因として，user D と user B の接触面積の特徴量が似ており，分類することが難しかったからなのではないかと考えられる．現時点での解決方法として，x 軸速度，y 軸速度のデータから特徴量として，特徴点の速度を用いているが，特徴点間の時間など特徴量を増やすことで，解決するのではないかと考えられる．以上の結果から，1 人の user を除いて，他の user の FAR は 0% に抑えることができたため，サポートベクターマシンを用いることで，強固で実用的な認証方式の実現に成功した．

参考文献

[1] 石黒司，福島和英，清本晋作，三宅優 ：“モバイル端末のロック解除向けパターン認証の安全性評価” 情報処理学会研究報告，Vol.2012-SPT-4 No.41，pp.273-278，2012

[2] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith,“Smudge Attacks on Smartphone Touch Screens.”Department of Computer and Information Science University of Pennsylvania

[3] Guixin Ye , Zhanyong Tang , Dingyi Fang , Xiaojiang Chen , Kwang In Kim , Ben Taylor , and Zheng Wang ,“Cracking Android Pattern Lock in Five Attempts.”NDSS ’17, 26 February - 1 March 2017, San Diego, CA, USA

[4] 牧野隆典，山田健一郎，納富一宏，斎藤恵一 ：“スマートフォンにおけるパターン認証の強化～軌跡情報および傾き情報に基づく生体認証～” 第 26 回バイオメディカル・ファジィ・システム学会年次大会講演論文集，pp.25-28，2013

[5] 小松哲幸 黒岩丈介 小高知宏 諏訪いずみ 白井治彦，”スワイプ操作における無意識的特徴量を用いた個人認証，”平成 28 年度電気関係学会北陸支部大会講演論文集，F2-42(2016.9)

電力使用量予測のための深層学習手法における最適なモデル選択に向けて

山名田 恭吾*　黒岩 丈介**　小高 知宏**　諏訪 いずみ***　白井 治彦****

An Investigation of Optimal Model Selection in Deep Learning Methods for Electricity Usage Prediction

Kyogo YAMANADA*,　Jousuke KUROIWA**, Tomohiro ODAKA**
Izumi SUWA*** and Haruhiko SHIRAI****

(Received February 1, 2023)

In this paper, we investigate deep learning methods to predict electricity consumption. We employ RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory), both of which are applicable in handling time-series data, to predict electricity usage. The target data are hourly, daily, and weekly electricity consumption data. The purpose of this paper is to investigate prediction accuracy for various structures of RNN and LSTM. The results show that the prediction accuracy surely depends on either RNN or LSTM and its structure for each of the hourly, daily, and weekly forecasts.

Key words :*Electric Usage Prediction, Deep Learning, RNN, LSTM*

1. はじめに

東日本大震災以降に，コストが低く，効率的に発電できる原子力発電が安全性に問題があると考えられている．原子力発電所の稼働の減少により，原子力発電よりもコストが高い火力発電を用いた発電が増えている．そのため，発電コストが増加し，電気料金の値上げが進んでおり，消費者の負担が増加している．近年，世界的に持続可能な開発目標（Sustainable Development Goals）という活動が取り組まれている．2030年までに持続可能でよりよい世界を目指すという目標を指しており，電力使用量の予測を行うことでコスト削減をはじめとするこれらのような取り組みに貢献ができるのではないかと考えた．

電力使用量の予測が可能になった場合には，長期的には季節ごとの電力使用量をもとに適切な時期の化石燃料の調達ができ，短期的には電力使用量が変動する時間帯に合わせて発電所の稼働状態を設定することができるようになる．これまでの電力使用量のデータから大まかな変動は予測が可能であるが，深層学習を用いて予測を行うことでより詳細に予測ができるようになり，適切な量の化石燃料の調達や，発電量の調節に役に立つ[1]．適切な電力供給を行うことができれば，非効率な発電などを行う必要がなくなり，消費者の負担軽減にも繋がる[2]．

人工知能に関する多くの分野で知的処理技術である深層学習が用いられている．深層学習は機械学習の一つの手法であり，ニューロンが多層に重ねられたニューラルネットワークを用いた分析を行うことである[3]．深層学習は，表情認識や物体認識などの画像認識や，普段人間が用いる自然言語に対して解析する自然言語処理，音声で家電などを操作すること

*大学院工学研究科 知識社会基礎工学専攻
*Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering
**知能システム工学講座
**Department of Human and Artificial Intelligent Systems
***仁愛女子短期大学 生活科学学科
***Jin-ai Women's College
****工学部 技術部
****Technical Division

