Vol.2022-MUS-134 No.54
Vol.2022-SLP-142 No.54
2022/6/18

LSTM デノイジング変分オートエンコーダによるドラム演奏波形の特徴抽出精度向上

松川瞬[1]　竹沢恵[1]　稲垣潤[1]　真田博文[1]

概要：近年，音楽のリズムによる身体的高揚感＝グルーヴは音楽体験の重要な要素であり，グルーヴの定量的な解明が求められる．オートエンコーダによりリズム波形の特徴を取得する事はグルーヴ抽出において有効であるが，アクセントの弱い箇所がノイズと同一視され，特徴抽出に影響を及ぼしてしまう．本研究では，LSTM 変分オートエンコーダモデルに，予めノイズを含めた入力データからノイズなしの元データを復号するデノイジングオートエンコーダを組み込んだ LSTM デノイジング変分オートエンコーダモデルにより，弱アクセント箇所の特徴抽出への影響を抑えることを試みる．また，実際のグルーヴ抽出における本モデルの有効性を確認する．

キーワード：機械学習，LSTM，VAE，ドラムグルーヴ解析

A Study of Improved Groove Extraction from Drum Sounds Using LSTM Denoising Variational Autoencoder

SHUN MATSUKAWA[†1]　MEGUMI TAKEZAWA[†1]
JUN INAGAKI[†1]　HIROFUMI SANADA[†1]

Abstract: A groove that the feeling of physical exuberance caused by the rhythm of music has become an important element of the musical experience recenly, and quantitative clarification of groove is required. While acquiring features of rhythmic waveforms with an auto-encoder is effective for groove extraction, weak accents in the performance are considered noise and affect feature extraction. In this study, we attempt to suppress the influence of weak accents on feature extraction by using an LSTM variational autoencoder model that incorporates a denoising mechanism that decodes noise-free original data from input data that includes noise in advance, and evaluate the effectiveness of this model in actual groove extraction.

Keywords: Machine Learning, LSTM, VAE, Drum Groove Analysis

1. はじめに

1.1 背景

近年, プロジェクションマッピングや VR（Virtual Reality：仮想現実）など，コンピュータによる現実世界の拡張や仮想空間を用いたサービスの提供に注目が集まっている．それにより，音楽と合わせて視覚・聴覚を刺激するサービスが展開される等，音楽シーンにおいても音楽を「聴く」から「見る」「感じる」へ移行するという考えが広がってきている[1]．その際，映像の視覚効果と音楽のリズムを合致させる演出（音ハメ）などによってリズムに「ノり」，身体的高揚感すなわち「グルーヴ」を得ることが楽しみの一部と言える．

そういった「グルーヴ」を得られるサービスを提供するには，音楽の旋律やリズムが視聴者に与える影響を解析する事が重要である．特にドラムにおいては，「マイクロタイミング」[2]に着目した演奏の時間的逸脱に着目され，メトロノーム時刻の±50 ミリ秒の時間的逸脱がグルーヴ感の表現において重要であると言われている[3]．この時間的逸脱においては実験的検証も良く行われており，ドラムスの基本 3 点（バスドラム，スネアドラム，ハイハットシンバル）を用いた基本的リズムにおけるマイクロタイミングのズレによるグルーヴ感の検証[4]もなされている．マイクロタイミングのズレについて，藤井は予測的符号化理論と結び付けて解説している[5]．この予測的符号化理論は人間が知覚する際の脳の計算原理を説明した理論で，いわば聴者の期待感をモデル化する事であると言える．その中で，脳の予測結果が逸脱しすぎず当てはまりすぎないときにグルーヴ感が大きくなるモデルについて紹介している．

以上より，グルーヴには音楽要素（リズムパターン，譜面）と演奏特性（マイクロタイミングのズレやダイナミクス等，奏者の意図）が大きく関わっている．特に演奏特性の影響は大きく，メタルバンド「聖飢魔Ⅱ」のドラム奏者であるライデン湯澤も，教則ビデオ[6]の中でグルーヴと演奏特性との関連が強いことを述べている．

しかし，記述統計量などのハンドクラフトな特徴量によるアプローチでは，ノリ・グルーヴ感の要となるそれら音

1 北海道科学大学
Hokkaido University of Science

Vol.2022-MUS-134 No.54
Vol.2022-SLP-142 No.54
2022/6/18

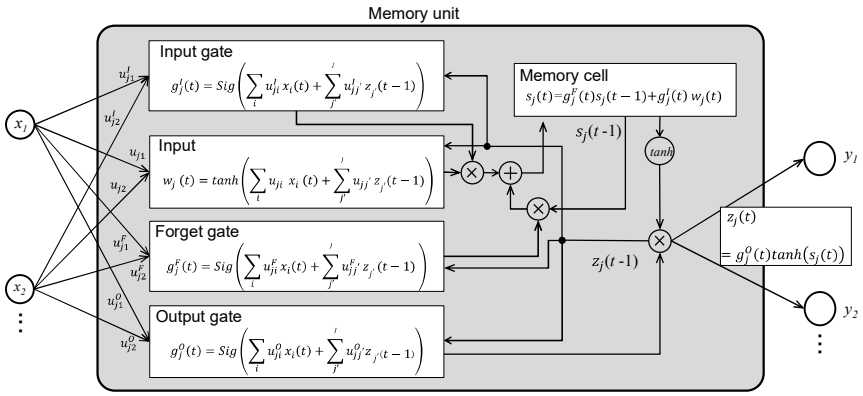

図 1　LSTM におけるメモリーユニットの構造

楽要素・演奏特性を複合的に扱う事が非常に難しい．また演奏特性は複雑であり，適切な特徴の選択そのものが困難である問題もあるため，ハンドクラフトなアプローチでは限界がある．

1.2 本研究の目的

そこで本研究では，機械学習的なアプローチにより，楽曲データを入力として音楽要素と演奏特性を同時かつ定量的に表現するモデルを構築し，ドラムグルーヴをデータドリブンに解析する．著者は以前，LSTM (Long Short-Term Memory)[7]を用いた変分オートエンコーダにより中間層出力分布(LSTM ユニット出力の隠れ状態分布)を 2 次元正規分布で表現する事により，演奏特性のある音源と無い音源における特徴差を情報量で抽出・再構成可能なことを検証した[8][9]．この結果は直感的に，藤井の「聴者の期待感」を情報量として捉えた場合得られるであろう結果と合うものである．しかし，LSTM 変分オートエンコーダにおいて振幅の小さい弱アクセント部分がノイズと捉えられ，演奏特性として抽出できない課題があった．

本稿では，LSTM 変分オートエンコーダモデルに，予めノイズを含めた入力データからノイズなしの元データを復号するデノイジングオートエンコーダを組み込んだ LSTM デノイジング変分オートエンコーダモデルにより，弱アクセント箇所の特徴抽出への影響を抑えることを試みる．また，演奏特性のある音源と無い音源における中間層出力分布差をカルバックライブラー情報量で表現する事により，実際のグルーヴ抽出における本モデルの有効性を確認する．

2. グルーヴ要素の定義

ある演奏者p，あるセクションs（イントロ，サビといった楽曲内の区分け）におけるグルーヴが

$$gr_{p,s} = \{notation, module, tone, dynamics, nuance\}$$

で構成されると定める．ここで，notation は音譜の配置，module はリズムパターン，tone は音色，dynamics はダイナミクス(強弱)，nuance はニュアンス(譜面とのずれ)を表す．このうち，notation・module・tone が音楽要素，dynamics・nuance が演奏特性である．なお，演奏者pの持つグルーヴは，セクションs内で不変とする．