ができる音声認識などがある．深層学習は株価の予測や電力使用量の予測にも用いることができる．近年は深層学習が盛んに行われているが，課題もある．学習の時にデータ数が多いほど精度が高くなる傾向があるが，計算時間が長くなってしまうことや，データが不足している場合には精度の向上が望めないことも挙げられる．また，ネットワークの構造によっても精度が異なってくることもあるため，複数のモデル構築を試したり，ハイパパラメータを手動で変更し予測を行う必要があるなどの問題も存在する．人間の脳では，一つの情報からではなく，周りの状況などの他の情報からも情報処理が行われている．そのような処理の方法を模した手法にマルチモーダル学習がある．マルチモーダル学習は複数の情報を入力として学習を行う手法のことである．電力使用量が変化する要因は気温や天気，降水量など様々考えられる．そこで，電力使用量だけではなく，他要因も組み合わせて入力として，学習することで予測精度が高くなることが考えられる．

本研究の目的は，電力使用量の時間変化量のみから電力使用量を予測する最適なモデル及びモデル構造を明らかにすることである．これにより，電力使用量が突然跳ね上がるような異常値の予測を実現したり，マルチモーダル入力による高精度な電力使用量予測を実現する研究へ繋げていくことを目的としている．

2. 時系列データ処理のための深層学習手法

2.1 ニューラルネットワーク

ニューラルネットワークは，人間の神経回路網をもとにして作られた数学モデルのことである．人間の脳は情報伝達や記憶の定着のために神経細胞の結びつきであるニューロンを用いており，脳神経系の強力な学習能力を画像認識などの分野へ応用する研究が近年盛んに取り組まれている．ニューラルネットワークは，入力層と中間層，出力層の構造をもつ．層と層の間にはシナプス結合荷重 W があり，ニューロン同士を繋ぐ役割を果たしている．シナプス結合荷重のことを重みと呼ぶ．ニューラルネットワークの中間層は多数重ねることができる．中間層が多層構造になったニューラルネットワークのことを深層学習（Deep Learning）と呼ぶ．

2.2 RNN

RNN（Recurrent Neural Network） とは，ある時刻における中間層の出力を次の層の入力とすることができる再帰的構造を持つニューラルネットワークモデルのことである．このニューラルネットワークモデルは，David E. Rumelhart[4] の研究を元にして開発されたものである．過去の情報を扱うことができるように，前の時刻での中間層の出力を用いることにより，それ以降の出力に影響を与えていく．これにより，過去の情報を保持していくことができるため，時系列データを扱うことが可能になっている．データの形状は，$\boldsymbol{x}(1), \cdots, \boldsymbol{x}(T)$ という T 個のデータが入力データ群となっている．RNN では，入力層と中間層の間の重み U，時刻 t での入力を $x(t)$，f, g を活性化関数として中間層の出力値 $\boldsymbol{s}(i)$ とネットワークの出力 $y(t)$ は以下のように示される．

$$y(t) = f(U\boldsymbol{x}(t) + W\boldsymbol{s}(t-1)) \tag{1}$$

入力 x と $s(t-1)$ により $s(t)$ が更新されていく．

$$y(t) = g(W\boldsymbol{s}(t)) \tag{2}$$

2.3 LSTM

LSTM（Long Short-Term Memory）は，RNN の欠点を補うようにして構築されたニューラルネットワークモデルである．これは，1997 年に Sepp Hochreiter や Jürgen Schmidhuber が発表したモデルであり，時系列データを扱うことができる [5]．これは，長期記憶（Long Term Memory）と短期記憶（Short Term Memory）の二つを組み合わせて作られており，RNN では時刻が離れているデータ間の依存関係を学習することが難しかったという課題を克服するために開発されたモデルである．近い過去を扱う短期記憶と遠い過去を扱う長期記憶が可能であるため，時系列データが長い場合には有効な場合が多い．LSTM は RNN の中間層を LSTM block というメモリと入力ゲート及び忘却ゲート，出力ゲートの三つのゲートを持つブロックに置換することで実現されている．メモリは入

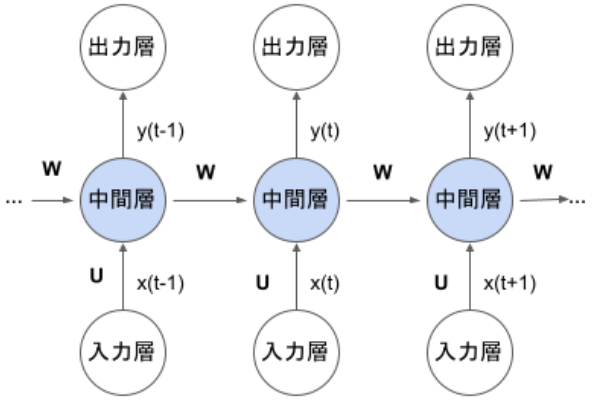

図 1 RNN モデル図

力の依存性を記憶している．入力ゲートはメモリへ渡す入力を調整し，忘却ゲートはメモリの値を破棄する量を調整，出力ゲートはメモリの値を活性化関数に渡す量を調整する役割を持っている．