本稿では，演奏特性の dynamics（特にゴーストノーツと呼ばれる「聞こえるか聞こえないか」と言う程度の弱音），nuance（特にルース：ゆったりとしている／タイト：明快であると表現されるタイミングズレ[10]）に着目し，それらが奏者により表現されている演奏音源と，音の大きさ・タイミングが一定の打込音源との差異抽出を行う．

3. デノイジング LSTM 変分オートエンコーダモデル

3.1 モデルの概要

本研究では，時系列の回帰式を近似するリカレントニューラルネットワーク，その中でも長期的な時系列に対応した LSTM （Long Short-Term Memory）に着目する．LSTM は，通常のニューラルネットワークにおける中間層のノードを図 1 に示すようなメモリーユニットに置き換えたもので，通常の入力受容部（Input）のほか，入力・忘却・出力ゲート（Input / Forget / Output gate），メモリーセル（Memory cell）を持っている．自身の状態を保持しているメモリーセルを再帰的に参照しつつ，忘却ゲートで不要な情報の忘却を行うことで，長期的な時系列特徴の保持が期待できる．各所での入出力を一般化して記述すると，時刻tにおけるi番目の入力$x_i(t)$，再帰される値を$z_i(t-1)$，入力部の重さu_{in}，再帰部の重さu_{re}として

$$g(x) = f\left(\sum_i \{u_{in}x_i(t) + u_{re}z_i(t-1)\}\right) \quad \cdots\cdots\cdots \quad (1)$$

で行われる．fは活性化関数である．なお，バイアスの記述は割愛している．

LSTM 変分オートエンコーダは，式 1 の出力値を対象として，変分オートエンコーダ[11][12]により隠れ状態の分布を推定する．分布は正規分布とし，隠れ状態zの分布$P(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$を推定するエンコーダ部分(式 2)と，その分布からサンプリングした値を基に出力を決定するデコーダ部分

Vol.2022-MUS-134 No.54
Vol.2022-SLP-142 No.54
2022/6/18

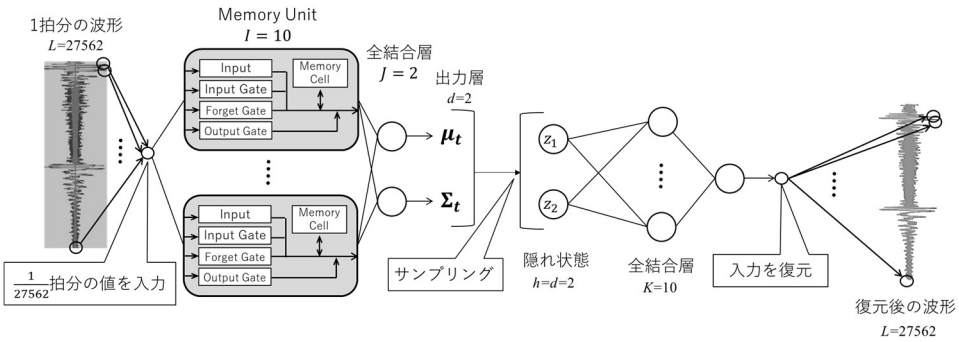

図 2　本稿で用いる LSTM 変分オートエンコーダモデル

(式 3)に分かれる．本稿で用いるモデルの概要図を，図 2 に示す．

$$\boldsymbol{\mu}(\boldsymbol{x_t}),\boldsymbol{\Sigma}(\boldsymbol{x_t}) = f\left(\sum_i \boldsymbol{w}_i \cdot g_i(\boldsymbol{x_t})\right) \tag{2}$$

$$h(\boldsymbol{z_t}) = f\left(\sum_k \boldsymbol{v}_k \cdot \boldsymbol{z_t}\right), \boldsymbol{z_t} \sim N\big(\boldsymbol{\mu}(\boldsymbol{x_t}),\boldsymbol{\Sigma}(\boldsymbol{x_t})\big) \tag{3}$$

これにより，楽曲特徴を分布$P(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$で取得する事が可能となり，演奏特性のある音源（演奏音源）とない音源（打込音源）を入力した際の分布差＝情報量によって，グルーヴの有無を表現する事が可能である．