$$\boldsymbol{f}(t) = \sigma_g(U_f\boldsymbol{x}(t) + W_f\boldsymbol{h}(t-1)) \tag{3}$$

$$\boldsymbol{i}(t) = \sigma_g(U_i\boldsymbol{x}(t) + W_i\boldsymbol{h}(t-1)) \tag{4}$$

$$\boldsymbol{o}(t) = \sigma_g(U_o\boldsymbol{x}(t) + W_o\boldsymbol{h}(t-1)) \tag{5}$$

$$\boldsymbol{c}(t) = \boldsymbol{f}(t)\odot\boldsymbol{c}(t-1)+\boldsymbol{i}(t)\odot\sigma_c(U_c\boldsymbol{x}(t)+W_c\boldsymbol{h}(t-1)) \tag{6}$$

$$y(t) = \boldsymbol{o}(t)\odot\sigma_h(\boldsymbol{c}(\boldsymbol{t})) \tag{7}$$

以上より，活性化関数 σ，入力層と中間層間の重み U，中間層と出力層間の重み W を用いて忘却ゲート $\boldsymbol{f}(t)$ や入力ゲート $\sigma\ (t)$，活性化ベクトル $\boldsymbol{o}(t)$，中間層の出力値 $\boldsymbol{h}(t)$，ネットワークの出力値 $y(t)$ を定義している．$\odot$ はアダマール積を表しており，ベクトルの要素ごとに積を求める演算子である．

3. 電力使用量予測実験

3.1 対象データ

本研究では，東京電力パワーグリッド社が提供している電力使用量のデータを用い，2016 年 4 年から 2022 年 3 月までの期間を用いる．時間ごとのデータは 24（時間）× 365（日）× 6（年）で 52560 個存在している．日ごとのデータは 24 時間ごと，週ごとのデータは日ごとのデータを 7 日ごとにまとめて使用した．電力使用量のデータの一部を図 3 に示す．2016 年 11 月 21 日から 2 週間分の電力使用量の波形を示している．縦軸は電力使用量の 10^4 kW，横軸は時間を表している．図は，日ごとのデータを示しており，一日の間で規則的な波が現れている．図より，日中は電力使用量が増え，夜間は減っていることが読み取れる．本実験ではこのような電力使用量のデータを用いた．

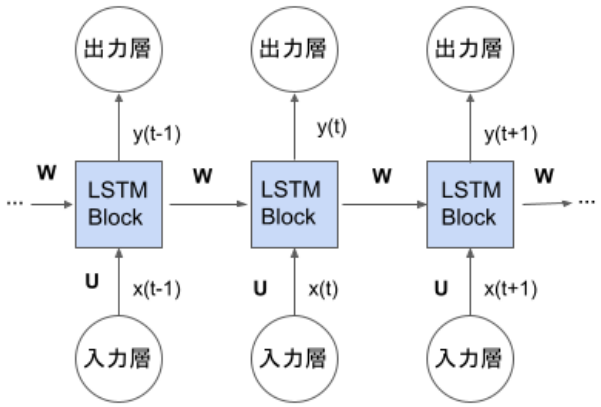

図 2 LSTM モデル図

3.2 異常値検出

電力使用量が，トレンドから外れるような急激な変化点を以降では異常値と呼ぶ．異常値は，台風や豪雪などの，異常気象による影響により現れる傾向が高い．そのような異常気象のときの電力使用量を予測することが大きな課題であるが，これまでの研究では精度の高い予測ができていない．用いた電力使用量には，異常値が多く含まれており，そのような点についても高い予測精度を示すことが可能になれば，適切な電力調整や化石燃料調達に役立つ．本実験では，実測データから異常値検出を行うが，実際の予測では，天気や降水量などの電力使用量に関わる他要因も含め，異常値が現れそうな位置を検出することを想定している．

時系列データの平滑化には，移動平均という手法が存在し，大きく分類すると単純移動平均と加重移動平均，指数移動平均の 3 種類に分類できる．移動平均とは，金融分野の分析や気象の分析に用いられる手法のことである．

単純移動平均とは，SMA（Simple Moving Average）と呼ばれ，直近の n 個のデータを平均をとり，単純な重み付けをすることなく，平滑化を行う手法のことである．しかし，重み付けがないため，過去のデータの影響を受けてしまう．

加重移動平均とは，WMA（Weighted Moving Average）と呼ばれ，重みを一定量ずつ線形に減らすことで平滑化を行う手法のことである．直近のデータの