また変分オートエンコーダは，その学習の際，以下の ELBO を目的関数$L(\boldsymbol{x})$とし，最大化する．

$$L(x) = E_{P_e(z|x)}[logP_d(x|z')] - KL[P_e(z|x)||P_d(z')] \tag{4}$$

ここで右辺の前半はエンコーダ部の確率分布によるデコーダ部の出力の期待値，後半はエンコーダ部とサンプリング部における隠れ変数の近似差である．サンプリング部分の逆伝播は通常不可能だが，ある時刻tにおける実際のサンプリング$z_t' \sim N(\mu_t,\Sigma_t)$を行わず，$z' = \mu_t + \epsilon\Sigma_t \ \ (\epsilon \sim N(0,I))$と見做して$\mu_t,\Sigma_t$を定数として扱うことで逆伝播が可能になる．すなわち単なるガウス雑音が乗った入力であると見做せ，ノイズに強い学習が可能である．

3.2 カルバックライブラー情報量による隠れ状態分布の差異抽出

時刻tにおける演奏音源と打込音源との差異は，多変量正規分布のカルバックライブラー情報量（KLD）[13]D_{KL_t}で求める（式 5）．

$$D_{KL_t} = \frac{1}{2}\left[\log\frac{\left|\boldsymbol{\Sigma}_{2_t}\right|}{\left|\boldsymbol{\Sigma}_{1_t}\right|} + tr\left({\boldsymbol{\Sigma}_{2_t}}^{-1}\boldsymbol{\Sigma}_{1_t}\right) + \left(\boldsymbol{\mu}_{2_t} - \boldsymbol{\mu}_{1_t}\right)^T {\Sigma_{2_t}}^{-1}\left(\boldsymbol{\mu}_{2_t} - \boldsymbol{\mu}_{1_t}\right) - d\right] \tag{5}$$

ここでμ_{1_t},Σ_{1_t}は時刻tにおける演奏音源の隠れ状態の平均・共分散，μ_{2_t},Σ_{2_t}は打込音源の隠れ状態の平均・共分散，dは分布の次元数を表す．この数値は打込音源が演奏音源に変化する際の情報量の増量の期待値を表すものであり，先に説明した通り，「演奏音源にあって打込音源にない」特徴すなわちグルーヴ（演奏特性）の差異を表す値であると考える．

3.3 デノイジング

本稿では，予めノイズを含めた入力データからノイズなしの元データを復号するデノイジングオートエンコーダの機構を導入する事により，元来得られなかった弱アクセント箇所の特徴抽出を試みる．しかし，変分オートエンコーダにおいて，通常のデノイジングオートエンコーダのように入力データにノイズを載せた場合，先のガウス雑音部分にノイズが吸収され，振幅が小さい箇所において平均値の学習が進まなくなる．すなわち先のガウス雑音と付与するノイズが似通っている場合，本来混合正規分布$\epsilon \sim \sum_k \pi_k N(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$で表されるべき部分が単に分散の大きい 1 つの正規分布として学習され，逆効果になると考えられる．

弱アクセント部分のようなノイズと捉えられてしまう個所を強調し先の問題に対応するため，本稿では，通常の正規分布に従うもの（弱ノイズ）に加え，振幅が一定値以下のデータにのみ弱ノイズを与えたもの(条件付きノイズ)，元の波形の逆位相となるもの（逆位相ノイズ）を付与してデノイジングする事を考える．条件付きノイズは元々の振幅が大きいところの影響を抑えるため，逆位相ノイズは振幅の小さい箇所を強調するために付与する．逆位相の値については直接求めるのが難しいため，ノイズを付与する点直近のデータを中心として移動平均を掛けた AR モデルにより推定した予測値に係数をかけ，符号を逆転させた値を用いる．