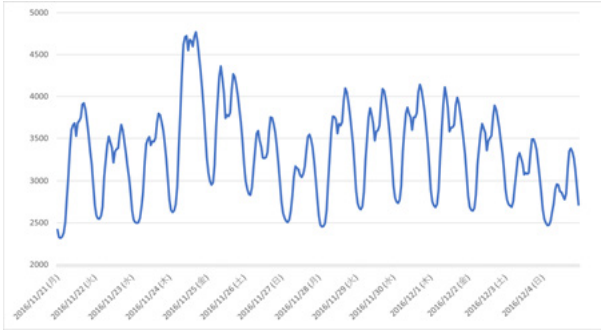

図 3 電力使用量の元データ

重みを n として，その一つ前のデータの重みを $n-1$ とする．このように重みを 0 まで小さくしていくことで古いデータの重みを減らし，古いデータの影響を小さくする手法である．

指数移動平均とは，EMA（Exponential Moving Average）と呼ばれ，加重移動平均では線形に減らした重みを指数関数的に減少させる手法のことである．単純移動平均と比べて直近のデータに影響を受けやすいという特徴がある．

まず，指数移動平均による時系列データの平滑化を行った．期間 14 日ごとのデータを用いて移動標準偏差（Moving Standard Deviation）を計算し，同様に 14 日ごとの指数移動平均を計算する．指数移動平均の計算結果が移動標準偏差から 1 倍以上大きい値を異常値として扱った．図 4 に異常値を検出した結果を示す．青色が元の波形，橙色の波形が EMA，橙色の点が異常値を表している．図から読み取れるように，トレンドから大きく外れているところが異常値となっている．評価は全区間と異常値のみを集めた区間の 2 区間に分けて行う．それぞれの区間について RMSE や分散の指標により評価を行う．

3.3 RNN による予測手法

本研究では，RNN に入力として時間ごとや日ごと，週ごとの電力使用量のデータを与える．このデータから時間ごと，日ごと，週ごとの電力使用量の予測を行った．用いたデータを，Train data と Test data が 8:2 になるように分割し，入力データとした．入力系列の長さは 12 とした．そのため，時間ごとの予測なら 12 時間の電力使用量のデータを入力し，次の 1 時間の電力使用量を予測するような予測方法である．同様に，日ごと，週ごとの予測も行う．最適化アルゴリズムは Adam を用い，学習率の初期値は 0.001，損失関数として平均二乗誤差を用いた．活性化関数は ReLU，バッチサイズは 64 とし，epoch 数は validation loss が 10 回以上改善しなくなるまで繰り返すこととした．層数と中間層の素子数を変化させることで予測を行った．層数は 1 層，2 層，3 層を変化させ，素子数は 100 個，200 個，300 個を変化させることによって予測精度を調べた．RMSE と分散を用いることによって，予測した電力使用量の予測結果の評価を行った．また，RMSE の式は以下で定義される．