4. 実験

4.1 実験環境

本稿では対象の演奏音源として，BPM96 となる Red Hot Chili Peppers の Dani California を用いた．セクションは A メロ約 25 秒（40 拍）を抜き出した．ドラムパートのみとなる WAV ファイルの作成にはパート分離ソフトを用いた．

Vol.2022-MUS-134 No.54
Vol.2022-SLP-142 No.54
2022/6/18

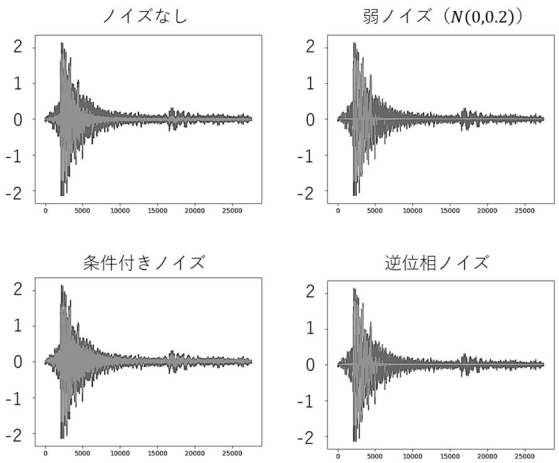

図 3　各ノイズを付与した際の波形復元の一例

打込音源は手作業で作成し, BPM やリズムパターンを元音源に一致させつつ, その他演奏特性を含まないものとした.

モデル構造は図 2 に示す通りで，使用データの都合上，BPM96 となる楽曲の 1 セクション（A メロ）の WAV データ約 25 秒分を対象として，エンコーダ部が入力 1／LSTM ユニット数 10／出力 2（2 次元），デコーダ部が入力 2／中間層 10／出力 1 とした．なお，入力時系列長は 1 拍分＝27562 とした．中間層の活性化関数は ReLU，出力層は恒等写像とし，最適化には Adam[14Ada]を利用，各重みの初期値は標準正規分布に従うランダムな値にした．学習 epoch は 10 とし，また，各セクションランダムな 32 拍分を学習データとして残りの 8 拍分を評価データとして用いた．

ノイズ付与に関して，$N(0,0.2)$に従う弱ノイズ，振幅 0.5 以下の際に$N(0,0.2)$に従う条件付きノイズ，試行錯誤的に決定した AR モデルから得られる逆位相ノイズを用いた．ある時刻tにおける点の逆位相ノイズは，その時のデータx_tに，±10 時刻での移動平均データ$\boldsymbol{X}^*$を求めたのち，時刻$t-100$からtまでのデータ$\boldsymbol{x}^* = \{x_{t'}^* | t' = t-100 \sim t,\ x_{t'}^* \in \boldsymbol{X}^*\}$を用い，ラグ 1〜10 でフィッティングした中で AIC[15]が最も高くなるモデルを用いて予測した$\hat{x}_t$に 0.3 を掛けた値を減算して求めた．

4.2 実験結果

まず，各ノイズを付与した際の LSTM 変分オートエンコーダによる演奏音源の復元波形の一例を図 3 に示す. なお，復元波形は隠れ状態のサンプリングを行わず分散を 0 とし，ガウス雑音を除いたものである．色の濃い部分が演奏音源の波形で，薄い部分が復元波形である．左上がノイズの無い通常の復元波形で，振幅が大きい箇所はよく再現出来ているが, 振幅が小さくなるにつれあまり再現出来ていない. 右上が弱ノイズを付与した際の復元波形で，元の振幅が一定以上小さい場合，復元時の振幅が 0 になっている．左下が条件付きノイズを付与した際の復元波形で，振幅が小さくなっていても上手く再現出来ている．右下が逆位相ノイズで，弱ノイズ同様，元が一定以下の振幅において復元後の振幅が 0 となっている.

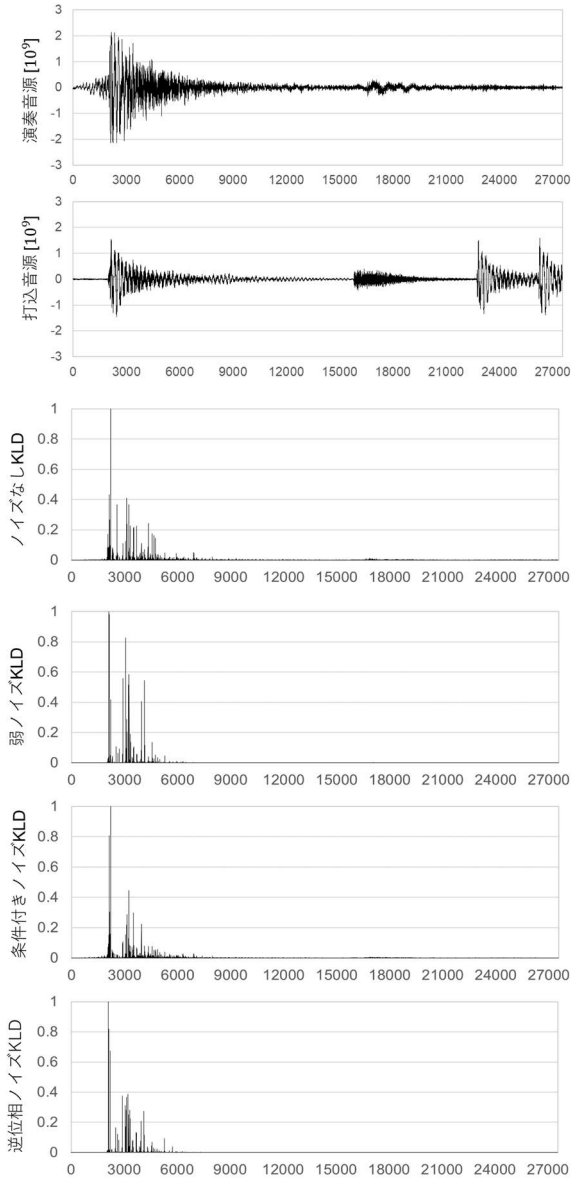

図 4　各ノイズを付与した際の KLD の一例

これらより, 3.3 にて述べた通り, 通常のデノイジングオートエンコーダのように入力データにノイズを含めた場合, 逆効果となることが分かった．また，元々の振幅が大きい箇所の影響を抑える事が重要な点であることが推察された.

次に，図 3 のデータにて算出した KLD を図 4 に示す. なお，見やすいように最大値・最小値でスケーリングしている．横軸約 18000 付近が弱アクセント部分だが，ノイズなしと条件付きノイズにおいてごくわずかにの KLD 上昇がみられるのみで，演奏特性の抽出に関してはどのノイズ付与データにおいてもデノイジングの有効性は見られなかった.

Vol.2022-MUS-134 No.54
Vol.2022-SLP-142 No.54
2022/6/18

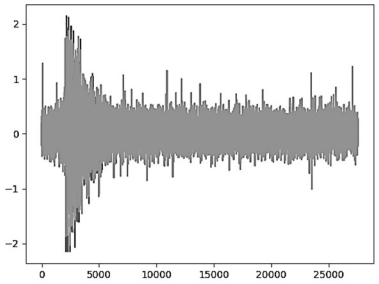

図 5　隠れ変数の分散を含め復元した波形の一例

5. 考察

図 5 は，図 3 において行わなかった隠れ変数のサンプリングを行って波形を復元した図である．これを見ると，元の音源において振幅が小さい箇所は，その振幅に比べ隠れ変数の分散が大きくなっていることが分かる．すなわち条件付きノイズによるデノイジングは，隠れ変数の分布の平均においては有効だが，分散においてはそうではない事が分かる．この分散の学習改善が，今後の課題である．