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

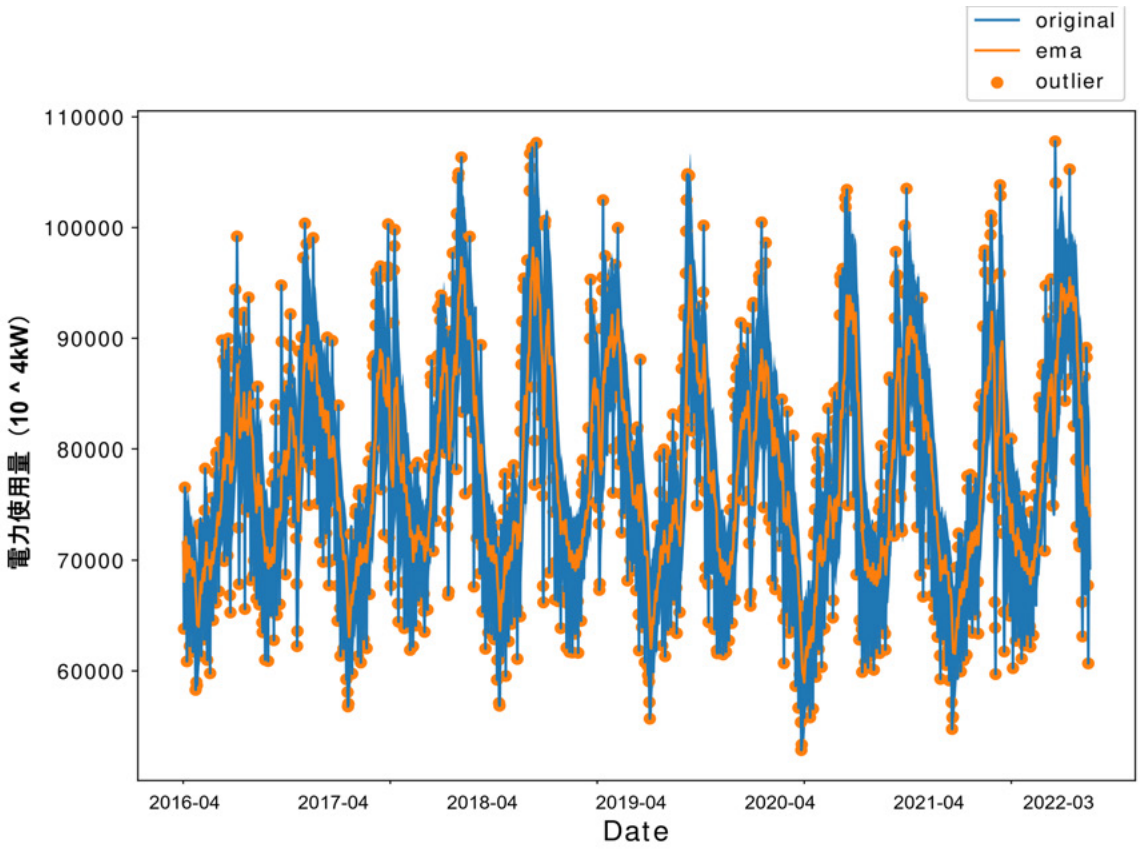

図 4 異常値検出結果

n はデータ数，y と $\hat{y}$ はそれぞれ入力値と予測値である．

3.4 LSTM による予測手法

RNN のときと同様に，LSTM に入力として時間ごとや日ごと，週ごとの電力使用量のデータを与えた．このデータから時間ごと，日ごと，週ごとの電力使用量の予測を行った．用いたデータを，Train data と Test data が 8:2 になるように分割し，入力データとした．入力系列の長さは 12 とした．そのため，時間ごとの予測なら 12 時間の電力使用量のデータを入力した. 同様に，日ごと，週ごとの予測も行う．最適化アルゴリズムは Adam を用い，学習率の初期値は 0.001，損失関数として平均二乗誤差を用いた．活性化関数は ReLU，バッチサイズは 64 とし，epoch 数は validation loss が 10 回以上改善しなくなるまで繰り返すこととした．層数と中間層の素子数を変化させることで予測を行った．層数は 1 層，2 層，3 層を変化させ，素子数は 100 個，200 個，300 個を変化させることによって予測精度を調べた．RMSE と分散を用いることによって，予測した電力使用量の予測結果の評価を行った．

表 1 RNN を用いた時間ごとの予測

RMSE / 分散	Test（全区間）
1 層（75）	0.0177 / 0.0323
1 層（100）	0.0168 / 0.0349
1 層（125）	0.0207 / 0.0319
2 層（100）	0.0173 / 0.0338
2 層（125）	0.0172 / 0.0329
2 層（150）	0.0176 / 0.0338
3 層（100）	0.0184 / 0.0320
3 層（125）	0.0163 / 0.0334
3 層（150）	0.0180 / 0.0338

3.5 結果

電力使用量の元データ，学習データとテストデータに対する予測結果を一例として図 5 に示す．電力使用量の元データは緑色，学習データの予測は青色，テストデータに対する予測は橙色で示している．

RNN と LSTM を用いて，時間ごと，日ごと，週ごとの電力使用量を予測した結果を表 1 に示す．これは，RMSE と分散を用いて，Test data について予測結果を評価した表である．