本モデルにおいて特に隠れ層の分散の学習を改善するならば，3.3 で述べた「本来混合正規分布$\epsilon \sim \sum_k \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$で表されるべき部分が単に分散の大きい 1 つの正規分布として学習される」点を考慮して対応する必要があると考える．今後，変分オートエンコーダの隠れ層からのサンプリングを混合正規分布から行うようにし，かつ誤差逆伝播を行えるようなモデルに修正する必要がある．

6. 結論

楽曲データを入力として音楽要素と演奏特性を同時かつ定量的に表現する機械学習モデルを構築するため，LSTM 変分オートエンコーダと入力データにノイズを含ませるデノイジングオートエンコーダを組み合わせたモデルについて検討した．デノイジング LSTM 変分オートエンコーダの中間層から，グルーヴがある演奏音源とグルーヴがない打込音源との差を取得することを試みたところ，1 拍分の音源を入力した際の中間層における平均出力分布のカルバックライブラー情報量（KLD）にて，演奏音源と打込音源との差が取得できた．

デノイジングにおいて，元の振幅が 0.5 以下の際に$N(0,0.2)$に従う条件付きノイズを付与し学習させたところ，デノイジングを行わなかった場合より，振幅が小さい箇所において元の波形が上手く復元できることが分かった．しかし，演奏音源と打ち込み音源の KLD においては，デノイジングを行わなかった場合と比べ差異が現れず，特徴の抽出においてはデノイジングの有効性は見られなかった．これは変分オートエンコーダの特性から現れる問題であると考えられ，今後，混合正規分布からのサンプリングを行え，かつ誤差逆伝播可能なモデルに修正する必要がある．

参考文献

[1] “しくみデザイン：「KAGURA」”，https://www.kagura.cc/jp/, (参照 2022-05-18).
[2] 宮丸友輔,江村伯夫,山田真司．ポピュラ音楽のドラムス演奏におけるグルーヴ感の研究．日本音響学会誌, 2017, vol. 73, no. 10, pp.625-637.
[3] C. Keil. Participatory discrepancies and the power of music. Cultural Anthropology, 1987, vol. 2, no. 3, pp.275-283.
[4] 奥平啓太,平田圭二,片寄晴弘.ポップス系ドラム演奏の打点時刻及び音量とグルーブ間の関連について．情報処理学会研究報告, 2004, vol. 56, pp.21-26.
[5] 藤井進也．巧みな音楽家の演奏にみられる時間のゆらぎとグルーヴ．バイオメカニズム学会誌, 2020, vol. 44, no. 4, pp.217-222.
[6] ライデン湯沢，ゼノン石川. RX モンスター・リズム・バトル．リットーミュージック. https://iss.ndl.go.jp/books/R100000002-I024194556-00, (参照 2022-05-18）.
[7] F.A. Gers, J. Schmidhuber, F.Cummins. Learning to forget: continual prediction with LSTM. Neural Computing, 2000, vol. 12, no. 10, pp. 2451–2471.
[8] 松川瞬. LSTM における中間層出力値の分布を用いたドラムグルーヴ抽出可能性の検討．令和 3 年度 電気・情報関係学会北海道支部連合大会, 2021.
[9] 松川瞬，竹沢恵，稲垣潤，真田博文．ドラムグルーヴ解析における LSTM 変分オートエンコーダ利用の検討．情報処理学会第 84 回全国大会講演論文集(分冊 2), 2022, pp. 91-92.
[10] 渡辺哲郎，近山隆．ドラム演奏のグルーヴ感の解析．情報処理学会研究報告(MUS), 2006, vol.67, no.6, pp.27-32.
[11] D. P. Kingma, M. Welling. Auto-Encoding Variational Bayes. arXiv:1312.6114, 2014.
[12] C. Doersch, Tutorial on Variational Autoencoders, arXiv:1606.05908, 2021.
[13] S. Kullback, R. A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, 1951, vol.22, no.1, pp.79-86.
[14] D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization. arXiv:1412.6980, 2017.
[15] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, vol.19, no.6, pp.716-723.