RNN を用いた時間ごとの予測結果は，3 層，ニューロン数 125 個のときが最も精度の良い予測となって

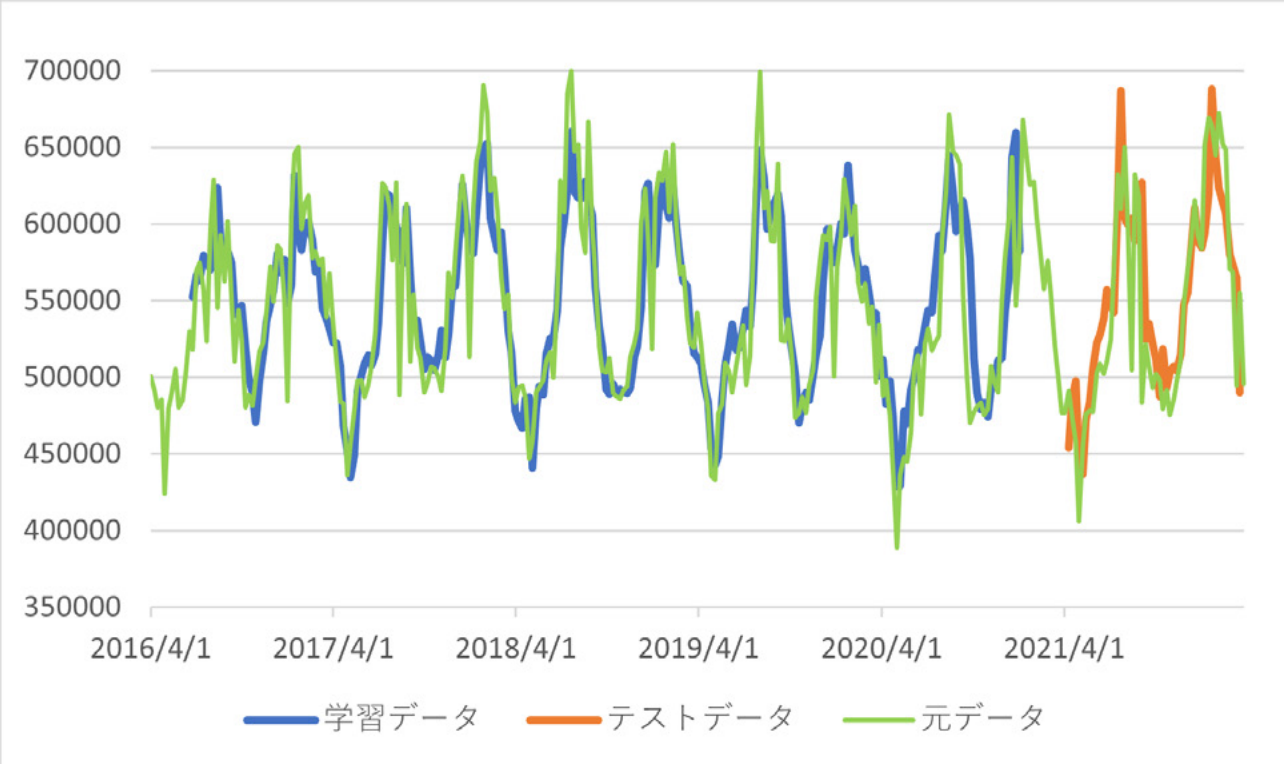

図 5 モデルの予測

表 2 LSTM を用いた時間ごとの予測

RMSE / 分散	Test（全区間）
1 層（75）	0.0158 / 0.0323
1 層（100）	0.0148 / 0.0320
1 層（125）	0.0176 / 0.0325
2 層（100）	0.0172 / 0.0316
2 層（125）	0.0134 / 0.0330
2 層（150）	0.0165 / 0.0323
2 層（100）	0.0163 / 0.0344
3 層（125）	0.0139 / 0.0333
3 層（150）	0.0148 / 0.0349

表 3 区間ごとの予測精度（時間）

RMSE/分散	異常値以外の区間	異常値区間
RNN	0.0186 / 0.0321	0.0148 / 0.0377
LSTM	0.0148 / 0.0309	0.0133 / 0.0373

いた．時間ごとのデータはデータ数が多く，層を重ねることで複雑な予測に対応しているためこの結果となっている．

表 2 より，LSTM を用いた時間ごとの予測結果は，2 層，ニューロン数 125 個のときが最も精度の良い予測となっていた．RNN のときと同様に，層を重ねて複雑な予測に対応しているためこの結果となっている．

また，時間ごとの異常値区間の予測精度を表 3 に示す．

RMSE の結果は異常値区間の方が低くなっていることが読み取れる．分散の値は大きくなっており，精度が高くないことがわかる．RNN と LSTM では LSTM の方が区間別にみても精度が高くなっている．

表 4 より，RNN を用いた日ごとの予測結果は，1 層，ニューロン数 125 個のときが最も精度の良い予測となっていた．時間ごとのデータと比べて，日ごとのデータはデータ数が少ないため，層数は 1 層が一番良くなっているが，1 層から 2 層までで精度に大きな差は生まれていない結果となった．

表 5 より，LSTM を用いた日ごとの予測結果は，2 層，ニューロン数 125 個のときが最も精度の良い予測となっていた．時間ごとの予測結果と同じ結果となった．また，RNN と LSTM どちらの予測も，時間ごとの結果と比べて RMSE や分散の値が大きくなっていることがわかる．

また，日ごとの異常値区間の予測精度を表 6 に示す．

表 4 RNN を用いた日ごとの予測

RMSE / 分散	Test（全区間）
1 層（100）	0.095 / 0.044
1 層（125）	0.091 / 0.041
1 層（150）	0.093 / 0.042
2 層（100）	0.094 / 0.042
2 層（125）	0.093 / 0.041
2 層（150）	0.098 / 0.041
3 層（100）	0.107 / 0.042
3 層（125）	0.098 / 0.042
3 層（150）	0.099 / 0.041

表 5 LSTM を用いた日ごとの予測

RMSE / 分散	Test（全区間）
1 層（100）	0.103 / 0.042
1 層（125）	0.098 / 0.040
1 層（150）	0.111 / 0.042
2 層（100）	0.128 / 0.042
2 層（125）	0.097 / 0.043
2 層（150）	0.109 / 0.043
3 層（100）	0.112 / 0.042
3 層（125）	0.109 / 0.044
3 層（150）	0.130 / 0.038

日ごとのデータの区間ごとの予測精度では，RNN と LSTM ともに異常値区間の RMSE が低くなっていた．分散は，異常値区間が大きくなり，予測精度が高くないことがわかった．全区間で高い精度を示したのは RNN だが，区間ごとに計算しても同じ結果になったと言える．

表 7 より，RNN を用いた週ごとの予測結果は，3 層，ニューロン数 125 個のときが最も精度の良い予測となっていた．データ数が一番少ない週ごとのデータであるが，予測結果は 3 層，ニューロン数 125 個であった．しかし，1 層から 3 層までの予測結果では大きな差はない結果となっていた．

表 8 より，LSTM を用いた週ごとの予測結果は，3 層，ニューロン数 125 個のときが最も精度の良い予測となっていた．1 層，2 層，3 層で結果を比べると 3 層のときの結果が 1 層，2 層に比べて精度が高くなっていた．

また，週ごとの異常値区間の予測精度を表 9 に示す．

週ごとのデータの区間ごとの予測精度は，RNN と LSTM ともに異常値区間の方が RMSE の値が大きく

表 6 区間ごとの予測精度（日）

RMSE/分散	異常値以外の区間	異常値区間
RNN	0.088 / 0.041	0.109 / 0.051
LSTM	0.088 / 0.043	0.116 / 0.055

表 7 RNN を用いた週ごとの予測

RMSE / 分散	Test（全区間）
1 層（100）	0.194 / 0.061
1 層（125）	0.172 / 0.065
1 層（150）	0.187 / 0.065
2 層（100）	0.178 / 0.059
2 層（125）	0.175 / 0.064
2 層（150）	0.182 / 0.062
3 層（100）	0.180 / 0.062
3 層（125）	0.157 / 0.062
3 層（150）	0.181 / 0.062

表 8 LSTM を用いた週ごとの予測

RMSE / 分散	Test（全区間）
1 層（100）	0.199 / 0.048
1 層（125）	0.167 / 0.050
1 層（150）	0.234 / 0.051
2 層（100）	0.214 / 0.049
2 層（125）	0.201 / 0.051
2 層（150）	0.223 / 0.045
3 層（100）	0.166 / 0.048
3 層（125）	0.162 / 0.052
3 層（150）	0.190 / 0.048

表 9 区間ごとの予測精度（週）

RMSE/分散	異常値以外の区間	異常値区間
RNN	0.171/ 0.071	0.211 / 0.072
LSTM	0.179 / 0.059	0.223 / 0.058

なり，予測精度が低くなっていた．RNN では分散は，僅かに異常値区間の方が精度が低く，LSTM では僅かに異常値区間の方が精度が高いという結果であった．

時間ごと，日ごと，週ごとの電力の予測を RNN と LSTM を用いて行ってきたが，予測精度が高かったものは全てニューロン数が 125 個であることがわかった．よって，電力使用量の予測にはどちらのモデルを用いるときでもニューロン数を 125 個に設定することで高い精度が得られる．

4. 考察

時間ごとの予測結果は RNN の場合では 3 層，ニューロン数 125 個のとき，LSTM では 2 層，ニューロン数が 125 個のときが最も精度の良い予測となっている．RNN は膨大な期間のデータを記憶することには向いていないため，少ない層数では予測できなかったものだと考えられる．LSTM では，RNN とは異なり，長期記憶が可能になっていることから，層数は 2 層で予測できている．RNN より LSTM の方が全体的に RMSE，分散どちらの数値も小さくなっており，LSTM の方が予測精度が高くなっていた．また，異常値については，分散が大きくなっているため，異常値区間の電力使用量の予測は精度が高くないことがわかった．

日ごとの予測結果は，RNN の場合は 1 層，ニューロン数 125 個のとき，LSTM では 2 層，ニューロン数 125 個のときが最も精度の高い予測となっている．RNN のときは，時間ごとのデータよりもデータ数が少ないために層数が 1 層で高い精度の予測ができたと考えられる．LSTM では 2 層のときが精度が高くなっているが，1 層の結果と値に大きな差はなく，層数は RNN と LSTM どちらの場合も複雑に重ねる必要がない．それぞれの結果を比較すると，RNN の方が全体的に RMSE と分散が小さくなっており，精度が高いといえる．異常値区間については予測精度が低くなっていた．時間ごとのデータの予測と同様に，日ごとのデータの予測も異常値区間の予測精度が高くないことがわかった．

週ごとの予測結果は，RNN の場合は 3 層，ニューロン数 125 個のとき，LSTM では 3 層，ニューロン数 125 個のときが最も精度の高い予測となっている．RNN，LSTM ともに層数が多くなっている．週ごとのデータでは，1 時間の電力使用量のデータを 24 時間分合わせ，さらに 7 日分合わせて作成している．天気や気温などが不規則なことに関連して，1 週間の電力使用量は複雑なデータとなっている可能性が高い．データ数が少ないため層数は少なくても予測できるとは考えられるが，週ごとのデータが複雑なデータとなっていることにより層数は RNN と LSTM どちらも 3 層が予測精度が高くなったものだと考えられる．異常値区間については，全区間に比べてかなり予測精度が悪くなっていることがわかった．これもデータ数が少ないこと，データが複雑なことにより精度

が下がったものだと考えられる．

データ量によって RNN と LSTM どちらのモデルを用いるか，層数の設定は変える必要があることがわかった．ニューロン数については，125 個がほとんどの結果において高い精度を示していた．しかし，異なるニューロン数の結果と比べて大きく精度が高くなっているとはいえない．今後は，用いた電力使用量のデータ量より長期のデータを用いて実験を行い，これが適切なニューロン数であるか検討する必要がある．

結果より，データ量が非常に多くなるほど LSTM が予測に向いており，反対にデータ量の少ない場合は RNN の方が予測に向いている傾向があることがわかった．用いるデータの量に合わせて適切なモデルとそのモデル構造を選択することで，より精度の高い電力使用量の予測の実現に繋がる．また，異常値区間については，精度が低くなっていたため，電力使用量以外の他要因を用いる必要があると考えられる．そうすることにより，天気や降水量など，電力使用量の変化に影響する他要因からも予測が行えるようになるため，集合型のマルチモーダル入力による学習を行う必要がある．集合型のマルチモーダル学習とは，電力使用量のみから電力使用量の予測を行うモデルと，天気から電力使用量の予測を行うモデルや降水量から電力使用量の予測を行うモデルをそれぞれ作り，それらの予測結果の多数決をとる手法である．ここでいう予測結果は，前の時刻から次の時刻にかけての電力使用量の変化量を予測することを表す．このように予測することで急激に電力使用量が変化する時刻を予測できるようになると考えられる．

5. おわりに

本研究では，発電所の稼働状態を調整することや適切な化石燃料の調達のための電力使用量の予測を行った．予測結果は RMSE や分散を用いて精度を評価した．RNN や LSTM を用いた予測では，時間ごと，日ごと，週ごとについてそれぞれ適切なモデルが存在し，電力使用量のトレンドを予測することができるようになった．今後は全区間においてより精度の高い精度を追求していく．また，異常値区間の予測精度については，全区間の予測精度に比べて低くなっていた．異常値は，天気や降水量，気温などの要因から起こる可能性が高い．今回の実験では，電力使用量のみの単入力で予測をしたため，異常値の予測には対応できていなかったと考えられる．

今後の課題は，電力使用量の単入力での予測精度の向上を図ることである．また，電力使用量のみを入力データとして扱うのではなく，天気や降水量などの他要因もマルチモーダルな入力として扱う深層学習モデルを構築することである．これにより，電力使用量のトレンドのみではなく，他要因により現れる異常値についても詳細な予測ができるようになる．

参考文献

[1] 市村匠. 鎌田真. **リカレント構造適応型** *Deep Belief Network* **による時系列データの学習**. 計測自動制御学会論文集, Vol.54, No.8, pp.628-639, 2018.

[2] 村上朋子. 松尾雄司, 永富悠. **有価証券報告書を用いた火力・原子力発電コスト構造の分析**. Journal of Japan Society of Energy and Resources, Vol.33, No.5, pp.1-2, 2012.

[3] 浅川伸一. *python* **で体験する深層学習**. pp.302-303, コロナ社, 2016.

[4] G.E.Hinton D.E.Rumelhart and R.J.Williams. *Learning representations by back-propagating errors.* Nature, Vol.323, pp.533–536, 1986.

[5] J.Schmidhuber. S.Hochreiter. *Long short-term memory.* Neural Computation, Vol.9, No.8, pp.1735–1780, 1997.

福井大学学術研究院工学系部門研究報告

http://www.eng.u-fukui.ac.jp/research/memoirs-2/index.html

福井大学学術研究院工学系部門研究報告　別冊　研究活動一覧

http://www.eng.u-fukui.ac.jp/research/researchactivities-2/index.html

国立情報学研究所　論文情報ナビゲータ（福井大学関連）のＵＲＬ（書誌情報のみ）

https://ci.nii.ac.jp/ncid/AA12208150

福井大学学術研究院工学系部門研究報告

2023年3月17日 発行

福井大学大学院工学研究科

〒910-8507　福井市文京3丁目9-1

電話（0776）27-8016（研究・地域連携推進部研究推進課）

印刷所　　能登印刷株式会社